EasyChair Preprint

№ 4258

Acoustic Pattern Recognition Technology Based on the Viola-Jones Approach for VR and AR Systems

Alexander Alyushin and Sergey Dvoryankin

September 25, 2020

Acoustic Pattern Recognition Technology Based on the Viola-Jones Approach for VR and AR Systems

A. M. Alyushin[1][0000-0003-1722-0598], S.V. Dvoryankin[2][0000-0001-6908-0676]

[1]National Research Nuclear University MEPhI (Moscow Engineering Physics Institute),
Kashirskoe shosse 31, Moscow, 115409, Russia
Plekhanov Russian University of Economics
st. Hook, 43, Moscow, 117997, Russia
[1]alyshin@list.ru
[2] Financial University under the Government of the Russian Federation,
Moscow, Russia
Leningradsky prospect, 49, Moscow, Russian Federation, 125993
[2]SVDvoryankin@fa.ru

Abstract. The ability to solve problems of graphic images recognition in VR, AR, MR and XR systems is highlighted as one of the most important. The urgency of solving problems of recognition and classification of acoustic images has been substantiated, which will bring the quality of VR, AR, MR and XR systems closer to real reality (RR). Independent solution of graphic and acoustic patterns recognition problems using heterogeneous algorithmic and software tools is attributed to the disadvantages of modern systems. The study proposes an approach that allows the use of unified methodological and software tools for the simultaneous solution of graphic and acoustic patterns recognition problems. The proposed approach is based on converting acoustic information into graphic information using 2D-images of dynamic sonograms. This allows the recognition of acoustic patterns using unified algorithmic and software tools. It is proposed to use the Viola-Jones technology as such a unified tool. It is shown that the implementation of a two-stage determination of similarity measures of primitives and areas of the original image makes it possible to increase the speed of algorithms. For this purpose, at the first iteration, it is proposed to use not the graphic primitives themselves, but their coordinate projections. In the study, by analogy with Haar's features, parametrizable acoustic primitives were developed, presented in the classical graphical version, as well as in the form of coordinate projections.

Key words: Recognition and Classification, Graphic and Acoustic Patterns, Viola-Jones Algorithm.

1 Introduction

One of the most important functions implemented in modern VR, AR, MR and XR systems and largely determines their capabilities, is the recognition and classification of graphic images [1]. For example, the basis of AR, MR and XR technologies is the

recognition of objects in each frame of the video stream and the addition of new graphic information to them. In addition, a number of AR, MR and XR technologies use the so-called graphic markers, which are necessary to determine the spatial characteristics of objects of real reality (RR) [2]. This function must be repeatedly performed in real time [3], which imposes strict requirements on the speed of algorithms used for this purpose, for example, various modifications of the well-known Viola-Jones algorithm [4-5].

One of the main trends in the modern development of AR, MR, and XR systems is the approach to RR, primarily due to the development of artificial intelligence technology, which allows simultaneous processing of video and audio patterns [6]. This allows you to implement a natural user interface for a person, to carry out acoustic navigation [7-9], to provide the necessary information interaction between characters, for example, in training systems [10, 11], to increase the efficiency of solving production tasks [12, 13].

The aim of the research is to unify algorithmic and technical means used for the recognition of graphic and acoustic patterns based on the Viola-Jones approach for VR, AR, MR and XR systems.

2 State of Research in This Area

Currently being developed VR, AR, MR and XR systems, as a rule, involve the use of independent channels for processing video and audio information. It should be noted that the applied algorithmic and methodological means of graphic patterns recognition are forced to operate with significant data streams, which is associated with the use of modern high-speed high-resolution video cameras. In this regard, the flow of acoustic data even when using multichannel systems [7] has a significantly smaller volume.

For this reason, it is relevant to use algorithmic and methodological tools developed for the recognition of graphic patterns for the recognition of acoustic patterns in VR, AR, MR and XR systems. The feasibility of this approach in practice is due to a fairly well-developed technology for converting acoustic information into a graphic representation in the form of a 2D-image of dynamic sonograms. This technology is widely used, for example, for the protection of documents against forgery based on the so-called speech signature [14, 15]. The specificity of a 2D-image of a dynamic sonogram is the presence of areas with different graphic structures, for example, linear and dotted, as well as low image contrast with a high noise level.

Of all the existing variety of approaches and algorithms for recognizing graphic objects in the image, which are also suitable for working with images of dynamic sonograms, the most suitable is the approach proposed by Viola-Jones for recognizing facial images [16]. Recognition of a graphic object in accordance with this approach is carried out on the basis of the similarity measures analysis of a large set of characteristic features typical of the analyzed image. At the same time, the features themselves, known as Haar features [16-18], characterize the properties of limited areas of the recognized object.

The disadvantages of this approach include a fairly large amount of calculations to determine the measures of features similarity and areas of the analyzed image, as well as a decrease in the reliability of the result obtained with a decrease in the contrast of the analyzed image.

3 The Essence of the Proposed Approach

To recognize acoustic patterns presented in the form of 2D-images of dynamic sonograms, the study proposes to use a two-stage analysis of similarity measures for a set of acoustic features and areas of a 2D-image. This allows you to significantly reduce the necessary computational costs of the approach. This approach assumes the presence of two forms of features representation and analyzed image areas − in the form of image fragments and in the form of their coordinate projections. A two-stage analysis of features similarity measures is applicable both in the analysis of acoustic signals and in the analysis of video stream frames.

At the first step, it is proposed to analyze the features similarity of the coordinate projections and the coordinate projections of the areas of the analyzed image, usually selected using a floating rectangular window.

At the second step, the analysis of the features similarity measures selected in this way and the corresponding image areas, presented as fragments of the corresponding images, is carried out. This operation is completely analogous to the procedures used in the Viola-Jones algorithm.

Fig. 1 illustrates this approach.

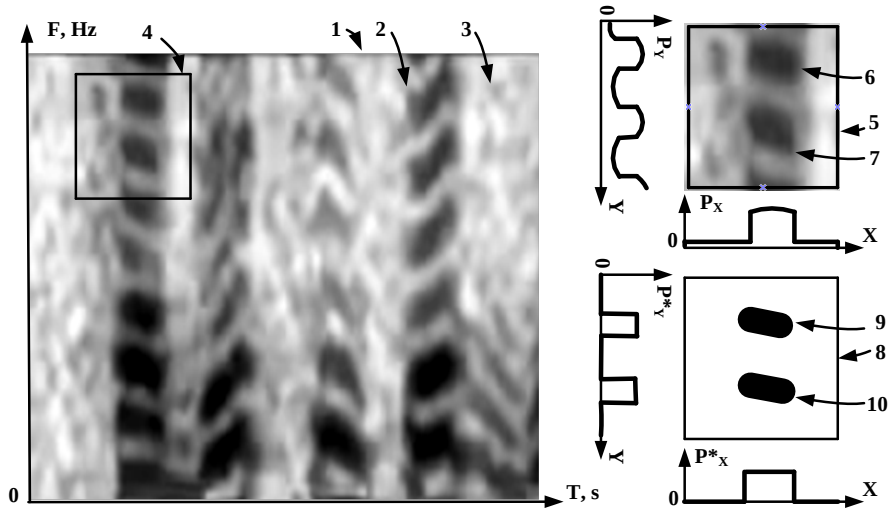

Fig. 1. Approach illustration.

The original image ($IMG[i, j]$, i=1,…, $IMAX$, j=1,…, $JMAX$, where $IMAX$ and $JMAX$ are the image size in pixels, respectively) of dynamic sonogram 1 contains

areas 2 with a linear structure corresponding to vowel sounds of human speech (harmonic signals), and area 3 corresponding to consonants (hissing) sounds. The X-axis of the sonogram corresponds to time, and the Y-axis corresponds to the frequency F.

At the first step, in accordance with the proposed approach, by means of a floating window 4 with a size of $N \times N$ pixels, a fragment of the image 5 is selected, for example, containing characteristic stripes 6 and 7.

For a given fragment of the image, its coordinate projections $P_Y(Y)$ and $P_X(X)$ are determined:

$$P_X(X) = \sum_{j=Y_0}^{Y_0+N} IMG(X, j), \quad X=1, \ldots, N, \tag{1}$$

$$P_Y(Y) = \sum_{i=X_0}^{X_0+N} IMG(i, Y), \qquad Y=1, \ldots, N, \tag{2}$$

where X_0 and Y_0 are the coordinates of the window anchor point.

These projections, as a rule, contain noise components that cause, for example, nonzero values of the functions $P_Y(Y)$ and $P_X(X)$ in the intervals between bands 6 and 7. To minimize the influence of noise components in two forms of information presentation, it is proposed to carry out, respectively, threshold discrimination for the functions $P_Y(Y)$ and $P_X(X)$ and contrasting for a fragment of the image 5. A typical result of these operations is shown in Fig. 1 in the form of new obtained coordinate projections $P^*_Y(Y)$ and $P^*_X(X)$, as well as a fragment of a contrast image 8, containing characteristic stripes 9 and 10.

Analysis of possible structures of image fragments of dynamic sonograms made it possible to form a basic set of characteristic acoustic features. To describe the image areas of a sonogram with a line structure, basic features *SL*(*A*, *D*, *W*) were formed, where *A* is the relative angle of inclination of the lines (-*AMAX* ≤ *A* ≤ *AMAX*, the value *AMAX*=10, which corresponds to the angle of inclination of the lines in 90^0), *D* is the relative distance between the lines (1≤ *D* ≤ *DMAX*, the value *DMAX*=10 corresponds to the maximum distance), *W* is the relative width of the lines (1≤ *W*≤ *WMAX*, the value *WMAX*=10 corresponds to the maximum width).

In Fig. 2 shows examples of the formed basic acoustic features, which include two forms of presenting information of the considered type − in the form of coordinate projections and a fragment of a high-contrast image.

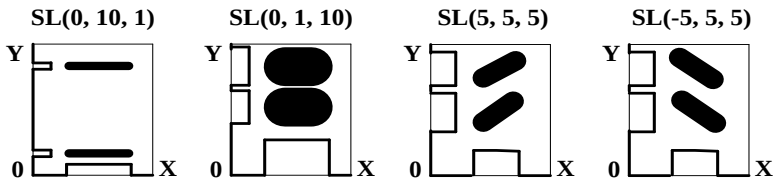

Fig. 2. Examples of *SL*(*A*, *D*, *W*) basic acoustic features.

To describe image areas with a pixel structure, basic features of the *NP*(*Q*, *G*) type were formed, where *Q* is the relative density of dark pixels ($1 \le Q \le QMAX$, the value $QMAX=10$ corresponds to the maximum density of dark pixels), *G* − relative density gradient with respect to the Y axis ($-GMAX \le G \le GMAX$, the value $GMAX=10$ corresponds to the maximum value of the gradient).

In Fig. 3 shows typical examples of formed basic features of the *NP*(*Q*, *G*) type.

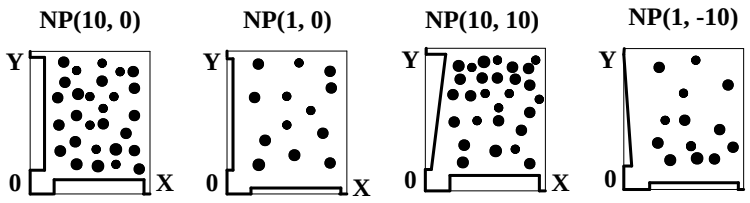

Fig. 3. Examples of *NP*(*Q*, *G*) basic acoustic features.

The basic acoustic features of *SL*(*A*, *D*, *W*) and *NP*(*Q*, *G*) types formed in this way were the basis for the creation of working features *SL**(*A*, *D*, *W*, *M*) and *NP**(*Q*, *G*, *M*), differing in the scaling factor *M* ($1 \le M \le MMAX$, for sonograms with a resolution of less than 1024x1024 pixels, it is sufficient to use the value $MMAX=10$).

In Fig. 4 shows an example of the created operating characteristics *SL*(*A*, *D*, *W*).

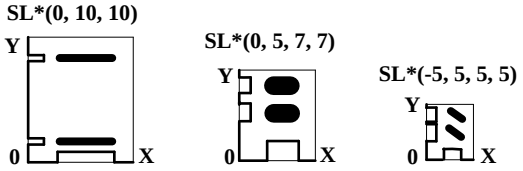

Fig. 4. Examples of working acoustic signs of *SL**(*A*, *D*, *W*, *M*) type.

Working acoustic signs allow you to get a graphic image of acoustic signals, speech and sounds based on recognition of the structure of the sonogram. To determine the similarity measure of coordinate projections at the first step, it is proposed to use the value *D*:

$$D = 1/(1+\sum_{Y=1}^{N}\left|P_F(Y) - P^*_Y(Y)\right| + \sum_{X=1}^{N}\left|P_F(X) - P^*_X(X)\right|), \qquad (3)$$

where $P_F(Y)$ and $P_F(X)$ are, respectively, coordinate projections for working acoustic features.

In Fig. 5 shows an example of structure recognition according to the proposed technique of 2D-image of a dynamic sonogram shown in Fig. 1 (X-axis time scale not saved).

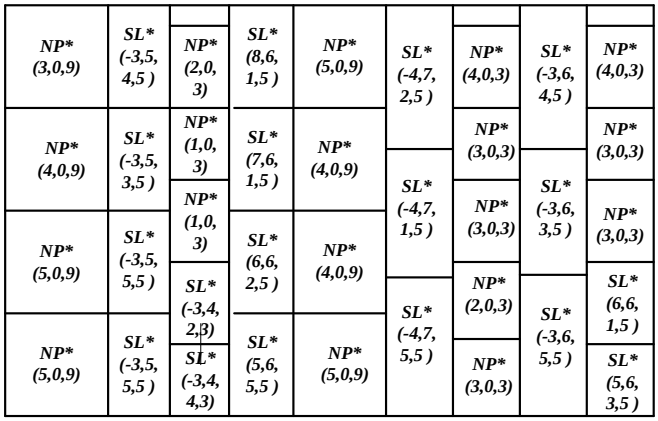

Fig. 5. Example of structure recognition.

The resulting structure of the sonogram is an image that is further processed by the methodological and algorithmic means inherent in the Viola-Jones approach. This makes it possible to use the existing graphic image processing tools for solving problems of recognition and classification of acoustic patterns in VR, AR, MR and XR systems.

4 Experimental Laboratory Approbation of the Approach

The conducted experimental laboratory testing of the approach confirmed the possibility of its implementation in practice using the classical Viola-Jones algorithms. The decrease in the performance of the tasks being solved for the recognition of graphic objects with the simultaneous processing of acoustic patterns did not exceed 10%. The implementation of a two-stage process for determining the measures of similarity of features and areas of an image made it possible to increase the performance by 15-30% for images with dimensions of 800x800 pixels - 1600x1600 pixels, respectively.

5 Areas of Possible Application of the Developed Technology

The proposed approach is primarily focused on expanding the functionality of modern VR, AR, MR and XR systems through the simultaneous processing of video and acoustic information. Another area of possible application of the approach is systems for protecting important documents based on the use of a speech signature [14,15], which involve solving the problems of searching for a sonogram on a document image, as well as recognizing the structure of a sonogram in order to identify the author of the document and his psycho-emotional state.

6 Conclusion

The approach proposed in the study makes it possible to use already available software and methodological tools for solving problems of recognizing and classifying acoustic patterns, initially focused on recognizing and classifying graphic images. The most prominent representative of such tools are algorithmic and software tools that implement the principles of processing graphic data in accordance with Viola-Jones technology. The implementation of a two-stage procedure for determining the similarity measure of features and image regions allows increasing the speed of the computational process.

7 Acknowledgement

The research was carried out by grant of the Russian Scientific Foundation (project №19-71-30008) in Plekhanov Russian University of Economics

References

1. Park, H., Jeong, S., Kim, T., Youn, D., Kim, K.: Visual representation of gesture interaction feedback in virtual reality games. In: IEEE Proc. of the 2017 International Symposium on Ubiquitous Virtual Reality (ISUVR), Nara, Japan, 27-29 June 2017, pp. 20–23 (2017).
2. Kato, H., Billinghurst, M.: Marker tracking and HMD calibration for a video-based augmented reality conferencing system. In: Proceedings 2nd IEEE and ACM International Workshop on Augmented Reality (IWAR'99), San Francisco, CA, USA, 20-21 Oct. 1999, pp. 85–94 (1999).
3. Prince, S., Cheok, A. D., Farbiz, F., Williamson, T., Johnson, N., Billinghurst, M., Kato, H.: 3D Live: real time captured content for mixed reality. In: IEEE Proc. of the International Symposium on Mixed and Augmented Reality (ISMAR'02), Darmstadt, Germany, 1−1 Oct. 2002, pp. 7–317 (2002).
4. Lee, Y. J.: Effective interface design using face detection for augmented reality interaction of smart phone. In: International Journal of Smart Home, vol. 6(2), pp. 25-32 (2012).
5. Lee, Y. J., Lee, G. H.: Augmented Reality Game Interface Using Effective Face Detection Algorithm. In: International Journal of Smart Home, vol. 5(4), pp. 77−88 (2011).
6. Godin, K. W., Rohrer, R., Snyder, J., Raghuvanshi, N.: Wave acoustics in a mixed reality shell. In: Proc. of the 2018 AES International Conference on Audio for Virtual and Augmented Reality (AVAR), Redmond, USA, 20-22 August, pp. 7−3 (2018).
7. Tylka, J. G., Choueiri, E. Y.: Fundamentals of a parametric method for virtual navigation within an array of ambisonics microphones. In: Journal of the Audio Engineering Society, vol. 68(3), pp. 120−137 (2020).
8. Tylka, J. G.: Virtual navigation of ambisonics-encoded sound fields containing near-field sources. In: Computer Science, 246 p. (2019).
9. Tylka, J. G., Choueiri, E. Y.: Performance of linear extrapolation methods for virtual sound field navigation. In: Journal of The Audio Engineering Society, vol. 68(3), pp.138−156 (2020).
10. Yuen, S. C.-Y., Yaoyuneyong, G., Johnson., E.: Augmented reality: an overview and five directions for AR in education. In: Journal of Educational Technology Development and Exchange, vol. 4(1), pp. 119−140 (2011).

11. Abdoli Sejzi, A.: Augmented reality and virtual learning environment. In: Journal of Applied Science Research (JASR), vol. 11(8), pp. 1−5 (2015).
12. Lahti, H., Bahne, A.: Virtual Tuning – A mixed approach based on measured RTFs. In: Proc. of the AES 2019 International Conference on Automotive Audio, Bavaria, Germany, September 11–13. Paper Number: 12. (2019).
13. Malbos, F., Bogdanski, M., Strauss, M. J.: Virtual reality experience for the optimization of a car audio system. In: Proc. of the AES 2019 International Conference on Automotive Audio, Bavaria, Germany, September 11–13. Paper Number: 13. (2019).
14. Alyushin, M. V., Alyushin, A. M., Kolobashkina, L. V.: Human face thermal images library for laboratory studies of the algorithms efficiency for bioinformation processing. In: IEEE Proc. of the 11th IEEE International Conference on Application of Information and Communication Technologies, (AICT 2017), Russia, Moscow, 20-22 September, (2017).
15. Alyushin, A. M.: Document protection technology in the digital economics using cognitive biometric methods. In: Procedia Computer Science, vol. 169, pp. 887-891 (2020).
16. Viola, P., Jones, M. J.: Robust real-time face detection. In: International Journal of Computer Vision, vol. 57(2), pp.137−154 (2004).
17. Alyushin, M. V., Alyushin, V. M., Kolobashkina, L. V.: Optimization of the data representation integrated form in the viola-jones algorithm for a person's face search. In: Procedia Computer Science, vol. 123, pp. 18−23 (2018).
18. Kolobashkina, L. V., Alyushin, M. V.: Analysis of the possibility of the neural network implementation of the Viola-Jones algorithm. In: Advances in Intelligent Systems and Computing, vol. 948, pp. 232−239 (2020).