Developing operation algorithms for vision subsystems in autonomous mobile robots

M V Shikhman[1,2] and S V Shidlovskiy[2]

[1]National Research Tomsk Polytechnic University, Tomsk, Russia

[2] National Research Tomsk State University, Tomsk, Russia

E-mail: Sikhmar@gmail.com

Abstract. The paper analyzes algorithms for selecting keypoints on the image for the subsequent automatic detection of people and obstacles. The algorithm is based on the histogram of oriented gradients and the support vector method. The combination of these methods allows successful selection of dynamic and static objects. The algorithm can be applied in various autonomous mobile robots.

1. Introduction

Currently, robotic systems are rightly considered as the base for automation of industry, medicine, military industry, space exploration and other spheres of human activity. Their joint work allows achieving various goals and solving a wide range of technological problems.

An autonomous mobile object is a kind of robotic system, which is based on an autonomous control system provided by optical sensors, computer algorithms and radar. An autonomous mobile object can be used to move passengers or cargo, to deliver materials, technological or other types of equipment, and to collect visual information when working with additional equipment. The main two tasks for any autonomous object are as follows. First, to be able to plan the path independently; second, to be able to move around the dynamic environment successfully, including the ability to move among other moving objects. Development of satellite navigation and electronic maps facilitated successful solution of the first task. The second task is much more difficult; many scientists around the world are trying to solve it using different methods and algorithms. The use of original path planning algorithms makes it easy to ensure the safety of movement in an area with dynamic obstacles.

2. Distribution of keypoints on the image

Any image represents complex and difficult-to-extract structured information about the observed scene. Therefore, we need a method that will allow extracting information from the real-time video data stream, identifying and recognizing objects using this information.

The main problem during the process of object recognition is to compare the image obtained from the camera with the etalon sample stored in the database. This problem can be solved using a number of tools allowing image recognition and matching them to databases. However, the main method is to establish a correspondence between the keypoints on the initial image and the etalon image.

A keypoint is the simplest geometric element of the discrete representation of the mathematical function for describing the object recognition. To define these points, we introduce a concept of a neighborhood. That is, we consider a point p_i to be a key (reference) point for some image; the neighborhood $O(p_i)$ of this point can be distinguished from the neighborhood $O'(p_i)$ of any other keypoint of the image p_i. The process of detecting the given point is called *detection*, and the program that carries out this process is called *a detector*. After this process, it is necessary to describe this particular point; and descriptors are responsible for this operation. A descriptor is a description of a keypoint that determines the characteristics of its neighborhood; it represents a numerical or binary vector of certain parameters [1].

There are many different methods for selecting keypoints and descriptors, but in this paper, we will use the histogram of oriented gradients (HOG). This choice is associated with a number of advantages of the HOG descriptor over others. First, HOG operates locally; it allows maintaining invariance in relation to geometric and photometric transformations of the object on small fragments of the image, but orientation of the object is an exception here. Second, a clear space partition, accurate calculation of directions, and strong local photometric normalization allow to not considering the movement of people if they are in vertical position. Therefore, this detector is a good tool for determining pedestrians on images.

3. Histogram of oriented gradients

The histogram of oriented gradients (HOG) means descriptors of keypoints used in computer vision and image processing systems for object recognition. This descriptor is based on the method of counting the number of gradient directions in local areas of the image.

The basis of this method is the assumption that an object description in the image can be achieved by specifying the distribution of intensity gradients or edge direction. The method is realized by partitioning the image into elementary regions (cells) and further calculating the histograms of the gradient directions or the edge directions for pixels for each region. The collection of histogram data for all cells is a descriptor [2].

Besides, the image is normalized in contrast; this is necessary in order to increase the accuracy. For this purpose, the intensity measure is determined for a large area of the image (block), the resulting value is used for normalization. This allows increasing the invariance of the descriptor in relation to illumination.

The steps of the HOG algorithm realization include the following: calculating the gradient, forming cell histograms, forming descriptor blocks, normalizing blocks, and classifying descriptors [3, 4]. The last step is the most difficult because it is based on machine learning. To implement it, the support vector method can be used. This method allows a binary classification, that is, a division into two classes: 1) the object belongs to the required category or 2) does not belong.

4. Supervised learning

The essence of the supervised learning is as follows. Let us suppose that we have a task to determine the membership of an object on the image. The decision-making process will be like a “black box” because we do not know how it works [5]. The task is to get a result or, in other words, to decide if an object belongs to a given group.

Let us consider the given process by an example of image recognition when it is necessary to define whether the given object is a person or not. In this case, we ourselves form the so-called “learning set”, that is a set of examples and correct answers for them. We denote a set of examples as *X*, and a set of solutions as *Y*. Therefore, to solve our problem, it is necessary to introduce a function *f(X)*, which will allow transforming the set *X* into the set *Y* (Figure 1).

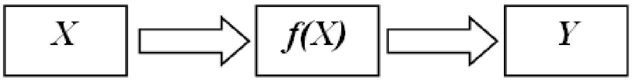

Figure 1. Function extraction *f(X).*

These pairs of sets can be called as “stimulus-response” or precedent, and the aggregate of sets can be called as a training sample. The supervised learning is also often called as the learning by precedents [6].

Thus, the use of the function found and the software allows solving an example task, which was not presented in the training sample.

5. The support vector method

As we have already mentioned, the last step in the process of object detection on an image is the descriptor classification, which is implemented using the support vector method, which, in turn, is based on supervised learning.

The support vector method (SVM) is a set of algorithms for solving classification problems based on supervised learning; this method belongs to a family of linear classifiers. Its main advantage is in constant decrease in the empirical classification error and increase in the gap, that is, a more confident work of a classifier. Therefore, the method is also called as maximum margin classifier [7–9].

The main idea of the method is to interpret the initial vectors into a higher dimension space, and to allocate a hyperplane with the maximum gap in a given space.

5.1 Description of the support vector method using the optimal separating hyperplane

Let there be a linear separable sample, that is, a sample that can be divided into two classes (the case of binary classification). We divide this sample using a linear hyperplane; however, several planes may exist (Figure 2). In this case, each plane will divide the sample, but changes in the plane coefficients will be observed. Thus, here is a question. Which of these hyperplanes will be optimal?

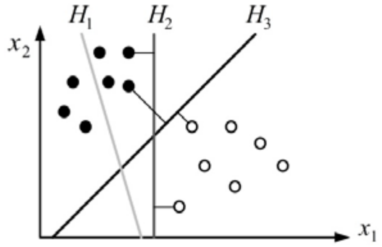

Figure 2. Construction of a separating hyperplane.

The optimal separating hyperplane will be the one that allows fulfilling the following condition: the distance between two nearest points lying on the opposite sides of the hyperplane (that is, between points belonging to different classes) should be maximum. The corresponding classifier is called as the optimal separating classifier.

To find the optimal hyperplane, we consider the formal construction, which is obtained as a result of the maximum shift of the separating hyperplane to to one or the other classes (Figure 3).

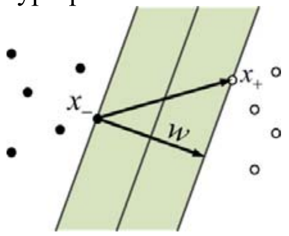

Figure 3. Construction of a separating hyperplane.

In this case, there is a question. What values will the scalar product of the weight vector *w* and *x* have taking into account the shear coefficient at the extreme points of each class. These extreme points belong to one or the other classes simultaneously, and the boundaries of the separating line pass through them.

$$a(x) = sign\left(\sum_{i=1}^{n} w_i x_i - w_0\right) = sign\left(\langle w,x\rangle - w_0\right), \langle w,x\rangle - w_0$$

where w_0 is a shear coefficient (the probability distribution parameter, it allows establishing a relationship between the value of this parameter and the choice of the reference point of the scale measurement); *sign* is a function that allows determining the sign of a number or a mathematical / trigonometric function.

If we multiply the vector w and the coefficient w_0 by some arbitrary number, then the separating surface will not change at all. The sign of the expression will not change because the function will only be multiplied by some constant. In this case, we can require that the values of this function on the extreme objects modulo equal to unity

$$\min_{i=1,\ldots,l} y_i\left(\langle w, x_i\rangle - w_0\right) = 1.$$

By requiring such normalization, it is possible to calculate the width of the separating strip. To do this, we take the direction that defines the normal to the separating surface, that is, the vector w, and consider the projection of the vector of extreme element difference of each class to this direction. In order to find this direction, it is necessary to take the scalar product of the *«x+»* and *«x-»* difference and the vector w normalized to its length. If we write this operation in details, we obtain the difference of scalar products of vectors:

$$\left\langle \left(x_+ - x_-\right), \frac{w}{\|w\|} \right\rangle = \frac{\langle w, x_+\rangle - \langle w, x_+\rangle}{\|w\|}.$$

At the same time, we should not forget that a certain relationship resulting from the normalization occurs to these extreme objects. If we use it, we get the following:

$$\left\langle \left(x_+ - x_-\right), \frac{w}{\|w\|} \right\rangle = \frac{\langle w, x_+\rangle - \langle w, x_+\rangle}{\|w\|} = \frac{(w_0+1)-(w_0+1)}{\|w\|} = \frac{2}{\|w\|}$$

As a result, the width of the separating strip will be equal to $2/\|w\|$. Now we can put down a task that will describe the maximization of the width of the separating strip. To do this, we minimize the scalar product of two vectors w taking into account the fact that the vector w length is the root of this value. We also should take into account that the indention on this object must be greater or equal to 1.

$$\begin{cases} \langle w, w\rangle \to \min \\ y_i\left(\langle w, x_i\rangle - w_0\right) \geq 1,\ i = 1,\ldots l \end{cases}$$

5.2 Transition to linear inseparable sample

Let us consider the case of a linear inseparable sample. In the previous case, we could require that the indention was greater or equal to 1, where 1 was a positive value, and the indention was positive when the classification was correct. Now, in the case of a linear inseparable sample, there will be objects that are incorrectly assigned to the class by this classifier. This means that it is necessary to allow this algorithm to be mistaken, that is to allow making indention, which will not be greater or equal to 1, but will be greater than or equal to 1 minus ξ_i, where ξ_i is the error on the ith object. In this case, it is necessary to add "fines" for these errors to the function being minimized because in case of their absence, it is possible to make a classifier with any arbitrary indention

$$\begin{cases} \frac{1}{2}\langle w, w\rangle + C\sum_{i=1}^{l}\xi_i \to \min\limits_{w, w_0, \xi} \\ y_i\left(\langle w, x_i\rangle - w_0\right) \geq 1 - \xi_i,\ i = 1,\ldots,l \\ \xi_i \geq 0,\ i = 1,\ldots,l \end{cases}$$

Thus, it is possible to obtain an optimization problem for the support vector method. We need to introduce a link to a linear classifier. First, it is necessary to minimize the sum ξ_i, with $\xi_i \geq 0$ and $\xi_i \geq$ 1- M_i, where M_i is an indention. This means that the error ξ_i will be greater or equal to the maximum of the two values - 0 and 1 minus indention. On the other hand, because the sum ξ_i is minimized, then ξ_i is exactly equal to this value, that is, the maximum of 0 and 1 minus M_i

$$\begin{aligned} &\xi_i \geq 0 \\ &\xi_i \geq 1 - M_i \quad \Rightarrow \xi_i = \max\{0, 1 - M_i\} = (1 - M_i)_+ \\ &\sum_{i=1}^{l} \xi_i \to \min \end{aligned}$$

In this case, you can simply substitute these ξ_i for the optimization task, namely for the first expression; then we will get an unconditional optimization problem in SVM. That is, an optimization task without additional conditions.

$$Q(w, w_0) = \sum_{i=1}^{l} \left(1 - M_i\left(w, w_0\right)\right)_+ + \frac{1}{2}\|w\|^2 \to \min_{w, w_0}$$

Here, we can clearly see the loss function, which is piecewise linear, and the regularizer, which is a regular L2-regularizer.

Thus, the support vector method is a linear classifier with a piecewise linear loss function (hinge loss) and an L2-regularizer. This method is necessary to maximize the gap between classes. In the case of a linear separable sample, this means simply maximizing the width of the separating strip. In the case of a linearly inseparable sample, the possibility of objects' hitting into the strip and “fines” for this hitting is simply added.

6. Conclusion

This algorithm, which includes a set of HOG descriptors and the support vector method, allows not only recognizing objects on a static image but also extracting them from the video data stream. The algorithm is optimal for recognizing people by means of an autonomous mobile object; it can also be used for recognizing other moving and static objects.

Acknowledgments

The research was carried out in Tomsk State University with the financial support of the Ministry of Education and Science of Russia, a unique project identifier: RFMEFI57817X0241.

The authors are grateful to Tatiana B. Rumyantseva from Tomsk State University for her assistance in preparing the paper.

References

[1] Finogeev A, Chetvertova M 2017 *Molodoj uchenyj* **11** 214–121

[2] Dalal N, Triggs B 2015 *Telecommunication Systems* **60** 337–339

[3] Shashev D, Shidlovsky S 2017 *Journal of Physics: Conference Series* **881(1)** 012029

[4] Hila M, Yannawar P, Gaikwad A 2017 *Proceedings - 1st International Conference on Intelligent Systems and Information Management, ICISIM 2017* 22–25

[5] Panin S V, Shakirov I V, Syryamkin V I, Svetlakov A A 2003 *Avtometriya (1)* 37–53

[6] Vyugin V 2014 *Matematicheskie osnovy mashinnogo obucheniya i prognozirovaniya* (Moscow, MCMNO)

[7] Gevan N, Ivanov V 2012 *Informacionnye tekhnologii i problemy matematicheskogo modelirovaniya slozhnyh sistem* **10** 81–84

[8] Wang Z, Qiao H 2017 *Proceedings of 2017 IEEE 7th International Conference on Electronics Information and Emergency Communication, ICEIEC 2017* 572–575

[9] Yurchenko A V *et al* 2015 *IOP Conf. Ser.: Mater. Sci. Eng.* **81** 012112