BIO Web of Conferences **138**, 02023 (2024) https://doi.org/10.1051/bioconf/202413802023
AQUACULTURE 2024

Creating an original technique for identifying objects in video streams

Islam Alexandrov[1*], *Andrei* Ruslantsev[1], *Tagirbek* Aslanov[1], and *Viktor* Gorelov[1]

[1]Institute of Design-Technology Informatics of Russian Academy of Sciences, 127055, Moscow, Russia

Abstract. In the course of practical operation in automated systems that work with multiple video streams, it often becomes necessary to identify certain objects or scenes in video frames. In this regard, this work is aimed at analyzing and improving computer vision methods that provide identification of specified images in video streams. The original method for identifying images in video streams presented below implements the calculation of descriptors and comparison of similarity measures with the search object. The presented technique is characterized by high performance and can significantly reduce the time of analyzing a video stream. A feature of the technique is to narrow the search range by clarifying the intersection of identifiers. The developed method has shown high efficiency in relation to a large part of modern analogs. In the process of creating methodology, it was compared with such identifying scene methods as, for example, the "SURF" and "Kudan" methods. The comparison results allow us to conclude that the presented method is approximately 7-23% more effective than that of analogues.

1 Introduction

We can confidently state that information on the vast majority of occurring phenomena can be adequately displayed in the format of scalar fields. Automated systems providing security, surveillance, and restrict access act as systems that analyze a wide variety of video streams. At the same time, one of the current problems in this subject area is the problem of identifying images in a video stream. This includes tracking a particular object, comparing an image with an information array, identifying copies of images, their combinations, etc.

Identifying objects is often hampered by a number of projection and affine distortions, partial overlapping as well as noise. In addition, for many practical tasks the identification technique must process the video sequence online, i.e. almost in real time.

The predominant part of universal as well as neural network techniques which involve preliminary training, are used to interact with categories of objects or scenes that are part of training collections. At the same time, training the systems under consideration is characterized by a number of difficulties. Universal methods often use dynamic learning which provides for the presence of minor projection distortions in an image of an object against the background of neighboring frames, while the identification parameters are

* Corresponding author: alexandrov@ikti.ru

adjusted to changes in the projection qualities of an object image while tracking it. Consequently, this category of techniques is ineffective for tracking objects in dynamically changing video sequences [1-6, 11].

In relation to probabilistic methods, it is required to indicate the initial location of an object in the starting frames. For this reason, the techniques under consideration are used, as a rule, only for tracking objects in video streams. Of all the existing methods, the greatest resistance to projection deformations is demonstrated by methods based on identifying key areas of an image. At the same time, their weakness is the increased complexity of information processing. Thus, the urgent task that has determined the goal of this work is to develop and improve methods of identifying scenes in a video stream that are free from the drawbacks described above.

2 Methods and materials

The functioning of the automated complex being developed should be based on identifying main sections of a video frame. Those sections, in turn, have their own specifics:

1. The main areas of an image may not belong to the area of an image or the only object in an image. They can be detected in images of other objects because the main area is local, and therefore it has a small image dimension and can also be duplicated by "similar" copies. As a result, here comes a lack of information.

2. The main identification areas may not be located throughout the entire search subject but only in certain areas of it. In other words, by recognizing the main parts of the scene, it is impossible to make a conclusion on the location of the scene itself in an image. In this regard, identifying the exact location of an object in an image solely by the location of its main sections will, at least, not be entirely correct.

In addition, the existence of indirect evidence of an object in an image does not yet serve as the ground for guaranteed identification, as well as classification, since there is a possibility of repeating the outlines of an object in other images and introducing distortions. Therefore, in order to increase the accuracy of detection and classification, it is necessary to make the gradient of the main characteristics that will ensure the classification of an object. In the list of these characteristics, we ought to highlight the discrete spectrum because it can be calculated in a short time and will not depend on the rotations and the scale of an image.

The image of an object as the common gathering of all images of an object, can be described by the following formalism:

$$F(x,y) \rightarrow \begin{cases} K^{ref} = \left\{K_0^{ref},...,K_{m-1}^{ref}\right\} \\ H^{ref} = \left\{H_0^{ref},...,H_{n-1}^{ref}\right\} \end{cases} \tag{1}$$

in this case: $F(x,y)$ – is a second-order tensor containing all point intensity values, K^{ref} – is the gradient of the main image areas with size m, H^{ref} – is the gradient of the main image areas with size n samples.

To make the system more dynamic, we suggest implementing several steps before evaluating video streams:

1. Generate a list of images that have projection distortions of the standard.
2. Identify the main areas in all generated images.
3. Calculate identifiers for all main areas.

The list of these activities for calculating the identifiers of the main areas in images with a distorted standard opens up the possibility of using techniques that can detect and classify the main areas demonstrating instability to projection transformations.

In accordance with the conditions imposed on the video monitoring complex, the technique should be based on identifying the main areas of the object in an image. The results of numerous studies demonstrate that ASIFT technique displays the greatest resistance to the listed types of influence. Note that “ASIFT” is based on “SIFT” technique which has a dynamic analogue in the form of “SURF” technique.

The SIFT technique is main-area oriented, it represents the maximum in the topic under consideration. At the same time, for the SURF technique which is described in detail in [7], the main section is the functional determinant proposed by Hesse. Practice shows that the SURF technique makes it possible to identify a much smaller number of main areas in an image of an object, but it quickly processes frames, which is discussed in detail in [8]. Therefore, in order to identify the main areas and calculate identifiers within a short time, we further suggest choosing the “SURF” method.

At the same time, a technique that allows you to narrow the range of identifying an object analyzes the image scale based on the identifiers of its main areas. This requires:

1. finding the closest similar areas from the list of deformed standards for all main areas of the image;
2. not considering the main image areas that have a similarity less than the Threshold value;
3. adding rectangular areas to all other main areas of an image;
4. considering only those main areas that are in contact with the others, with no less than 50% of the rectangular areas of the main area being studied.

The similarity of the image identifiers with the standard is calculated using the indicator proposed by Bhattacharya:

$$p = \sum_{i=0}^{n-1} \sqrt{a_i \cdot b_i} \ . \qquad (2)$$

In this case, a, b – are gradients of dimensions; n, p denote the level of similarity in the interval [0, 1] .

The presented method is characterized by a significantly lower computational complexity than the RANSAC and Cooney-Mankeers methods. But the weakness of the proposed method is the need to calculate *Thr*.

The identification technique must accurately determine whether an object in a specific image area is the image of an object or not. For that purpose, the technique must calculate the characteristics of windows in an image, taking into account the identified areas, or main areas along a discrete spectrum, which is discussed in detail in [10]. This is necessary so that minor window shifts cause minor changes to erroneous mappings:

$$K(x) = \begin{cases} 1 - x^2, |x| \le 1 \\ 0, |x| > 1 \end{cases} \qquad (3)$$

It turns out the colors of the x-points will introduce some weights $K(x)$ into the spectral bar diagram. Considerinf that, the algorithm of searching images in a video sequence can be presented in the form of a block diagram shown in Figure 1.

When identifying and classifying scenes, it is further suggested to use the “Mean Shift” approach which is described in detail in [12]. This approach is based on identifying peak densities in relation to certain functions that describe the main characteristics.

In this work, it is advised to use a four-parametric scheme for identifying windows in images of the wanted object. This scheme is described in [13]. In order to classify an object with a high probability against the background of minor changes in the spectrum, and to reduce the dimensionality, the spectrum is discretized, which is an original feature of the presented combined method for identifying scenes in video streams. Next, the SURF technique searches for the main areas using a 1:10 scale. We propose to change this scale, making it 1:20, because the main sections can only be calculated using a single and more convenient scale.

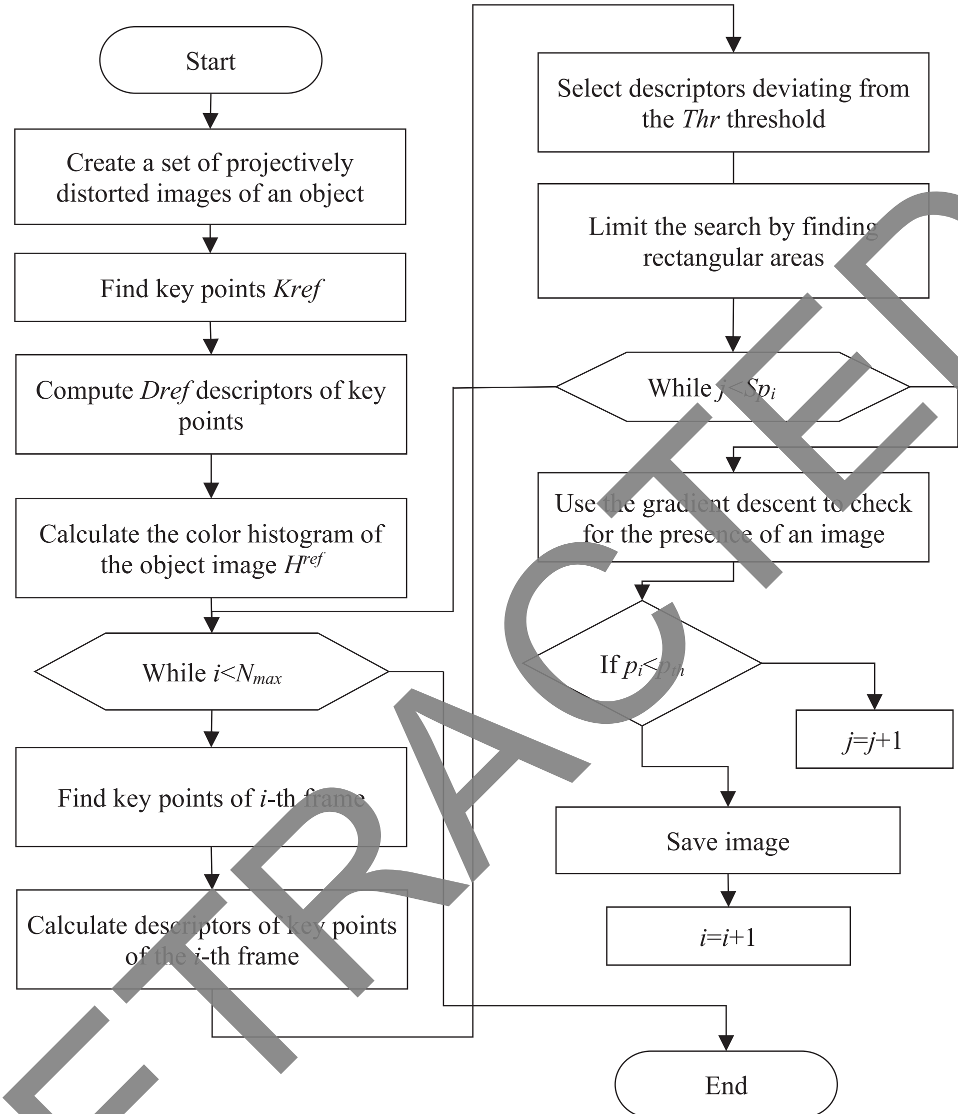

Fig. 1 – Algorithm of object search in video stream

3 Results

To test the invariance of the developed methodology, several video files were used where the rotation angle and the object scale were changed. In addition, to check for invariance, a dozen photographs of different objects were used. These images became the basis for a video sequence with a changed angle. Test results for FPR and TPR were made average. Next, Figure 2 shows the corresponding ROC dependencies.

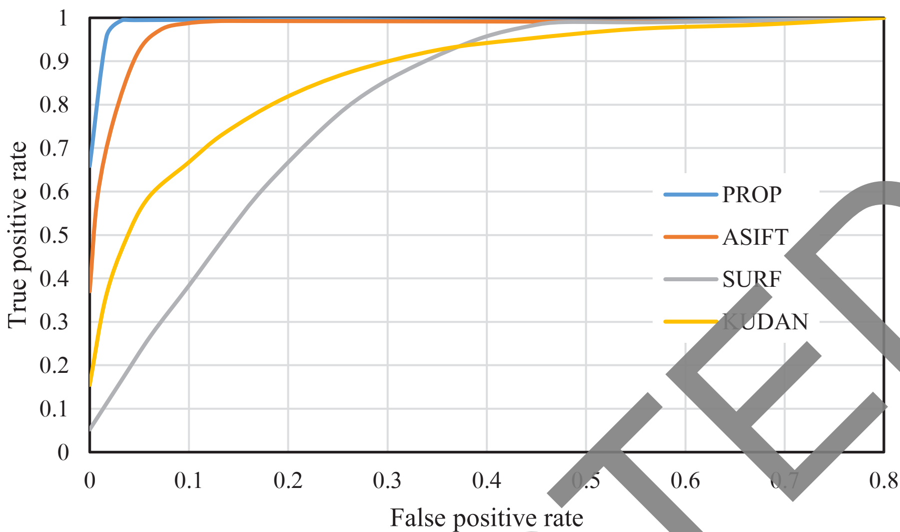

Fig 2 – ROC test performance curves for checking invariance to object rotation

From Figure 2 it is clear that the ROC curve of the created ASIFT technique passes in close proximity to the desired point. So we can conclude that we are able to detect and identify the majority of video chains. Next, it is advisable to compare methods based on the probability of identification "*p*":

$$p = \frac{TP + TN}{N} \quad (4)$$

It should be noted that, on using the above relation, it is necessary to take into account that the video sequence must have an identical number of frames the required scene is present in, or frames the desired scene should not be present in.

In order to determine the statistical indicators of each method, the *TP* and *TN* testing results were summed up. Threshold similarity values for each method were selected, taking into account tests for invariance and the presence of changes when a smaller number of errors are detected. Figure 3 shows the detection probabilities for each method. The total number of images *N* is 944.

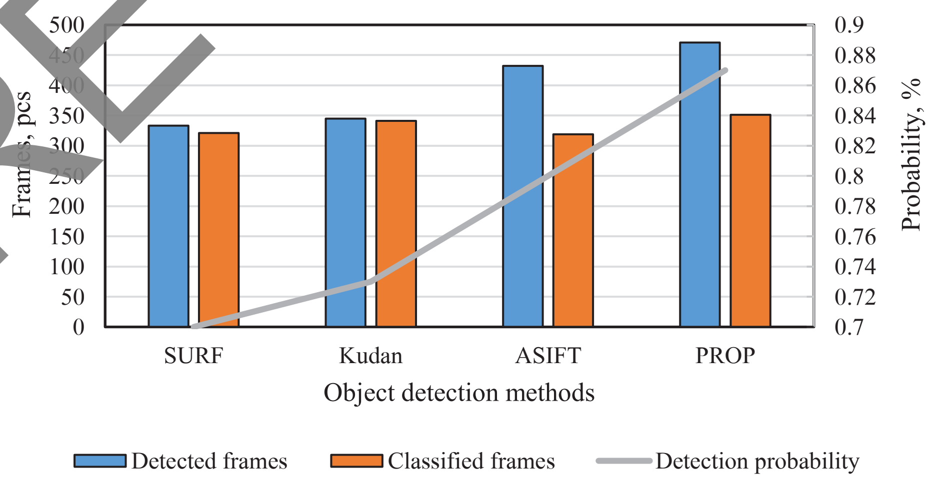

Fig 3 – probabilities of detecting objects in video sequences using various methods

4 Conclusion

The end result of the presented method for identifying scenes demonstrates that it should be based on identifying the main sections of the scene, calculating their descriptors to make a decision on the presence or absence of the subject to identify.

The presented method for identifying an object in a video sequence is based on assessing its main sections, as well as on the classification of distortions of each image in video frames. It ultimately increases the effeciency of the method by 7 to 24% compared to analogs. The presented method for identifying an object in video streams involves calculating descriptors of the main sections, considering a distorted standard and analyzing the degree of similarity of frame descriptors, as well as descriptors of the wanted scene. It ensures more accelerated information processing relative to the well-known analogue – the "SURF" technique.

5 Acknowledgements

The presented results are part of the work performed under the Subsidy Agreement dated April 8th , 2022 No. 075-11-2022-029 on the topic: "Creation of an infrastructure digital complex for storing and processing data, followed by the formation of a dynamic flow using artificial intelligence technologies" with the Ministry of Science and Higher Education of the Russian Federation.

References

1. Kuklin, V. Zh., Tatarkanov, A. A., & Umyskov, A. A. (2023). Trainable Regularization in Dense Image Matching Problems. HighTech and Innovation Journal, 4(3), 617–629. https://doi.org/10.28991/hij-2023-04-03-011
2. Zhang, F., Xu, X., & Qiao, Y. (2015). Deep classification of vehicle makers and models: The effectiveness of pre-training and data enhancement. 2015 IEEE International Conference on Robotics and Biomimetics (ROBIO). https://doi.org/10.1109/robio.2015.7418772
3. Alexandrov, I. A., Kirichek, A. V., Kuklin, V. Z., & Chervyakov, L. M. (2023). Development of an Algorithm for Multicriteria Optimization of Deep Learning Neural Networks. HighTech and Innovation Journal, 4(1), 157–173. https://doi.org/10.28991/hij-2023-04-01-011
4. Uijlings, J. R. R., van de Sande, K. E. A., Gevers, T., & Smeulders, A. W. M. (2013). Selective Search for Object Recognition. International Journal of Computer Vision, 104(2), 154–171. https://doi.org/10.1007/s11263-013-0620-5
5. Zulkifley, M. A., Rawlinson, D., & Moran, B. (2012). Robust Observation Detection for Single Object Tracking: Deterministic and Probabilistic Patch-Based Approaches. Sensors, 12(11), 15638–15670. https://doi.org/10.3390/s121115638
6. Inglada, J. (2007). Automatic recognition of man-made objects in high resolution optical remote sensing images by SVM classification of geometric image features. ISPRS Journal of Photogrammetry and Remote Sensing, 62(3), 236–248. https://doi.org/10.1016/j.isprsjprs.2007.05.011
7. Krishna, R. S. B., Bharathi, B., Ahamed. A, M. U., & Ankayarkanni, B. (2019). Hybrid Method for Moving Object Exploration in Video Surveillance. 2019 International

Conference on Computational Intelligence and Knowledge Economy (ICCIKE). https://doi.org/10.1109/iccike47802.2019.9004330

8. Rashid, M., Khan, M. A., Sharif, M., Raza, M., Sarfraz, M. M., & Afza, F. (2018). Object detection and classification: a joint selection and fusion strategy of deep convolutional neural network and SIFT point features. Multimedia Tools and Applications, 78(12), 15751–15777. https://doi.org/10.1007/s11042-018-7031-0
9. Alvarez-Meza, A. M., Lee, J. A., Verleysen, M., & Castellanos-Dominguez, G. (2017). Kernel-based dimensionality reduction using Renyi's α -entropy measures of similarity. Neurocomputing, 222, 36–46. https://doi.org/10.1016/j.neucom.2016.10.004
10. Khokher, A., & Talwar, R. (2016). A fast and effective image retrieval scheme using color-, texture-, and shape-based histograms. Multimedia Tools and Applications, 76(20), 21787–21809. https://doi.org/10.1007/s11042-016-4096-5
11. Petrovna, V. L. (2020). Convolution of Images Using Deep Neural Networks in the Recognition of Footage Objects. Studies in Computational Intelligence, 171–191. https://doi.org/10.1007/978-3-030-51920-9_9
12. Zhou, H., Yuan, Y., & Shi, C. (2009). Object tracking using SIFT features and mean shift. Computer Vision and Image Understanding, 113(3), 345–352. https://doi.org/10.1016/j.cviu.2008.08.006
13. Iqbal, S., Qureshi, A. N., & Lodhi, A. M. (2018). Content Based Video Retrieval Using Convolutional Neural Network. Intelligent Systems and Applications, 170–186. https://doi.org/10.1007/978-3-030-01054-6_12