One Approach to intellectual image analysis

Nikolai Bellustin[1], *Konstantin* Moiseev[2], *Olga* Shemagina[3], *Sergey* Starkov[2*] and *Aleksandr* Telnykh[3]

[1]Radiophisical Research Institute, Nizhny Novgorod, Russia
[2]National Research Nuclear University MEPhI, Obninsk, Russia
[3]Institute of Applied Physics, Russian Academy of Sciences, Nizhny Novgorod, Russia

Abstract. This study investigated the method of semantic image analysis by using a set of neuron-like detectors of foreground objects. This method is intended to find different types of foreground objects and to determine properties of these objects. As a result of semantic analysis the semantic descriptor of the image is created. The descriptor is a set of foreground objects of the image and a set of properties for each object. The distance between images is defined as distance between their semantic descriptors. Using the concept of distance between images, "semantically similarity" between images or videos is defined.

1 Statement of the problem

Fast development of computers, Internet and robotics technologies last years leads to significant increase of interest in image understanding systems. Here we discuss the systems of understanding images which are presented by the array of numbers and contain the description of the scene in "non-visual" form. The type of the scene description depends on the specific task. In some cases, it is desirable to have the description as an indication of the presence or absence of the specified object in the scene, or the appearance of an unexpected object, for example, this approach is useful for use in security systems. Or perhaps it is necessary to track the movement and direction of some object, this task can be used in surveillance systems. In more complex cases, such a system could create a general description of the scene, which contains the main elements of the scene and the list of objects in it. Modern researches focus on the development of such programmable systems that could process a wide class of images arising in a wide variety of applications, including compression, storage and video analysis. There are already created systems, which perform relatively simple images description for limited class of images [1-2].

The image analysis process is carried out in several stages. In the first phase in the observed image the simplest features are extracted, the features may be natural in the sense that they are established by visual analysis of the image, while other artificial features are obtained as a result of its special processing or measurements. Natural features include such features as brightness of the image elements, the coordinates of contour points, form objects, etc. To highlight these characteristics in the different methods of contour extraction

* Corresponding author: sergeystarkov56@mail.ru

are used, making two level images (binarization), or N-level images [1-2]. The artificial features are histograms of brightness distribution, spectra of spatial frequencies, the results of the processing of Haar cascades [3]. Then this feature set goes to the block symbolic representation, which forms the characters of the features. For example, contour points are grouped into line segments or closed curves, elements with the same brightness etc. These objects description already has semantic representation.

A coherent semantic description of the object is formed by using these semantic characteristics, and it allows making the categorization and identification of detected objects. The information obtained is stored in a general database and can be used for further search. The comparison of proximity between the images is already made on the basis of these semantic descriptions.

But this approach gives a number of issues making the procedure very difficult and sometimes essentially impossible. There are changes in brightness, scale, angle of view and the registered image angle. Serious limitations also appear in case of registered movement of the camera relative to the scene objects. In this case, the so-called semantic gap – the absence of adequate matching between the low-level descriptive features of the graphical object (color, brightness, relative positions and sizes of fragments, etc.) and its semantic description. Schematically this is presented in Figure 1. [2012 Hewlett-Packard Development Company, L. P.].

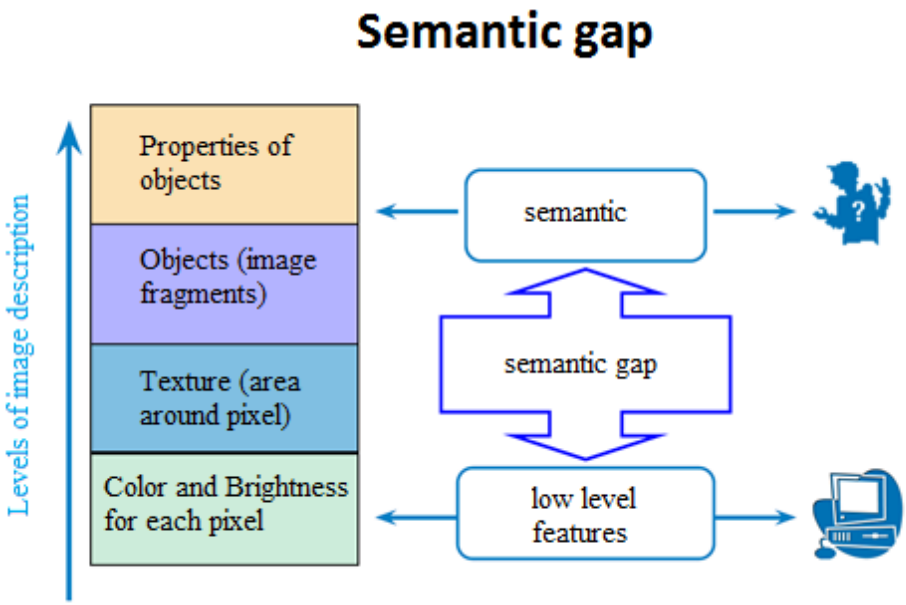

Fig. 1. The semantic gap the analysis of the graphic images.

2 Key idea

In this paper we consider one possible approach to constructing semantic descriptions of registered scenes image, for its subsequent use for effective analysis of graphic images and video scenes.

2.1 Video data signatures

The basic idea is to use the totality of the intellectual video detection, which allows to carry out detection and identification of particular classes of objects at the stage of primary video recording. An example of such a video detector, for instance, may be a video face detector, widely used currently in a various applications: smartphones, authentication system, etc. If we assume that we have setup VMD under a unique object (a person, object, figure, angle, shape/s, action, moving, etc.), the descriptors of individual objects, their relative positions would accurately reflect the specific scene.

Often, to help you find the desired graphic information in large databases of images and videos, formal external description of the data, called metadata are used. Signatures of images and videos in a sense can also be considered as metadata, as are formal appearance, relative to the source image information description. Video data signatures, in the form of metadata, depend only on the algorithm for generating signatures and on the data.

In this work, we propose to overcome the semantic gap between mathematical and semantic description of video data using the technique of "semantic analysis of video information".

In the application to static images the semantic analysis consists of the designation of certain areas or the whole mage by units of a natural language, and in the application to video – dynamic scenes marking by units of natural language.

A tool that is able to automatically designate a certain area of the image is a detector of objects of the specified type and tracker, which provides support for found objects from frame to frame [3, 4]. The nature of the detector can be different - neural network, cascade connection, based on the color or the contour analysis, texture analysis and so on. The same applies to the nature of the tracker. Important property of the detector is the space localization of object in the image with which it is associated, and for the tracker the main property is the localization of the object in time. The detector indicates the presence of something described by the noun with which it is associated.

Significant expansion of functionality of the proposed system gives us possibilities for analysis of localized detector regions property.

The properties of image area found by the detector is represented by the detector attributes [5], which describes some of the field characteristics by an adjective. Thus, we have a relationship between the noun, its properties, and the detector, the detector attributes and the region on the image. This relationship is the basis of semantic analysis of the static image. And localization of the object found in time using a tracker can be described by a verb of natural language.

The Bank of various detectors, there are a few detectors that analyze the image at the same time, correspond to the controlled vocabulary system for semantic image analysis. The volume of the dictionary says about possibilities of semantic analysis of images. For larger size dictionary semantic analysis will be more pure and complete.

2.2. The primary result

The primary result of the semantic analysis of image is the following table:

Table 1. Semantic signature of image.

ID	Object type	Object size	Object angle	Object location	Attribute 1	…	Attribute N
	Object identifier. Corresponds to type of detector (face, pedestrian, auto, etc.)	Object size feature	Object angle feature	In center, in border, left, right, up, down, et cetera	Object attribute. Takes values (-2,-1,0,1)	.	Object attribute. Takes values (-2,-1,0,1)

The Table 1 is named as semantic signature of image.

On figure 2 an example of semantic signature forming is shown.

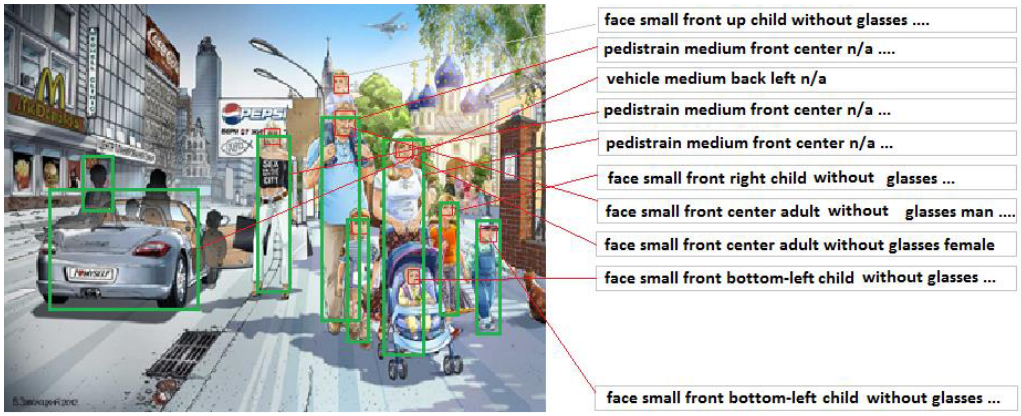

Fig. 2. The formation of semantic signatures image using a Bank of videodetection objects and attributes detectors.

2.2.1 Distance between the images

As shown above, to describe the image so-called "semantic signatures" are used, which can be represented in the form of a table. The semantic description of the image is a set V={v ⃗_1..v ⃗_n }, which every element v ⃗_i={⟦x_1⟧^i..⟦x_k⟧^i },i=(1,n) describes the properties of a specific fragment of an image. The elements of these vectors form a set for description of the properties of the fragment obtained using available detectors and object detectors attributes. To build the system, we need to introduce the definition of distance between semantic descriptions of images, the example is presented in Figure 2. To solve the problem, we introduce the following definitions:

Recoding. This encoding is used to represent each element of the vector $\vec{v} = \{x\_1..x\_k\}$, in the form of a number.

The **distance** between two vectors $\vec{v} = \{x\_1..x\_k\}$ that describes some fragment of an image.

The **distance** between the sets $V = \{\vec{v}\_1..\vec{v}\_n\}$ that describes the entire image and is its semantic description.

2.2.2 Decoding and comparison of codes

This encoding is used to represent each element of the vector $\vec{v} = \{x\_1..x\_k\}$ in the form of a code number corresponding to a particular property of the found object. Consider the elements of a vector in more detail. They are in general, correspond to the columns of Table 1. In particular:

- Object type
- Object size
- Object location
- Object angle
- Object attributes

The recoding of an object type is presented in Table 2.

Table 2. The recoding of an object type.

Detector name	Object name	Code
Face detector	People face	1
Pedestrian detector	People pedestrian	2
Car detector	Car	3

The recoding of an object scale is presented in Table 3.

Table 3. The recoding of an object scale.

Object scale	Square criterion	Code
Very large	>50%	1
Large	>20% <50%	2
Medium	>5% <20%	3
Little	<5%	4

The recoding of an object location is presented in Table 4.

Table 4. The recoding of an object location.

Object location	Code
In the center	1
In the left	2
In the right	3
At the top	4
At the bottom	5
Top-left	6
Top-right	7
bottom-right	8
bottom-left	9

The recoding of an object angle is presented in Table 5.

Table 5. The recoding of an object angle.

Object angle	Code
Straight	1
Straight-left	2
Straight-right	3
Left	4
Right	5
Behind	6
Behind-left	7
Behind-right	8

The recoding of an attribute is presented in Table 6.

Table 6. The recoding of an attribute.

Attribute disponibility	Code
It is impossible to determine attribute	-2
Attribute is absent	-1
Attribute is present	1
Attribute is not determined	0

For comparison the corresponding codes x_i describe the fragment of an image use the binary function (1).

$$h_i^{km} = \begin{cases} 1, x_i^k \neq x_i^m \\ 0, x_i^k = x_i^{m'} \end{cases} \quad (1)$$

where k, m correspond to semantic descriptions of image K and image M. note that if the attribute cannot be computed, the function returns 1.

2.2.3 The distance between the fragments of images

As we already know, the image fragment is described by a vector $\vec{v} = \{x_1 .. x_k\}$, where each element is a code describing some properties of this fragment (tables 2-6). Being able to compare the elements of this vector (1), we can introduce a function of distance between the vectors k,m, that describes a rectangular fragment of the image (2).

$$\rho^{k,m} = \sum_{i=1}^{N} h_i^{km}, \quad (2)$$

Using (2) the distance between identical fragments, in terms of their semantic descriptions, will be equal to zero, and the maximum distance between the fragments is equal to N – the number of elements in the vector, which describes any fragment of an image.

2.2.4 A measure of the closeness between semantic descriptions of images

In order to establish the degree of "similarity" of images in terms of their semantic descriptions, it is necessary to establish a measure of the closeness between them, which, as in the case of the description of image fragments for identical descriptions would give 0, and for all the other a number > 0.

The Hausdorff Measure. Hausdorff distance (HD) [6] is a metric between two sets. As an example, consider two sets of points in the plane $A = \{a_1 \dots a_m\}$ and $B = \{b_1 \dots b_n\}$. In this case, the Hausdorff distance between these two sets of points is defined as:

$$H(A,B) = \max\big(h(A,B), h(B,A)\big), \tag{3}$$

$$h(A,B) = \max_{a \in A} \min_{b \in B} \|a - b\|. \tag{4}$$

Here h(A,B) is called a "directed Hausdorff distance" from A set A to the set B. A modified Hausdorff measure. For many applications, including for our search task measures the closeness between the semantic descriptions of the images, it is possible to apply the so-called "a modified Hausdorff measure" (5), it differs from (3) is that instead of the maximum distance it uses the average distance from set A to set B.

$$h_{mod}\ (A,B) = \frac{1}{|A|} \sum_{a \in A} \min_{b \in B} \|a - b\|, \tag{5}$$

Substituting in (5) expression (2) by using (1), we get the expression (6), which we will be used in the expression for the proximity measure between the semantic descriptions of the image (3), which can be used to assess the degree of similarity of two images.

$$h_{mod}\ (A,B) = \frac{1}{|A|} \sum_{a \in A} \min_{b \in B}\ \sum_{i=1}^{N} h_i^{ab}\ , \tag{6}$$

where h_i^{ab} is i-th code corresponding to the vectors a and b describing the fragments of two different images.

3 Summary

1. Using the concept of distance between images, we can find the "semantically similar" between images or videos.
2. Using the semantic analyzer of images presented in the paper gives the opportunity to create an indexed database of video information that include not only images, but also an additional set of metadata that can be obtained automatically during analysis of the image or video semantic analyzer, and which characterize an image or video.
3. Using the thus obtained additional information, you can organize a quick search for the desired video or image based on its semantic characteristics only.

References

1. R. Duda, P. Hart, «Pattern Classification and Scene Analysis», NY, 1973.
2. U.Prett «Digital Image Processing».
3. Paul A. Viola, Michael J. Jones Rapid Object Detection using a Boosted Cascade of Simple Features// CVPR (1) 2001: pp. 511-518
4. Bellustin N.S., Kalafati Y.D., Kovalchuck A.V., Telnykh A.A., Shemagina O.V., Yakhno V.G Neuropodobny detector litsa. Technicheskie osobennosti realisatsii I obuchenia// X Vserossiaiskaya nauchno-technicheskaya konferentsija "Neiroinformatika-2008", Sbornik trudov, Tschast 2 MIFI, Moskva, yanv. 2008, p 123-132
5. N.Bellustin, A. Kovalchuck, A. Telnykh, O. Shemagina, V.Yakhno, Y. Kalafati, Abhishek Vaish, Pinki Sharma, Shirshu Verma Instant Human Face Attributes Recognition System // (IJACSA) International Journal of Advanced Computer Science and Applications, Special Issue on Artificial Intelligence, pp.112-120

6. Oliver Jesorsky, Klaus J. Kirchberg, and Robert W. Frischholz Robust Face Detection using the Hausdorff Distance // Third International Conference on Audio- and Video-based Biometric Person Authentication, Springer, Lecture Notes in Computer Science, LNCS-2091, pp. 90–95, Halmstad, Sweden, 6–8 June 2001