ОБЗОРНЫЕ СТАТЬИ

MSC 68T10, 68T45 DOI: 10.14529/mmp210401

EVOLUTION OF THE VIOLA–JONES OBJECT DETECTION METHOD: A SURVEY

V.V. Arlazarov[1,2,3,4]***, Ju.S. Voysyat***[1]***, D.P. Matalov***[1,3]***, D.P. Nikolaev***[1,2,4]***, S.A. Usilin***[1,2,3]

[1]Smart Engines Service LLC, Moscow, Russian Federation
[2]Moscow Institute of Physics and Technology, Moscow, Russian Federation
[3]FRC CSC RAS, Moscow, Russian Federation
[4]Institute for Information Transmission Problems, Moscow, Russian Federation
E-mails: vva@smartengines.com, u.voisiat@smartengines.com, d.matalov@smartengines.com, d.p.nikolaev@smartengines.com, usilin@smartengines.com

The Viola and Jones algorithm is one of the most well-known methods of object detection in digital images. Over the past 20 years since the first publication, the method has been extensively studied, and many modifications of the original algorithm and its individual parts have been proposed by researchers and engineers. Some ideas popularized by Paul Viola and Michael Jones became the basis for many other algorithms of object localization in images. This paper presents a description of Viola and Jones algorithm, the history of its development and modifications in the context of various problems of object localization in images, as well as a description of the current state of affairs: the method's place in the era of convolutional neural networks extensive application.

Keywords: Viola–Jones algorithm; pattern recognition; machine learning; object classification; object localization; object detection.

Introduction

Almost every resident of a modern metropolis encounters computer vision systems [1–3]. For example, facial recognition systems [2, 4] are one of the most widely known practical applications of computer vision algorithms today. One of the problems solved by such systems is the problem of object localization in an image. Despite significant progress in the development of localization methods achieved by the end of the 20th century, the algorithms back then [5, 6] could hardly be called a solution to the problem: the quality characteristics (the ratio of correctly detected objects to all detections and the number of falsely detected objects) and performance were a significant obstacle to the industrial-scale application. The milestone at which object detection systems began their widespread use was reached in 2001. Paul Viola and Michael Jones published an object detection framework, which was used to obtain a high-quality detector of frontal upright faces [7] suitable for real-time face detection in images. The concepts laid down by Viola and Jones in their original work have been independently researched and developed by many scientists around the world, and the method itself has become de facto one of the main tools for building high-performance object detectors.

The Viola and Jones method [7] is a combination of four approaches in a single design: computationally lightweight features; boosting classifiers; a cascade of classifiers that

provides high performance; and the sliding window method to locate objects. The Viola–Jones detector belongs to the family of sliding window methods, where rectangular areas of an image are "scanned" with some step by the classifier for the content of the target object. In the Viola–Jones method, the classifier is binary, and the positive responses of this classifier are aggregated and form the parameters of one or more rectangles – the result of the object localization algorithm. Intuitively, with this approach, especially when considering the detection of objects at multiple scales, the number of classifier runs for a single image can reach millions of times. To ensure extreme performance, Viola and Jones proposed the use of efficiently computed Haar features, the concept of an integral image, and a cascade structure of the classifiers. Viola and Jones method itself is a scheme for building such detectors using the classical machine learning algorithm based on AdaBoost [8] training. AdaBoost was used both to select the most efficient features and to build the binary classifiers. For this, the authors associated each feature with a so-called weak classifier, a decision stump classifier. A set of such weak classifiers is subsequently fed to the AdaBoost algorithm, which forms a so-called strong classifier. The last important part of the Viola and Jones method is a way to combine a set of strong classifiers into a cascade structure, which provides high detector performance on "empty" sub-windows of images that do not contain the target object. The successful combination of these ideas allowed for the building of the face detector capable of "processing" up to 15 images per second on a publicly available general-purpose processor in 2001. Below, we take a closer look at each fundamental part of this method.

1. Original Viola and Jones Algorithm

1.1. Rectangular Haar Features and Integral Image

In the original Viola and Jones paper, each weak classifier is associated with a Haar feature, which is based on the Haar wavelets [9]. This choice is primarily due to the fact that Haar features allow for the determining of the characteristic features associated with local brightness variations within the object. For example, with Haar features, it is easy to reflect the fact that the eye area is darker than the nose area in an image of a human face (Fig. 1). The original Viola–Jones approach uses three types of features: two-rectangle features, three-rectangle features, and four-rectangle features. The value of Haar features is computed as the difference between the sum of pixels within black and white rectangles of equal size within the image areas (Fig. 1). Two-rectangle and three-rectangle Haar features can be oriented either vertically or horizontally. An integral representation of the image is proposed to calculate the value of Haar features effectively.

Fig. 1. Haar features

Definition 1. *Let a grayscale image $f(y, x)$ be of the size $M \times N$. $I_f(y, x)$, an integral representation of the image $f(y, x)$ be a digital image of the size $(M + 1) \times (N + 1)$, where pixel values are calculated as follows:*

$$I_f(y,x) = \begin{cases} 0, & if\ y = 0\ or\ x = 0; \\ \sum_{y'<y,x'<x} f(y',x'), & if\ x > 0\ and\ y > 0. \end{cases} \quad (1)$$

Using the integral representation of an image, the sum of the pixel intensities of an image within a certain rectangular area can be calculated in four arithmetic operations, regardless of the size of that area, by the following formula:

$$\sum_{y_1 \leqslant y \leqslant y_2} \sum_{x_1 \leqslant x \leqslant x_2} f(y,x) = I_f(y_2,x_2) - I_f(y_1,x_2) - I_f(y_2,x_1) + I_f(y_1,x_1). \quad (2)$$

Therefore, the concept of integral image allowed the authors to construct an extremely efficient algorithm for calculating the feature description of localizable objects.

1.2. Training Strong Classifier by AdaBoost Algorithm

Each weak classifier is represented by a decision tree with the root and two leaves, a decision stump, which compares the numerical value of Haar feature to a threshold value:

$$b(x) = \begin{cases} +1, & if\ h(x) < t \\ -1, & if\ h(x) \geq t, \end{cases} \quad (3)$$

where t is the threshold value, $h(x)$ is a value of Haar feature. Viola and Jones use the *AdaBoost* algorithm [8] to determine optimal thresholds and select the most effective features. In its original form, the *AdaBoost* learning algorithm was proposed to improve classification efficiency by multiple simple (sometimes called weak) classifiers. The algorithm constructs a composition of such weak classifiers (a strong classifier) as a linear combination of weak classifiers:

$$S(x) = \sum_{j=1}^{J} \alpha_j \cdot b_j(x), \quad (4)$$

where α_j is the weight of the classifier, $b_j(x)$ is the response of the classifier. Besides theoretical estimates of generalizability and convergence [8], an important property is the ability of the algorithm to construct "very quickly" a composition that exhibits much better classification performance than individual weak classifiers. In [7], it is claimed that it is possible to consrtuct a strong classifier consisting of 2 weak classifiers, which correctly classify 100% of face images and yield false positives for 40% of non-face images. The number of computed features required for classification closely correlates with the performance of the detector.

1.3. Cascade of Strong Classifiers

The object localization process is based on the scanning window method, a hypothesis on the presence of the target object within this window is tested across all parts of the image at several scales. To ensure high detector speed, Viola and Jones suggested using a cascade of strong classifiers. The strong classifiers are used sequentially, and the next strong classifier is applied only if the previous one yielded a positive response. The classical cascade of strong classifiers can be represented as a conjunction of the responses of strong classifiers:

$$Cascade(x) = \prod_{n=1}^{N} \Big[S_i(x) > 0 \Big]. \tag{5}$$

The intuition behind this idea is the property of object localization by the scanning window method: the number of regions containing the target object is orders of magnitude smaller than the number of windows that do not contain it. Therefore, the first strong classifiers in the cascade are trained in such a way as to provide 100% correct answers on the windows containing an object and minimize false positives on the set of "simple" windows that do not contain the object, by calculating the smallest possible number of weak classifiers. The subsequent strong classifier is trained on examples that have "passed" all previous strong classifiers: on false positives "that have reached" this stage, and on the remaining correctly classified true positive windows containing object. As the learning process progresses and another strong classifier is added to the cascade, the classifier includes more and more features. The final cascade consists of 32 strong classifiers with 4297 features. Evaluated on the MIT+CMU test set, an average of 8 features are computed per sub-window.

2. Further Development of Viola and Jones Algorithm Ideas

Viola and Jones revolutionized the methods of object localization in images and provided a high-quality solution to the relevant problem of facial detection. In fact, it was the first method to provide industrial quality recognition at a rate close to the video camera frame rate. The method yielded similar or better detection quality on the widely used MIT+CMU set of real data compared to previously proposed methods while being 15 times faster.

The approach proposed by Viola and Jones to construct object detectors could not go unnoticed by the scientific community. The researchers proposed many modifications to each basic element of the original algorithm: in order to improve the qualitative characteristics and to adapt the approach to the detection of objects of other nature.

2.1. Weak Classifiers Feature Space Modifications

The original Haar rectangle feature space describes the objects with local brightness variations along the vertical and horizontal directions well. However, there are object types in which the contrast structure is significantly more complex. To construct high-quality detectors for such objects, many modifications of the original feature space were proposed by various researchers.

Lienhart and Maid [10] proposed expanding the feature space by additional 45-degree rotated Haar features (first row in Fig. 2). To compute these new features quickly, Lienhart and Maid suggested using an additional rotated integral image representation (Rotated Summed Area Table), which is defined as follows:

$$I_f^r(y,x) = \begin{cases} 0, & if\ y = 0\ or\ x = 0 \\ \sum_{x'<x-|y-y'|} f(y',x'), & if\ x > 0\ and\ y > 0. \end{cases} \tag{6}$$

The rotated integral representation of an image can be quickly computed by two passes over the pixels of the original image. Using the pre-calculated image $I_f^r(y,x)$, the sum of

pixels inside a rotated rectangle can be calculated in 4 arithmetic operations regardless of the size of the rectangle. The proposed features were later used to detect birds [11], cars [12], people and cars in aerial images [13].

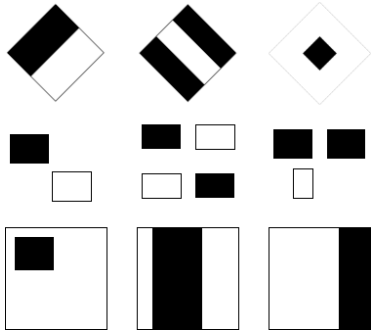

Fig. 2. Extended Haar features

To improve the arbitrarily located faces localization performance, [14] proposes to extend the feature space with so-called "non-contiguous" Haar features (second row in Fig. 2). The only difference between these features and ordinary Haar ones is that the rectangles can be at some distance from each other. This accounts for asymmetric contrast structure inherent in arbitrarily placed (rotated, in profile, etc.) faces in the image.

In 2003, Viola and Jones developed and published [15] an extension of their original work, which deals with the same problem of poor localization quality of randomly situated faces. One of the contributions of this work is a new type of feature that targets the diagonal contrast structure of target objects. A feature consists of 4 overlapping rectangles that together form blocky diagonal regions. The feature value is still effectively computed in 16 operations of an integral image reading at the corner points of all rectangles which contribute to the feature. Viola and Jones later employed these features to detect pedestrians [16].

Messom and Barczak further developed the concept of rotated Haar features by proposing a way to calculate feature values at arbitrary angles [17]. However, this type of features was rarely used in other works. This is most likely due to the low quality of the rotated features pixel coordinates approximation during feature value calculation.

In [18], the asymmetric Haar features (third row in Fig. 2) were proposed. These features increase the accuracy of the classifier and reduce the total number of features in the final classifier. Unlike in the original Haar features, the rectangles that form an asymmetric Haar feature may have different widths or heights from each other. The asymmetry of the feature-forming rectangles significantly expands the feature space and allows the learning algorithm to select a single feature, which is approximated by the combination of the original features.

In [19], it is proposed to optimize the values of weights of rectangles forming the feature. Experimental results for problems of facial recognition on digital images and of human hearts localization in magnetic resonance images are presented. Three methods for values of the weights optimization were investigated: full search on the fixed grid, genetic algorithm, and linear Fisher discriminant. The detectors that used features with optimized rectangular weights showed significantly better accuracy and speed, and the best gain in detector quality performance was demonstrated by the optimization using the

genetic algorithm. The disadvantage of this approach is that the detector training time increase by 10 times.

The work [20] describes the concept of polygonal features. To compute polygonal feature values effectively, the paper presents a generalization of the integral image concept for image integration along polygonal integrals of a general form. Also, the paper demonstrates a way to represent the integral image as a weighted sum of precomputed right-triangular integrals.

The main disadvantage of Haar features is the instability in the case of uneven illumination of the object to be localized, i.e. when one rectangular region has illumination, which is physically different from another. Different researchers proposed various solutions to this problem. For example, in their original work [7], Viola and Jones normalize the variance of pixel intensities in the studied image area. However, there are families of features that are robust to irregularities in illumination. One such family of features is the LBP family [21], which was originally proposed as one of the ways to represent textures.

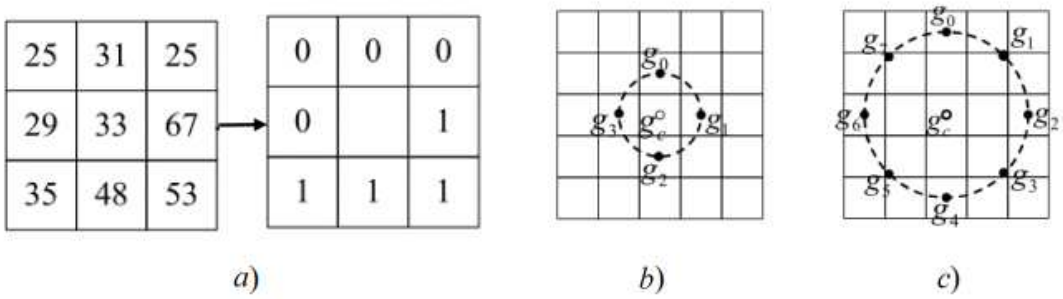

Fig. 3. An example of LBP operator calculation

The original LBP feature calculation operator works up a 3×3 neighborhood of the considered pixel. The pixel intensities values of the neighborhood are mapped to 1 if the brightness value is not less than the intensity value of the central pixel, otherwise, they are mapped to 0 (see Fig. 3). The resulting binary sequence characterizes the neighborhood of the central pixel and encodes an 8-bit number, a feature over which a classifier can be constructed. LBP-like features are used in [22–25].

Another widely used family of features resistant to the uneven illumination of objects includes features based on histograms of oriented gradients. This type of feature was proposed based on the hypothesis that some classes of objects are fully described in the form of a characteristic distribution of boundary and contour directions for this class.

Therefore, if the image area containing the object is divided into small areas, and if a histogram of directional gradients for each such area is plotted, the combination of the obtained histograms can describe the object well. Initially, this type of features was actively used for image matching and became popular due to the SIFT method [26], but later such features was used to build a pedestrian detector [27]. Unlike Haar features, histograms of Oriented Gradients (HOGs) themselves are not capable of describing the spatial structure of the object. The work [28], using the example of the cat heads recognition, proposes a new kind of features, which combines the ability of Haar features to describe local contrasts and invariant description of the boundary features distributions to illumination in the form of a histogram of oriented gradients (Haar of Oriented Gradients (HoOG)).

One of the most widely known series of works on the modification of the Viola and Jones method is the work [29] and the consequential work [30]. The latter describes a way to build a pedestrian detector, which demonstrated state-of-the-art localization quality at that time, since the method was 1-2 orders of magnitude faster than previously proposed methods. The key concept behind the claimed high quality of localization is the integral channel-wise features. The classifier takes into account features of different nature. For this purpose, the input image is subjected to various linear and nonlinear transformations (splitting into color channels, mapping to another color space, gradient norm image, Gabor filter, etc.), and the so-called map of registered transformations of the original image is formed. For each transformed image, an integral image is calculated, which enables effective calculation of such features as local sums of pixel intensities, histograms, Haar features, etc. Each feature is also associated with a weak classifier (the best quality was demonstrated by the depth-2 decision tree), and the soft cascade [31] is employed as a detection classifier. Note that the classifier is trained to take into account the mutual relations between features of different nature.

2.2. Modifications of the High-Level Classifier Structure

The original cascade structure of the classifier allows for a significant improvement of the performance of object detection and running AdaBoost on a pseudo-balanced training set. However, this architecture has several disadvantages. First, according to the cascade architecture, the information obtained at the current cascade level is not passed to the next levels in any form. Accordingly, the decision to reject or to perform further image region recognition at the current cascade level does not depend on how well the image region was recognized at the previous levels. This approach can lead to the construction of a fragile cascade, which yields rejection for small changes in the feature description of the target objects. To pass information from the previous level of the cascade, the papers [32, 33] suggest using a decision stump over the value of the linear combination of the previous level as the first weak classifier in each strong classifier. Such a high-level structure is called a Nesting-Structured cascade.

The detectors proposed in [32, 33] demonstrated better facial detection quality compared to the classical Viola and Jones cascade [7] while having significantly fewer features.

Another disadvantage of the classical cascade is the absence of an optimal learning algorithm. The final qualitative characteristics of the cascade are influenced by the training parameters of each strong classifier, which in their turn correlate with each other. The detection speed of the final cascade is another performance indicator that should be taken into account during the training phase. In [7], Viola and Jones use the average number of computed features per window on a given set of images to estimate the detection speed. Viola and Jones do not provide an exhaustive explanation on how to limit the number of features in the strong classifier. However, they provide data on the number of features they used for the first 5 levels of the final cascade. Therefore, the training of the classical cascade essentially takes place in "manual" mode: usually by adjusting the training parameters of each strong classifier.

A widely known high-level architecture of the Viola and Jones classifier is SoftCascade [31]. The image region classification algorithm combines the concepts of a

strong classifier and classical cascade. In terms of the classical cascade, it is still a tree with one terminal positive leaf and many negative leaves. In fact, SoftCascade is a large strong classifier, which is able to give a negative answer after computing another successive weak classifier. The classification rule can be represented as follows 7:

$$SoftCascade(x) = \prod_{n=1}^{N} \left[\sum_{n=1}^{N} \alpha_n \cdot h_n(x) > t_n \right], \tag{7}$$

where α_n is the weight of the classifier, $b_n(x)$ is the response of the classifier, t_n is the value of the partial sum cutoff threshold for the n-th classifier.

The SoftCascade training algorithm is divided into 2 stages: 1) training the strong classifier, which potentially contains an excessive number of features, and 2) calibration, i.e. selection of the highest quality features and threshold cutoffs of partial sums, providing the set indicators of detection completeness and performance, minimizing false positives. To avoid a significant imbalance in the distribution of positive and negative samples during the training of the strong classifier, the authors use a variant of bootstrapping and then iteratively update the negative samples set after training the next weak classifier. The calibration algorithm is based on the ROC-surface model, which, unlike the classical ROC-curve, uses the detection speed as an additional dimension.

In [34], another algorithm for SoftCascade training and calibration was introduced. This paper proposes the training algorithm which requires significantly less parameter tuning and uses weight-trimming [35] to form a negative subset in addition to bootstrapping. The authors performed a direct comparison with the original Soft Cascade [31] learning algorithm. They fixed the true positive rate training parameter, ran both of the mentioned methods, and compared the performance of the resultant detectors.

One of the important features required for industrial recognition systems is the possibility of "retraining" and tuning considering the constant expansion of the training samples set. The cascade classifier training method and the modifications of the high-level classifier structure described above are based on the assumption that the set of training samples is fixed. Since the procedure of training the classifier "from scratch" is quite time consuming (depending on the complexity of the object and the size of the training sample, it can take from several hours to several dozens of days), the task of building a high-level structure of the classifier and developing an algorithm for its training, providing a fast procedure of additional training when expanding the training set is in high demand. The work [36] proposes a tree structure and an algorithm for its construction, which allows for a relatively "cheap" retraining.

The proposed decision tree of strong classifiers is a binary decision tree. A tree node contains a strong classifier, which sends the sub-windows classified as positive to the right edge, and negative ones to the left one. The final tree verdict is given only in the leaves. Fig. 4 illustrates an example of a decision tree containing 3 nodes and 4 leaves.

2.3. Boosting Algorithm Variations for Constructing Composition of Weak Classifiers

In the original paper of Viola and Jones, the training of a strong classifier is performed using the AdaBoost algorithm [8], which was called Discrete AdaBoost later. Boosting later turned out to be not only an algorithm but a methodology for constructing a high-quality

algorithm that solves various problems by a composition of low-quality algorithms. A detailed theoretical justification and experimental results can be found in the works [37,38]. Subsequently, many "boosting algorithms" were proposed, which mainly differ by loss functions. In this paper, we consider two of the most popular boosting algorithms used for object localization: Real AdaBoost and Logit Boost.

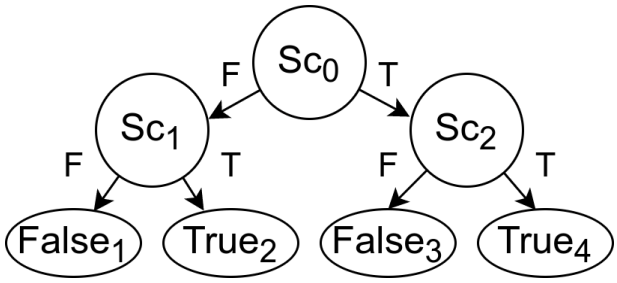

Fig. 4. Tree classifier

A large study of the family of AdaBoost algorithms, which is characterized by an exponential loss function, was presented in the work [39]. The Real AdaBoost algorithm for binary classification problems proved to be successful in the Viola–Jones framework. The Real AdaBoost algorithm represents a generalization to the case when the response of the weak classifier is a real valued number, rather than a binary one from the set $\{0, 1\}$. RealAdaBoost for the Viola and Jones method was employed in [14,15,34]. One section of the work [39] considers the case where the range of values returned by the weak classifier is discretized into a fixed number of periods. For this case, Vector Boost [40] was proposed using the rotated faces localization problem as an example. Vector Boost was also used in [18, 33].

Another popular boosting algorithm for the object localization problem is Logit Boost [35], which treats boosting as additive logistic regression. The main features of this algorithm are that the trained classifiers are optimal according to the Bayesian classification theory, and the value returned by the classifier represents a numerical estimate of the probability that the input belongs to a certain class. Moreover, the optimized logistic function is less "aggressive" than the exponential one. This means that the distribution of the weight over the training set is less subjected to potential bias to the "more complex" examples, which are often outliers (labeling mistakes). Therefore, Logit Boost is less prone to overfitting than AdaBoost in cases of the noisy training set. Logit Boost is quite actively used for object localization [29, 41–43].

2.4. Specifics of Training Cascade of Strong Classifiers

The original strong classifier training algorithm [7] adds weak classifiers until the target false negative rate FNR and false positive rate FPR for that particular strong classifier on a separate validation dataset are achieved. This approach does not consider the speed of the detector and potentially yields a slow detector unsuitable for use in real-time systems. The paper [44] proposes a greedy algorithm for searching training parameters of a strong classifier to optimize the cascade accuracy and speed. The optimization of the following functional was suggested:

$$F(E_1, E_2, D) = \beta_1 \times \log(E_1 + \delta_1) + \beta_2 \times \log(E_2 + \delta_2) + \beta_3 \times D, \tag{8}$$

where E_1 is the false negative rate, E_2 is the false positive rate, D is the average number of Haar features calculated to classify an image sub-window; the parameters β_1, β_2, β_3 are the weights of the corresponding characteristics, δ_1 and δ_2 are used for regularization. The total weight of positive training samples w and the number of weak classifiers N are the strong classifier training parameters. To search for the parameter w, a grid of weights is fixed based on the exponential nature of the optimization function:

$$w_{i+1} = \begin{cases} 1 - m \times (1 - w_i), & \text{if } w_i > 0,9, \\ w_i - \delta, & \text{if } w_i \leq 0,9, \end{cases} \tag{9}$$

which remains unchanged as strong classifiers are added to the cascade. The search for the parameter N depends on the level of the cascade since the first several levels of the cascade usually contribute the most to the value of the functional F. Therefore, a grid $N_i \in 1, 2, ..., \min(C, C_0 \times 2^i)$ is formed for the i-th strong classifier. The training is performed until the value of the characteristic E_2 on the separate validation set falls below a given threshold value. The experimental results presented in [44] indicate that the use of a validation set can significantly improve the quality of the resulting detectors. In addition, the detector produced by involving a highly experienced expert to select the training parameters for each strong classifier showed performance, which is only slightly better than the detector trained by the proposed algorithm in a fully automatic mode.

The first decade of the 21st century can be fairly called the decade of the Viola and Jones method. During this period most of its most important modifications were proposed, and its effectiveness in many different tasks of object detection was demonstrated. The Viola and Jones method de facto set the bar for qualitative characteristics and performance for other algorithms, and its individual basic elements were used for many other methods [45–47].

3. Place of the Viola and Jones Method in Era of Convolutional Neural Networks

The landmark when the Viola and Jones method began to gradually step aside was the revival of the era of neural network methods for classification problems [48]. Many researchers, impressed by the significant classification quality increase compared to previously proposed methods, began to develop convolutional neural network architectures for object localization. This was achieved in 2014, when on the well-known dataset *PASCAL VOC* [49] the purely neural network approach [50] demonstrated much better quality than many ensemble methods proposed previously. One year later, a cascade scheme of the convolutional neural network [45] was proposed for facial detection. Recall that the original method of Viola and Jones was presented on the example of detection frontal faces. Remarkably, the speed of the method is equivalent to 14 fps on a modern 2012 server processor, while the speed of the Viola and Jones method was equal to 15 fps on a consumer processor of 2000.

Therefore, 10 years after the publication of the Viola and Jones method, the vector of research on object detection methods was changed significantly. Due to the significant increase in the performance of personal computers and deep learning technologies as well as the availability of many open data sets, most research in the last decade addresses the problems of multiclass object localization. It turned out that convolutional neural networks

are flexible enough to be configured and scaled for simultaneous detection of several dozens of different object types [51–55]. However, this approach, unlike many traditional methods, including the Viola and Jones method, has some significant limitations.

The paper [56] compares modern convolutional network architectures and "traditional" algorithms, including the Viola and Jones method, in terms of the memory consumption and computational efficiency on a well-known dataset for facial detection *FDDB* [57]. The numerical characteristics of the required computational resources are presented in Table.

Table

Computational efficiency comparison
(reprinted from [56])

Detector	Time, GFLOPS,	Memory consumption, GB
Viola–Jones	0,6	0,1
HeadHunter DPM	5,0	2,0
SSD	45,8	0,7
Faster R-CNN	223,9	2,1
R-FCN 50	132,1	2,4
R-FCN 101	186,6	3,1
PVANET	40,1	2,6
Local RCNN	1206,8	2,1
Yolo 9000	34,9	2,1

According to the experimental data presented in Table 1, the Viola and Jones method shows 7 times greater efficiency in terms of the required memory and more than 8 times greater computational efficiency than the closest method based on "traditional" machine learning [58, 59]. As for the comparison with convolutional neural networks, the "closest" universal architecture in terms of computational efficiency requires almost 60 times faster performance and requires 21 times more memory. From the above comparison, it can be concluded that among the listed algorithms, the Viola and Jones method is the most preferable in cases when strict energy efficiency requirements are imposed. Therefore, the research direction for the Viola and Jones method has shifted towards solving localization problems on energy-efficient devices [60–62].

Another significant limitation of the convolutional neural network approach is the inevitable requirement for a large amount of training data. Therefore, more than 25 thousand images of faces were used to train the neural network [45]. However, there are many tasks for which such amounts of data are not available not only for tuning and training the algorithm but also for its testing. For example, the localization tasks related to identity documents [63]. In contrast to convolutional neural networks, the Viola and Jones method requires a much smaller amount of data, and with some modifications, it demonstrates the industrial quality of recognition. In [64], various Viola and Jones classifiers were trained to detect several types of identity documents in images obtained with a scanner, and the positive training set for each class consisted of several hundred examples. Pre-rectification of the document plane by a separate algorithm allowed for the

Viola and Jones method to recognize significantly projectively distorted documents [65], and in [66], the Viola and Jones classifier was trained to localize seals in Russian passports.

4. Future Directions and Discussion

In this section, we point out some areas of further development of the Viola and Jones method, which still have many unsolved problems and are of interest to modern researchers. In addition, we discuss the current place of the Viola and Jones method in the modern context of object detection problems.

4.1. Cascade Architecture Issues

In general, the cascade architecture of algorithms is motivated by performance in some "average scenario". Usually, in the case of object localization by the scanning-window method, the number of windows without objects is greater than the number of windows containing the object by orders of magnitude. The task of the cascade, in this case, is to yield the fastest "refusal" and finish the computation on the windows that do not contain an object. However, in the Viola and Jones method, cascade also solves another, equally important problem, which relates to the learning process. The unbalance of the positive and negative sub-windows is also inherent in the training phase. But due to its design, each new level of the cascade is trained by the AdaBoost algorithm not on the entire negative training set, but only on a limited subset of errors from previous levels of the cascade. And the iterative nature of the cascade structure allows for looking through the entire negative sample, the number of negative examples in which can reach tens of billions of examples.

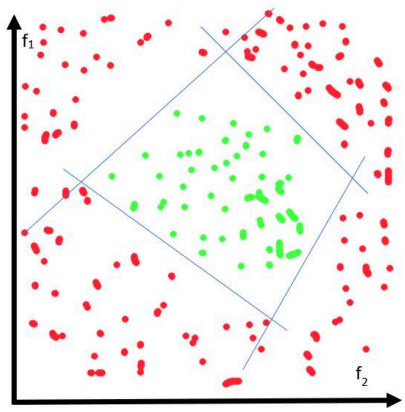

Fig. 5. An example of a set of samples not separable by a single linear classifier

Another important aspect of the cascade architecture of classifiers is their conjunctive form. Conjunctive classifiers have more expressive power than, for example, a single linear classifier. Fig. 5 illustrates an example where the feature description element of an object consists of a pair of numeric values, and positive and negative objects are highlighted in green and red, respectively. Obviously, there is no single linear classifier that splits the presented sample without errors. However, it is possible to build a cascade that includes 4 levels. Such a cascade is able to divide this sample without errors. Therefore, in addition to increasing the speed of the localization step and balancing the distribution of training samples, the cascade architecture provides more expressive power with fewer features than

a single linear classifier. This aspect is crucial when the number of computable features is limited.

Unfortunately, the scientific community does not paid sufficient attention either to the expressiveness of cascade architectures or to the learning algorithms for them. The works [31, 32, 44] consider applied problems of the computationally efficient classifier training, but, to our best knowledge, there are no works devoted to the cascade classifiers expressive ability specificity and algorithms that take the latter into account.

4.2. Stability with Respect to Different Distortion Models

Another important area of further research is the issue of robustness to various distortions. To improve the completeness of localization on prospectively distorted objects, several "engineering solutions" concerning Haar features [14,18,33] were proposed, and the task of choosing "stable" features was delegated to a learning algorithm using a training sample containing such distortions. However, the question of creating a feature space and deriving a mathematical proof of its invariance to a certain class of distortions, such as rotations, allows for significant progress in improving the quality of arbitrarily situated objects localization.

Intuitively, it is clear that the classical Haar features, which characterize the level of local contrast, are unstable in terms of local changes in object illumination. However, the question of stability of other efficiently computed features used in the Viola and Jones method to other distortions, for example, related to the specificity of the digital imaging [67], remains open.

4.3. Feature Space Issues

Even though many modifications of the feature space and a method of aggregating information from heterogeneous features in a single classifier were proposed [29, 30], the question of choosing the feature space that characterizes a certain class of objects in the best way remains open.

4.4. Energy Consumption of Computation

Another direction of the Viola and Jones method development is computational optimization for different processor architectures. In [68], an example of vectorization of two-rectangle Haar feature calculation for Elbrus VLIW design is introduced, but it can hardly be claimed that any significant part of the Viola and Jones method optimization in terms of calculations on modern processor architectures is investigated. The issues of various computational optimization for such processor architectures as ARM, x86-64, MIPS, Elbrus, etc. are still not addressed.

4.5. Scope of the Viola and Jones Method Application in Current Time

Summing up the review, the Viola and Jones method plays an important role in solving many classes of object detection problems today.

First, due to its computational efficiency, the method is highly demanded in edge computing, autonomous robotic devices, computing on smartphones, and other compact devices with low battery capacity, i.e. areas where energy efficiency plays a critical role.

Secondly, there is a huge layer of object detection tasks in which the complexity of the detected objects is low. In such cases, it is often unnecessary to train complex computational recognition models. To ensure high quality detection and recognition of such objects in practice, it is enough to apply the Viola and Jones method right out of the box, which is usually included in many image processing and analysis libraries (for example, OpenCV).

Finally, the Viola–Jones method occurs to be indispensable in object detection in multispectral images. The multispectrality of the input data significantly increases their dimensionality, which becomes an almost insurmountable obstacle for the algorithms which train feature extraction functions (filters) in an end-to-end manner, e.g. the CNN approach does. Due to the greedy nature of the AdaBoost algorithm, the algorithm trains and selects the most informative features from a set of semi-handcrafted features considering the distribution of various spectral characteristics of the detected class of objects.

Conclusion

This paper reviewed the most significant studies of the Viola and Jones method over the past 20 years since its publication. The Viola and Jones method [7] undoubtedly revolutionized the field of pattern recognition: for the first time, a high-quality method of complex objects localization was introduced, which on a widely available general-purpose central processor demonstrated the performance close to the frame rate of a video camera. For a long time, the concepts described in the original paper, the construction of such detectors captured a large part of the scientific community and were independently researched and developed. However, with the increase in computing power of personal computers and the development of deep learning methods, in particular, convolutional neural networks, the research vector and area of applicability of the Viola and Jones method shifted towards power-efficient devices and embedded systems, and towards tasks for which it is not possible to collect representative data in the quantities required for deep learning methods. Important future research also lies within the area of cascade architecture of classifiers, stability of the method with respect to different distortion models, and choice of feature space that best characterizes the objects to be localized.

Acknowledgments. *This work was partially financially supported by the Russian Foundation for Basic Research, project no. 20-17-50223.*

References

1. Henderson C. *Driving Crime Down: Denying Criminals the Use of the Road.* Available at: https://popcenter.asu.edu/sites/default/files/Henderson.pdf (accessed 21 July 2021)
2. *China's Watchful Eye* (2021). Available at: https://www.washingtonpost.com/news/world/wp/2018/01/07/feature/in-china-facial-recognition-is-sharp-end-of-a-drive-for-total-surveillance/ (accessed 21 July 2021)
3. Du S. et al. Automatic License Plate Recognition: A state-of-the-Art Review. *IEEE Transactions on Circuits and Systems for Video Technology*, 2012, vol. 23, no. 2, pp. 311–325.
4. *Law Enforcement's Use of Facial Recognition Technology.* Available at: https://www.fbi.gov/news/testimony/law-enforcements-use-of-facial-recognition-technology (accessed 21 July 2021)

5. Sung K.K., Poggio T. Example-Based Learning for View-Based Human Face Detection. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 1998, vol. 20, no. 1, pp. 39–51.
6. Schneiderman H., Kanade T. A Statistical Method for 3D Object Detection Applied to Faces and Cars. *Proceedings IEEE Conference on Computer Vision and Pattern Recognition*, 2000, no. 1, pp. 746–751.
7. Viola P., Jones M. Robust Real-Time Object Detection. *International Journal of Computer Vision*, 2001, no. 4, pp. 34–47.
8. Freund Y., Schapire R.E. A Decision-Theoretic Generalization of On-Line Learning and an Application to Boosting. *Journal of Computer and System Sciences*, 1997, vol. 55, no. 1, pp. 119–139.
9. Papageorgiou C.P., Oren M., Poggio T. A General Framework for Object Detection. *Sixth International Conference on Computer Vision*, 1998, pp. 555–562.
10. Lienhart R., Maydt J. An Extended Set of Haar-Like Features for Rapid Object Detection. *Proceedings of International Conference on Image Processing*, 2002, vol. 1, pp. I–I.
11. Huang C.C., Tsai C.Y., Yang H.C. An Extended Set of Haar-Like Features for Bird Detection Based on AdaBoost. *International Conference on Signal Processing, Image Processing, and Pattern Recognition*, 2011, pp. 160–169.
12. Wen X. et al. A Rapid Learning Algorithm for Vehicle Classification. *Information Sciences*, 2015, no. 295, pp. 395–406.
13. Gaszczak A., Breckon T.P., Han J. Real-Time People and Vehicle Detection from UAV Imagery. *Intelligent Robots and Computer Vision XXVIII: Algorithms and Techniques*, 2011, no. 7878, pp. 78780B.
14. Li S.Z. et al. Statistical Learning of Multi-View Face Detection. *European Conference on Computer Vision*, 2002, pp. 67–81.
15. Jones M., Viola P. Fast Multi-View Face Detection. *Mitsubishi Electric Research Lab TR-20003-96*, 2003, vol. 3, no. 14, pp. 2.
16. Viola P., Jones M.J., Snow D. Detecting Pedestrians Using Patterns of Motion and Appearance. *International Journal of Computer Vision*, 2005, vol. 63, no. 2, pp. 153–161.
17. Messom C., Barczak A. Fast and Efficient Rotated Haar-Like Features Using Rotated Integral Images. *Australian Conference on Robotics and Automation*, 2006, pp. 1–6.
18. Ramirez G. A., Fuentes O. Multi-Pose Face Detection with Asymmetric Haar Features. *2008 IEEE Workshop on Applications of Computer Vision*, 2008, pp. 1–6.
19. Pavani S.K., Delgado D., Frangi A.F. Haar-Like Features with Optimally Weighted Rectangles for Rapid Object Detection. *Pattern Recognition*, 2010, vol. 43, no. 1, pp. 160–172.
20. Pham M.T. et al. Fast Polygonal Integration and Its Application in Extending Haar-Like Features to Improve Object Detection. *2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, 2010, pp. 942–949.
21. Ojala T., Pietikainen M., Harwood D. A Comparative Study of Texture Measures with Classification Based on Featured Distributions. *Pattern Recognition*, 1996, vol. 29. no. 1, pp. 51–59.
22. Zhang L. et al. Face Detection Based on Multi-Block Lbp Representation. *International Conference on Biometrics*, 2007, pp. 11–18.
23. Nikisins O., Greitans M. Local Binary Patterns and Neural Network Based Technique for Robust Face Detection and Localization. *2012 BIOSIG-Proceedings of the International Conference of Biometrics Special Interest Group (BIOSIG)*, 2012, pp. 1–6.

24. Suri P.K., Verma E.A. Robust Face Detection Using Circular Multi Block Local Binary Pattern and Integral Haar Features. *International Journal of Advanced Computer Science and Applications, Special Issue on Artificial Intelligence*, 2011, pp. 67–71.

25. Jammoussi A.Y., Masmoudi D.S. Joint Integral Histogram Based Adaboost for Face Detection System. *International Journal of Computer Applications*, 2011, vol. 23, no. 5.

26. Lowe D.G. Object Recognition from Local Scale-Invariant Features. *Proceedings of the Seventh IEEE International Conference on Computer Vision*, 1999, no. 2, pp. 1150–1157.

27. Dalal N., Triggs B. Histograms of Oriented Gradients for Human Detection. *2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, 2005, no. 1, pp. 886–893.

28. Zhang W., Sun J., Tang X. Cat Head Detection-How to Effectively Exploit Shape and Texture Features. *European Conference on Computer Vision*, 2008, pp. 802–816.

29. Dollar P. et al. Integral Channel Features. *Proceedings of the British Machine Vision Conference*, 2009, pp. 91.1–91.11.

30. Dollar P., Belongie S., Perona P. The Fastest Pedestrian Detector in the West. *Proceedings of the British Machine Vision Conference*, 2010, pp. 68.1–68.11.

31. Bourdev L., Brandt J. Robust Object Detection Via Soft Cascade. *2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05)*, 2005, no. 2, pp. 236–243.

32. Xiao R., Zhu L., Zhang H. J. Boosting Chain Learning for Object Detection. *Proceedings Ninth IEEE International Conference on Computer Vision*, 2003, pp. 709–715.

33. Wu B. et al. Fast Rotation Invariant Multi-View Face Detection Based on Real AdaBoost. *Sixth IEEE International Conference on Automatic Face and Gesture Recognition*, 2004, pp. 79–84.

34. Zhang C., Viola P. Multiple-Instance Pruning for Learning Efficient Cascade Detectors. *Advances in Neural Information Processing Systems*, 2007, no. 20, pp. 1681–1688.

35. Friedman J., Hastie T., Tibshirani R. Additive Logistic Regression: a Statistical View of Boosting (with Discussion and a Rejoinder by the authors). *The Annals of Statistics*, 2000, vol. 28, no. 2, pp. 337–407.

36. Minkina A. et al. Generalization of the Viola–Jones Method as a Decision Tree of Strong Classifiers for Real-Time Object Recognition in Video Stream. *Seventh International Conference on Machine Vision*, 2015, no. 9445, pp. 944517. DOI: 10.1117/12.2180941

37. Mason L. et al. Boosting Algorithms as Gradient Descent in Function Space. *Advances in Neural Information Processing Systems*, 1999, no. 12, pp. 512–518.

38. Friedman J.H. Greedy Function Approximation: a Gradient Boosting Machine. *Annals of Statistics*, 2001, pp. 1189–1232.

39. Schapire R.E., Singer Y. Improved Boosting Algorithms Using Confidence-Rated Predictions. *Machine Learning*, 1999, vol. 37, no. 3, pp. 297–336.

40. Huang C. et al. Vector Boosting for Rotation Invariant Multi-View Face Detection. *Tenth IEEE International Conference on Computer Vision*, 2005, no. 1, pp. 446–453.

41. Duan S., Wang X., Wan W. The Logitboost Based on Joint Feature for Face Detection. *2013 Seventh International Conference on Image and Graphics*, 2013, pp. 483–488.

42. Gualdi G., Prati A., Cucchiara R. Multi-Stage Sampling with Boosting Cascades for Pedestrian Detection in Images and Videos. *European Conference on Computer Vision*, 2010, pp. 196–209.

43. Wang L., Zhang Z. Automatic Detection of Wind Turbine Blade Surface Cracks Based on UAV-Taken Images. *IEEE Transactions on Industrial Electronics*, 2017, vol. 64, no. 9, pp. 7293–7303.

44. Poljakov I.V. et al. [Training Optimal Viola–Jones Detectors Using Greedy Algorithms for Selecting Control Parameters with Intermediate Validation on Each Level]. *Sensory Systems*, 2016, vol. 30, no. 3, pp. 241–248. (in Russian)

45. Li H. et al. A Convolutional Neural Network Cascade for Face Detection. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2015, pp. 5325–5334.

46. Owusu E., Abdulai J. D., Zhan Y. Face Detection Based on Multilayer Feed-Forward Neural Network and Haar Features. *Software: Practice and Experience*, 2019, vol. 49, no. 1, pp. 120–129.

47. Cai Z., Vasconcelos N. Cascade r-cnn: Delving into High Quality Object Detection. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 6154–6162.

48. Krizhevsky A., Sutskever I., Hinton G. E. Imagenet Classification with Deep Convolutional Neural Networks. *Advances in Neural Information Processing Systems*, 2012, no. 25, pp. 1097–1105.

49. Everingham M. et al. The Pascal Visual Object Classes (VOC) Challenge. *International Journal of Computer Vision*, 2010, vol. 88, no. 2, pp. 303–338.

50. Girshick R. et al. Rich Feature Hierarchies for Accurate Object Detection and Semantic Segmentation. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2014, pp. 580–587.

51. Redmon J. et al. You Only Look Once: Unified, Real-Time Object Detection. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2016, pp. 779–788.

52. Liu W. et al. SSD: Single Shot Multibox Detector. *European Conference on Computer Vision*, 2016, pp. 21–37.

53. Lin T.Y. et al. Feature Pyramid Networks for Object Detection. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017, pp. 2117–2125.

54. Jiao L. et al. A Survey of Deep Learning-Based Object Detection. *IEEE Access*, 2019, vol. 7, pp. 128837–128868.

55. Zou Z. et al. Object Detection in 20 Years: A Survey. Available at: https://arxiv.org/abs/1905.05055 (accessed 21 July 2021)

56. Granger E. et al. A Comparison of Cnn-Based Face and Head Detectors for Real-Time Video Surveillance Applications. *2017 Seventh International Conference on Image Processing Theory, Tools and Applications*, 2017, pp. 1–7.

57. Jain V., Learned-Miller E. A Benchmark for Face Detection in Unconstrained Settings. *UMass Amherst Technical Report*, 2010, vol. 2, no. 4, pp. 5.

58. Yan J. et al. Real-Time High Performance Deformable Model for Face Detection in the Wild. *2013 International Conference on Biometrics*, 2013, pp. 1–6.

59. Felzenszwalb P., McAllester D., Ramanan D. A Discriminatively Trained, Multiscale, Deformable Part Model. *2008 IEEE Conference on Computer Vision and Pattern Recognition*, 2008, pp. 1–8.

60. Usilin S.A., Slavin O.A., Arlazarov V.V. Memory Consumption and Computation Efficiency Improvements of Viola–Jones Object Detection Method for Remote Sensing Applications. *Pattern Recognition and Image Analysis*, 2021, vol. 31, no. 3, pp. 571–579. DOI: 10.1007/978-3-030-68821-9_23

61. Xu Y. et al. A hybrid Vehicle Detection Method Based on Viola–Jones and HOG+SVM from UAV Images. *Sensors*, 2016, vol. 16, no. 8, p. 1325.
62. Irgens P. et al. An Efficient and Cost Effective Fpga Based Implementation of the Viola–Jones Face Detection Algorithm. *HardwareX*, 2017, no. 1, pp. 68–75.
63. Skoryukina N., Arlazarov V., Nikolaev D. Fast Method of ID Documents Location and Type Identification for Mobile and Server Application. *2019 International Conference on Document Analysis and Recognition*, 2019, pp. 850–857. DOI: 10.1109/ICDAR.2019.00141
64. Usilin S. et al. Visual Appearance Based Document Image Classification. *2010 IEEE International Conference on Image Processing*, 2010, pp. 2133–2136.
65. Tropin D.V. et al. Localization of Planar Objects on the Images with Complex Structure of Projective Distortion. *Informatsionnye Protsessy*, 2019, vol. 19, no. 2, pp. 208–229. (in Russian)
66. Matalov D.P., Usilin S.A., Arlazarov V.V. Modification of the Viola–Jones Approach for the Detection of the Government Seal Stamp of the Russian Federation. *Eleventh International Conference on Machine Vision*, 2019, vol. 11041, p. 110411Y. DOI: 10.1117/12.2522793
67. Polevoy D. et al. Key Aspects of Document Recognition Using Small Digital Cameras. *RFBR Journal*, 2016, vol. 4, no. 92, pp. 98–105. doi: 10.22204/2410-4639-2016-092-04-97-108 (in Russian)
68. Limonova E.E. et al. [Recognition System Efficiency Evaluation on VLIW Architecture on the Example of Elbrus Platform]. *Programming and Computer Software*, 2019, vol. 45, no. 1, pp. 15–21. DOI: 10.1134/S0132347419010047 (in Russian)

Received August 3, 2021

УДК 004.021 **DOI: 10.14529/mmp210401**

ЭВОЛЮЦИЯ МЕТОДА ВИОЛЫ – ДЖОНСА: ОБЗОР

В.В. Арлазаров[1,2,3,4]***, Ю.С. Войсят***[1]***, Д.П. Маталов***[1,3]***, Д.П. Николаев***[1,2,4]***, С.А. Усилин***[1,2,3]

[1]ООО «Смарт Энджинс Сервис», г. Москва, Российская Федерация
[2]Московский физико-технический институт, г. Москва, Российская Федерация
[3]Федеральный исследовательский центр «Информатика и управление» РАН, г. Москва, Российская Федерация
[4]Институт проблем передачи информации имени А.А. Харкевича РАН, г. Москва, Российская Федерация
E-mails: vva@smartengines.com, u.voisiat@smartengines.com, d.matalov@smartengines.com, d.p.nikolaev@smartengines.com, usilin@smartengines.com

Метод Виолы и Джонса является одним из самых известных методов локализации объектов на цифровых изображениях. За минувшие 20 лет со дня первой публикации метод был существенно изучен, исследователями и инженерами было предложено множество модификаций оригинального алгоритма и отдельных его частей. Отдельные популяризованные Полом Виолой и Майклом Джонсом идеи встали в основу множества других алгоритмов локализации объектов на изображениях. В этой работе представлено описание метода Виолы и Джонса, история его развития и модификаций в контексте различных задач локализации объектов на изображениях, а также описание современного состояния дел – какое место метод занимает сейчас, в эпоху обширного применения сверточных нейронных сетей.

Ключевые слова: метод Виолы и Джонса; распознавание образов; машинное обучение; классификация объектов; локализация объектов; детектирование объектов.

Работа проводилась при частичной финансовой поддержке РФФИ, проект № 20-17-50223

Владимир Викторович Арлазаров, кандидат технических наук, генеральный директор, ООО «Смарт Энджинс Сервис» (г. Москва, Российская Федерация); заведующий отделом, Федеральный исследовательский центр «Информатика и управление» РАН (г. Москва, Российская Федерация); и.о. ведущего научного сотрудника, Институт проблем передачи информации имени А.А. Харкевича РАН (г. Москва, Российская Федерация); преподаватель, Московский физико-технический институт (г. Москва, Российская Федерация), vva@smartengines.com.

Юлия Сергеевна Войсят, лаборант-программист ООО «Смарт Энджинс Сервис» (г. Москва, Российская Федерация), u.voisiat@smartengines.com.

Даниил Павлович Маталов, аспирант, ведущий программист, Федеральный исследовательский центр «Информатика и управление» РАН (г. Москва, Российская Федерация); научный сотрудник-программист ООО «Смарт Энджинс Сервис» (г. Москва, Российская Федерация), d.matalov@smartengines.com.

Дмитрий Петрович Николаев, кандидат технических наук, технический директор, ООО «Смарт Энджинс Сервис» (г. Москва, Российская Федерация); Заведующий лабораторией № 11 «Зрительные системы», Институт проблем передачи информации имени А.А. Харкевича РАН (г. Москва, Российская Федерация); преподаватель, Московский физико-технический институт (г. Москва, Российская Федерация), dimonstr@iitp.ru.

Сергей Александрович Усилин, кандидат технических наук, исполнительный директор, ООО «Смарт Энджинс Сервис» (г. Москва, Российская Федерация); старший научный сотрудник, Федеральный исследовательский центр «Информатика и управление» РАН (г. Москва, Российская Федерация); преподаватель, Московский физико-технический институт (г. Москва, Российская Федерация), usilin@smartengines.com.

Поступила в редакцию 3 августа 2021 г.