Shelia Guberman, Vadim V. Maximov, Alex Pashintsev

Gestalt and Image Understanding

I stand at the window. Theoretically I might see
there were 327 brightnesses and nuances of color.
Do I have "327"? No. I have sky, house, and trees.
M. Wertheimer

Was ist das Schwerste von allem?
Was dir das Leichteste dünket,
Mit den Augen zu sehn,
Was vor den Augen dir lieget.
Goethe

Introduction

The last two decades were marked by the appearance of computer programs that partly succeeded in solving some sophisticated problems imitating human decision making (Artificial Intelligence). At present the hot spot of Artificial Intelligence is image search.

There was great success in developing Google's search for particular words on the internet. Now there is a big demand for image search on the internet, which is populated with billions of images. Until now minimally acceptable solutions are not found. Google do it indirectly – the user defines key words for desirable images, Google's search engine finds sites which contain these words, and grabs adjacent images. The results are far from being adequate. During the last 50 years there have been many attempts at image recognition. The main tool was pattern recognition technique, and the images were restricted to a single object. In most cases solutions are based mainly on complete enumeration of possibilities plus a number of heuristic restrictions. If even partial success is achieved it is determined to a great extent by high speed of computers (so a huge number of possibilities could be analyzed) and tremendous size of memory (so

GESTALT THEORY
© 2012 (ISSN 0170-057 X)
Vol. 34, No.2, 143-166

a huge number of examples could be stored and used for comparison). There are three main obstacles in recognition of isolated objects: 1) dividing the image in adequate parts, 2) recognizing 3-dimensional objects at different angles and of different size, 3) recognizing generalized notions (like sky, tree, road, forest etc.).

The main cause of stagnation in that field was the neglect of knowledge accumulated in psychology of perception in general and in Gestalt psychology in particular. Too much was counted on mathematics and engineering and too little on laws of human perception, which has to be imitated. In this paper we demonstrate that by using principles of Gestalt psychology combined with basics of Linguistics (concepts of "language of meaning" and "adequate language") it is possible to come up with a computer program which works humanlike in quite a big domain of real images. But our goal is not only developing an application for image search engines. We believe that reaching that goal will satisfy the requirements Gestalt psychology stated in a modern review: 'Gestalt principles are usually illustrated with rather simple drawings. Ideally, it should be possible to apply them to an arbitrarily complex image and, as a result, produce a hierarchical parsing of its content that corresponds to our perception of its wholes and subwholes. This ambitious goal is yet to be accomplished' [Todorovic, 2008]. We also believe that a solution based on fundamental scientific principles of Gestalt psychology will facilitate solutions of a number of adjacent problems in AI (as was happening in the past). And even more: we are sure that implementing Gestalt laws in computer programs will make some notions and procedures of Gestalt psychology more formal and clear, and therefore make them easier to use in AI. We are optimistic about the path we chose, taking into consideration that a couple of other basic AI problems (for instance, abstract object detection [Guberman 2008], handwriting recognition [Andreevsky 1996], clustering analysis [Guberman 2002]) were resolved on the basis of Gestalt psychology after many years of unsuccessful attempts made by formal mathematical approaches.

1. Language

Linguistic Interpretation of Gestalt

In two papers in *Gestalt Theory* [Guberman, Wojtkowski 2001, Guberman 2007] the notion *Gestalt* was interpreted as ***short description.*** It was shown that such an interpretation is in agreement with views of Wertheimer, Köhler and Metzger expressed in their writings [Kohler 1975, Metzger 2006, Wertheimer 1923]. When they described our perceptions of different visual stimuli it was always a description ("circle", "two crossing lines", "Maltese cross in a quadrangle"). Wertheimer used expressions *good Gestalt* or *bad Gestalt* when he referred to images with simple or complicated descriptions (NB: not simple or complicated **images**, but simple or complicated **descriptions**). Metzger mentioned that when the stimulus distribution permits an organization in **simple** Gestalten, then these ‘‘good’’ forms prevail. Max Wertheimer’s point of view, expressed in his 1923 paper, was represented 60 years later by Michael Wertheimer [Brett et al, 1994]: *creating a meaningful configuration from lines and dots is governed by dynamic processes (based on their similarity, proximity, closure, continuity, and the like)* ***toward simple Gestalten***. In all cases when these terms were used they refer to images which have either a short and simple description (good Gestalten) or a long and complicated one (bad Gestalten). Such an interpretation of the notion of “Gestalt” does not contradict one of the basic meanings of the word *Gestalt* – shape. The good shape is the shape that can be described in brief, or easily reproduced or created. For example, the line, the circle, and the rectangle are described in brief, and can be easily recreated.

In the paper mentioned above it was shown also that the Gestalt, which one perceives from a particular image, represents not only the given image, but also a set of images, which our perception will refer to one class carrying the same pattern: the same Gestalt. The Gestalt percept from Fig. 1a as “two crossing lines” represents the set of images in Fig. 1b with the same Gestalt.

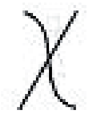

Fig. 1a

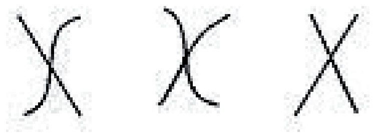

Fig. 1b

Following drawing 2 is Wertheimer’s example. Fig. 3 represents corresponding sets of drawings with the same Gestalt.

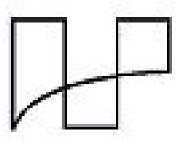

Fig. 2

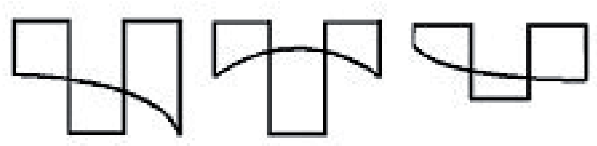

Fig. 3

It is trivial to say that the drawings in each line are similar. But now we can explain what we mean by the word *similar*: they have the same rough description, i.e. same Gestalt. More than that, we can build a model of recognition: *to recognize in a pool of drawings some, which are similar to a given one, we have to compare not the images themselves but the description of images: the Gestalts.*

Pattern Recognition vs. Adequate Language

Despite that the fathers of Gestalt psychology and most of their followers worked with linear drawings, there is no doubt that Gestalt principles are fundamental for our whole vision and are applicable to complicated images. In this paper we are trying to apply the above described approach to the problem of automatic image understanding. In the last 50 years in Artificial Intelligence (AI) was developed a fruitful method of generalizing experimental data by teaching the computer by a set of examples – pattern recognition method. Pattern recognition was often applied to image understanding but always met a lot of difficulties. Gestalt theory gives us two powerful tools to approach the problem of image understanding: 1) the ability of the Gestalt to generalize on the basis of a single example (contrary to traditional pattern recognition that demands a representative set of examples) and 2) the Gestalt has to be expressed on the linguistic level – as a description.

The words "pattern recognition" originating in the field of AI were considered in a visual context, as the recognition of visual patterns [Rosenblatt 1958]. Soon this was generalized to abstract patterns represented by a set of numbers [Bongard 1970, Guberman 1964]. Successful application of pattern recognition in the fields of (amongst others) geology, geophysics, medicine, and sociology over many years strengthens the belief that visual tasks were pattern recognition problems as well. This means that the approach was "learning through examples". Nowadays, for example, most programs which can distinguish human faces on photographs use this approach: a couple of hundred examples of full-face photographs are used for learning, producing a decision rule, then applying that decision rule to every rectangle of a given size on the image. Then the size is slightly modified, the learning and searching are repeated, and so on. Then a set of human-face images with head turned at an angle are taken and the procedure is repeated. Then human-face photos with head tilt are taken, continuing with the same approach.

At the same time another idea appeared and began to develop – the idea of an adequate language [Vasiliev 1969]. It started as a pure idea but very soon it began to find practical support through a series of applications. The first successes in implementing that idea were in medical diagnosis, earthquake prediction and oil exploration. The reason for these breakthroughs in each case was the use of a new language adequate for describing the phenomena being studied. But the algorithmic basis of all these solutions was pattern recognition based on learning through examples.

When considering visual objects, however, the choice of an adequate language looks different. In technical applications (as in geology, seismology, and medical diagnosis) the number of possible descriptions is sometimes very large but always finite. In the case of visual images the set of possible descriptions is practically infinite. Thus, in image processing, the task of finding an adequate language becomes the key problem.

In the early years of AI, most attention was attracted by printed and hand-written text recognition. Many descriptions were proposed to describe printed and handwritten characters, hundred of examples were used by the computer for learning, dozens of different algorithms were proposed: but with only partial success. The real breakthrough came when a general principle of constructing an adequate language was applied: the imitation principle as proposed by M. Bongard (Bongard 1970).

However, for more complicated images – images of the real world – it is difficult to apply the imitation principle to finding an adequate language of description. The imitation principle can be applied to man-made objects, but who made the sky? We will return to that crucial problem later.

Image Search Engine: Mel'chuk's approach

We discussed some general ideas of image recognition and understanding. Now let us define a practical problem that will give us the possibility to measure how good our ideas are. Let's consider the problem of finding in an image database a subset of images similar to a given image. This is how "state-of-art" was described in 2008 by scientists from Google Inc.

> "The majority of image searches use little, if any, image information to rank the images. Instead, commonly only the text on the pages in which the image is embedded is used. Although certain tasks, such as finding faces and highly textured objects like CD covers, have been successfully addressed, the problem of general object detection and recognition remains open" [Yushi Jing 2008].

We believe that an algorithm for finding images similar to a given one has to simulate the human ability to resolve this problem. How can one describe human behavior in solving this problem? The person looks at the given picture and then starts to take pictures out of the box ("the database") and looks at them one by one. The person examining the images from the box rejects most of them without looking at the given image more than once. Only occasionally does he look at the given picture more carefully. Some of these pictures still get rejected but some are moved to the pile of images similar to the original one. Such behavior can be explained by suggesting that humans compare not images themselves (the original one and ones taken from the box), but rather descriptions of images. This common sense suggestion has good support in the model of Gestalt-based recognition described in paragraph 1: to recognize in a pool of drawings some which are similar to a given one, we compare not the images themselves but the description of images: the Gestalts.

The first conclusion from the above analysis is as follows: the algorithm for image understanding has to produce descriptions of images. This approach poses two problems: 1) what is an adequate language of general image description, and 2) which procedure of comparison (metric) should be used.

Because we want to imitate and learn from a human approach, let us ask the person searching through the images why he rejected a particular image. His answer could be something like: "In the original image I saw a person in a park, but in this image taken from the box I see a car on the street, which is completely different". Such description seems similar to the approach taken with children in first grade school: "John, what do you see in this picture?" – "I see a boy swimming in the river" or "I see a table in a room and flowers on the table". It seems that the starting problem is to teach the computer to recognize notions like "person", "park", "sky", "car", "street", "river", "room", "table", "flowers" and so on.

This simple analysis shows that to resolve the image search problem, any algorithm has to be able to describe content using an adequate language, which means being able to understand images. That idea is similar to the more general idea of I. Mel'chuk: to translate a sentence from one language to another one must translate the initial sentence to the language of meanings, and then to another language (model "text-meaning" [Mel'chuk 2010]). It is worth remarking that the initial sentence expressed in "language of meaning" is a generalized description of the initial sentence, **its Gestalt**, and many sentences with similar meanings have the same Gestalt.

We want to use the notions mentioned above (person, sky etc.) as a primitive language of meanings of images. If meanings of two images are similar, the images are similar too. Of course, this is only the first level of understanding images. The second level will have to understand the relations between the objects in the scene. The next level will have to be able to predict the development of the scene either in the future or in the past. First step on the way to this goal is creating notions by means of computer software. To create the full list of notions that a human being possesses is not realistic at the moment. But some notions are extremely useful in image searching. The most informative notions are ones that divide the database in two equal, or roughly equal, parts. These notions include "human being", "indoors", "landscape", "greenery" (trees, bushes, grass), "sky", "sunny", "building", "road", "car", "perspective". Any of these would be extremely helpful in searching databases. A combination of two or three of them would reduce the number of possible images dramatically (e.g. "human being" + "park").

Despite the primitive nature and fuzziness of our reasoning, the practical recommendations are very valuable for developing an algorithm for image searching: even from the very first notion implemented in the software the volume of the database to be searched by the user will be significantly reduced, and each additional notion would reduce it even further.

2. Segmentation

Objects and Borders

The overwhelming majority of images in modern databases are in color. So, for the recognition of scenes it is natural to use the fact that the sky is blue, clouds are white, vegetation is mainly green, roads are mostly gray, faces are yellow-red, shadows are mainly black, seas and lakes are blue. Consequently, the initial algorithms of image segmentation frequently used colors of objects. But at the same time, behind all the activities of developing, testing and improving programs for image understanding, stands a simple fact: all these colored objects could be recognized on a black-and-white photograph. This simple fact persistently led us to look for structural, geometrical, and positional features which somehow identify sky, forest, trees, mountains, etc. Moving in that direction presents a question: is it enough to know only the gradients of brightness of the gray picture, rather than the value of brightness? It seems that in very many cases the answer is "Yes".

So, it seems that we arrived at the starting point in history of image processing: the gradients of brightness are the basic elements of finding objects in the image. But after our long journey in image processing we interpret the situation differently. We are convinced that the initial procedure of image processing is not finding borders of an object, but finding an object and then defining its borders. In other words, we are not going from bottom to top, but from top to bottom. As a matter of fact, we see and recognize many objects despite them being only partially confined by clear visible borders (i.e. with large gradients). That is how we see trees, or clouds in the sky. That is how objects look in X-ray photos. That is how geologists outline tectonic plates – with borders partially defined. That is how a water spot appears on pants – with no borders at all. That is why arts of pen drawing and engraving exist. Therefore the starting point of image processing and understanding has to be finding areas with clear borders not finding points with high gradients, connecting them in lines, and enclosing the lines.

Our current approach looks similar: we generate a small number of hypothetical objects and choose one with borders of the best quality. These hypothetical areas are generated using differences in brightness and color, and we need to do it knowing only points of big gradients. In other words, the set of points of big gradients have to serve at the same time as generator of hypothetical areas (future objects), and as a measure of quality of these objects.

All this above means that segmentation (in its precise meaning) is not adequate as an initial procedure for image processing and understanding, because it defines all borders of prospective objects. It has to be a more fuzzy procedure: define position of objects, show clearly expressed borders, but leave some areas between objects hazy.

Looking at the image with gradients one can see that points of high gradient form not only lines (potential borders of objects), but some kind of texture as well (consisting of short breaking lines). Such texture helps interpret the objects: it helps separate trees from the sky, and sky from the water. It is also obvious that texture can seldom help in defining borders. That is why we used some measure of texture for interpreting spots, and not for segmentation of the image. In the problem of finding objects in an image with gradients only, texture can help in initial outlining of potential objects.

Color

The use of color in the image segmentation is complicated more by the vector nature of the color space. Colors of individual pixels in digital images are usually specified in a coordinate system RGB, which is device dependent, being used in systems based on electronic displays (TV, video, computers). However, an independent use of the coordinates of this space is unsuited for image processing. The use of notions based on human perception, such as brightness, hue and saturation, instead of the amount of each primary color (red, green, or blue) is more fruitful. In particular, the brightness of the surface depends on the orientation of the surface with respect to the light source. Therefore, in order to locate on the image an area corresponding to the same surface, it is useful to abstract from the brightness. The same goes for saturation. In photos of open spaces saturation usually decreases with distance, and remote trees or mountains look unsaturated.

We have used a coordinate system of brightness (lightness), hue and saturation CIE-Lhs in the color space, developed by the International Commission on Illumination (CIE) specifically for classification of colors according to the human visual system [Ford 1998]. This color space is almost linear with visual perception, and the CIE-Lhs coordinate system is perceptually uniform, its brightness parameter having a good correlation with perceived brightness. A variety of simplified systems of color coordinates (HSL, HSV, etc.), developed for computer graphics, also describe colors using the same names: brightness, hue and saturation. These representations appear to be less useful because they suffer from perceptual nonlinearities and an uneven distribution of their components. Another reason why we have used CIE color space, specifically the coordinate system CIE-L*u*v*, is that it possesses a Euclidean metric and the notion of Euclidean distance between colors is determined. The color metric is necessary for calculation of the scatter of color in proximity to a particular point on the image (as a characteristic feature of textured surfaces) and for calculation of value of the gradient of color.

Finding Objects

There is an old technique of finding objects by using histograms of brightness. A simple example is shown in Fig. 4. The histogram of the brightness for that image has two spikes – Fig. 5. A slice of the image at any level of brightness $B=T$ between these two spikes produces a bitmap, which outlines the object. On the real photos the brightness of the objects is never a constant. Representation of the object on a histogram will not be a sharp spike but a bell-like curve as well as the representation of the background. Still the minimum of the histogram between the two maximums will provide a reasonable threshold T and will outline the object. The goal of this procedure is to single out clusters of brightness. M. Bongard considered this procedure as one of the basic tools of our intelligence [Maximov 1975]. He used the term "heap" for "cluster" and "breaking down into heaps" for "clustering". We use the following measure of clustering

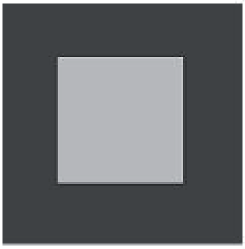

Fig. 4

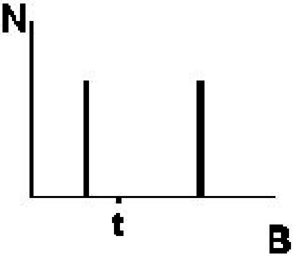

Fig. 5

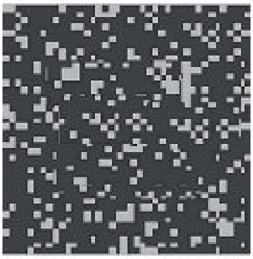

Fig. 6

$$k(T) = \frac{\sqrt{n_1 \cdot D_1} + \sqrt{n_2 \cdot D_2}}{\sqrt{n_0 \cdot D_0}}$$

where D_1 is dispersion of the left part of the histogram ($B < T$), D_2 is dispersion of the right part of histogram ($B \geq T$), n_1 and n_2 are number of pixels in each part of the histogram, and D_0 and n_0 are dispersion and total number of pixels for the entire histogram. When clusters overlap, the best dividing threshold corresponds to the minimum $k_{\min} = \min_T k(T)$.

The majority of real photos are more complex: the object is represented on a histogram by more than one maximum, the clusters are asymmetric and overlapping, the difference in brightness of different objects is small, etc. All that makes this tool useful in a limited number of cases. Not all these difficulties originate from the complexity of the reality. There is a shortcoming in the procedure alone: the algorithm does not care about ***position*** of pixels with particular brightness. The histogram for the image in Fig. 6 (known as "salt-and-pepper") is identical to the histogram of the image in Fig. 4 although there are no objects in Fig. 4.

To overcome this defect, a pair of spaces has to be introduced: one space is the one-dimensional histogram of brightness $H = H(B)$, the second space – the dual 3-dimensional space of the original image itself $B = B(x, y)$. The first space allows us to measure how compact is the distribution of the ***brightness*** of the image by calculating minimal clustering k_{min}.

Threshold brightness T corresponding to k_{min} defines the binary (black-and-white) image – bitmap $b = \varphi(x, y)$, where $\varphi(x, y) = 0$, if $B(x, y) < T$, and $\varphi(x, y) = 1$, if $B(x, y) \geq T$. The bitmap b is an object in dual space. On that bitmap a measure has to be defined reflecting how compact distributed black (or white) ***pixels*** are. For example, the measure of compactness for the bitmap in Fig. 4 has to be much higher than for the bitmap in Fig. 6.

A number of measures for compactness can be discussed.

1. *Number of spots N on the bitmap*. The fewer N the higher is the compactness of the bitmap. It works well for "salt-and-pepper" images – Fig. 6.
2. *Length of all borders L on the bitmap for a given threshold T*. That measure separated the Fig. 4 and Fig. 6 as well: the shorter the border the more is the compactness on the bitmap.
3. *Value of the gradients on the object's borders*. The ideal situation is when all objects have big gradients on their borders. But the reality is far from the ideal. That is why the common approach that starts with finding areas of high gradients and then proceeds to find objects has so many difficulties. The DC approach starts with finding objects (spots on bitmap b) and then estimates in dual space the "quality" of their borders. In other words, we are not looking for points of high gradient, but for objects with good borders. By the way, this measure, which is very useful in gray and color images, doesn't work on Fig. 6.

Because each of the proposed measures has its own pro and contras we construct a combination M_{DC} that reflects 1) difference in brightness between the object and the background measured by k, 2) length of all borders L reflecting the geometry of the object, and 3) mean gradient on the borders G, which reflects the quality of the border:

$$M_{DC} = \frac{G}{k \cdot L}.$$

The bigger the M_{DC} the better is the quality of segmentation.

Image Segmentation

Principles previously described were implemented in a program, which executes the following steps.

1. Input image is split in three channels: Hue, Saturation, and Brightness (Lightness).
2. Gray areas on the image are found (as areas with low saturation). These areas are excluded from the image in the Hue channel.
3. For segmentation the Dual Clustering (DC) procedure is applied to each channel (*H, S, L*), i.e. for each channel $M_{DC}(T)$ was calculated and the maximum M_{DC} and corresponding threshold were kept ({M^{B}_{DC}, T^{B}}, {M^{H}_{DC}, T^{H}}, and {M^{S}_{DC}, T^{S}}).
4. The largest of three M_{DC} values was chosen and appropriate *T* was used to create the bitmap representing a chosen segmentation. That bitmap divided the complete image into two segments: all black pixels and all white pixels. Each segment is then divided in non-overlapping connected sets of pixels – spots.
5. The algorithm continues by applying recursively the Dual Clustering procedure to each spot of the image obtained at the previous step.
6. At each step spots are eliminated if (a) the spot is too small, or (b) the measure of clustering M_{DC} for that spot sunk below some threshold.

Segmentation stops when all spots are eliminated.

The most time-consuming part of Dual Clustering is finding maximum calculating M_{DC} for each modality (*L, H*, and *S*). For that purpose 255 black-and-white bitmaps have to be generated (for each of 255 values of given modality). On each map borders of all spots have to be identified. Each pixel of an image has 4 neighbors. The brightness of that pixel and of all its neighbors is known. Having these values one can find at which thresholds *T* that pixel will be a border pixel. According to the definition, a pixel is a border pixel if at least one of its neighbors belongs to the spot and at least one of its neighbors belongs to the background. Let B_0 be the brightness of the given pixel. Let B_j ($j = 1, 2, 3, 4$) be the brightness of its neighbors. Let B_{min} be the minimum of B_j. First, it has to be noted that a pixel with brightness B_0 can be a border point of some spot on a bitmap only if the threshold *T*, that created that bitmap is less than B_0. Now, if the bitmap was created by the threshold *T*, which is smaller than B_{min} ($T < B_{min}$), then the central pixel and all neighboring pixels will belong to the spot, and the central pixel is not a border point. In case the threshold is between B_{min} and B_0 the central point will belong to the spot and at least one pixel (with brightness = B_{min}) will belong to the background, i.e. the central pixel will be a border point. That information has to be defined only once for each pixel and then it becomes known on which bitmaps (i.e. for which thresholds) it will be a border point.

3. Concepts

From Objects to Notions

As soon as the image is segmented into spots we can work on further interpretation: to find the notions. As was mentioned before, the list of notions which are useful in outdoor scenes without people or animals is as follows.

1) sky,
2) vegetation (trees, bushes, grass),
3) building,
4) road,
5) car,
6) mountains,
7) water (sea, lake, pool, river).

The above-mentioned notions are of a varied nature. Some of them are well-defined objects which could be described by a small number of features. For example, a car has four wheels and a body. Another example is the human face (two eyes, nose, mouth). Difficulties in recognition are caused by the fact that they are 3D objects and appear on scene at different angles and therefore look different. Nevertheless, it is possible to teach the computer to recognize these objects using a limited number of views at different angles and of different sizes for learning purposes. It is not very sophisticated but it could work.

What about sky, or vegetation, or water (seas, lakes, rivers)? They can not be represented by a limited number of views, as they are not physical objects but concepts. Sky does not exist as a physical object, sky is a universal background, it has no shape. W.Metzger noted it as early as 1936 [Metzger 2006]: "sky is not an object, but actually continues behind foreground objects". The concepts of vegetation, buildings, and human bodies have the same problem: too many appearances.

When we confronted the image understanding problem we decided to develop simple and reasonable algorithms to understand the reality of images, with a readiness to change our understanding of visual objects, colors, scenes, and recognition. And unlike our segmentation algorithm, in finding notions we use the simplified HSL color space. We had no difficulties in creating notions induced by color problems. We also modified our segmentation algorithm by replacing CIE representation of color with simplified HSL coordinates. At first glance it caused minor changes, which is crucial, and didn't change the list of found notions. We believe that it happened because we were trying to imitate the human perception at a very low level, and the simplicity of the tools turned out to be adequate for the simplicity of the task.

Here are the short descriptions of algorithms for finding notions. Examples are shown in Fig. 8 - 16, where descriptions generated by our software are given below each image.

Sky

Every spot found by segmentation could be described by shape of its borders, by color, by brightness, by position relative to the frame of the image, and by position relative to other spots. Appearance of sky varies dramatically in color and shape. Geometrical characteristics of the borders of sky spot in an image are borders of other objects: buildings, mountains, trees.

In the search for the sky the analysis began with spots (result of segmentation) of significant brightness and particular color (in HSL coordinates from H=130 to H=170). Usually it is connected to the upper border of the frame. As a rule, it covers a significant area of the image or touches a significant part of the top border of the image. It is often found at some distance from the bottom border of the image. Sky could be represented by one spot, or by a number of spots. Of course, for each of these "rules" a number of contrary examples exist, but still the rules cover the majority of real outdoor pictures. We would like to reiterate that in the beginning it is preferable to develop simple and reasonable algorithms and clarify the obstacles of real image understanding.

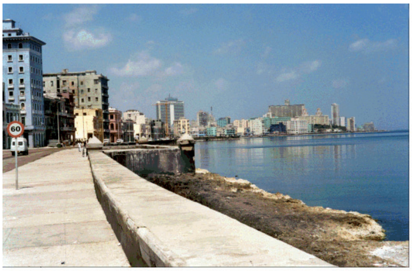

Fig. 7a

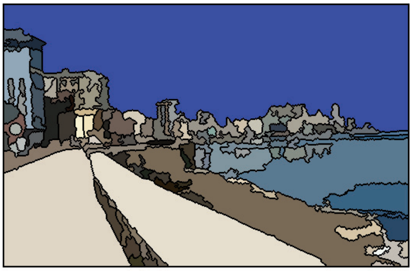

Fig. 7b

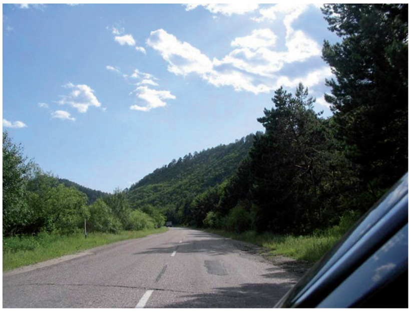

Fig. 8a

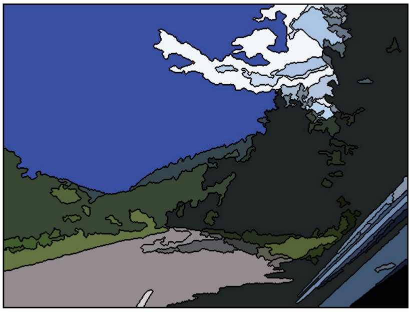

Fig. 8b

Figs 7 and 8 illustrate the operation of the algorithm. Original images were first subjected to segmentation, and each spot was painted in its average color. Segmentation results are shown in Figs 7b and 8b. Then there were found spots corresponding to the sky.

Our experience with the program has shown that in most cases of segmentation of outdoor images by this algorithm the first division occurs by hue channel. Typically, it is a division into two segments with warm and cool colors. In this case one or more spots, which form the segment of cold colors, meet the "rules" stated in this section, i.e., they represent the "sky". Thus, with a "top-to-bottom" scheme to find the sky it is unnecessary to carry the segmentation procedure to the end. All is revealed in the very first steps.

Clouds

Sometimes there are clouds in the sky – sometimes they are light with fuzzy borders and our segmentation fails to represent them as distinctive objects (Fig. 7). Sometimes clouds are well-defined and create objects, which are part of the "sky" – Fig. 8b. Commonly they are white or gray, and completely or partly surrounded by sky. There are some other objects that could appear in the sky and be misrecognized by the above-mentioned rules as clouds: balloons, airplanes, blimps. The distinctive features of such artificial objects are sharp borders, color, and texture.

Vegetation ("Green")

Green (trees, bushes, grass etc.) is a very common part of outdoor non-urban scenes. It is clear that not all trees belong to that notion: the trees without leaves (winter trees or burned trees) or fall trees covered with red and yellow leaves are excluded. Here once more we face the reminder of limitations of our approach to image understanding – dependence on color: the human eye can recognize vegetation on gray images.

There are three main difficulties in identifying the notion of "green":

1) objects belonging to that notion have no definite shape,
2) the same object appears quite different from different distances,
3) texture and color saturation are extremely variable.

Naturally, the first obstacle is the existence of artificial objects colored green. It could be overcome by measuring the smoothness of the green surface (in contrast to vegetation, which is characterized by sharp changes in brightness and saturation). That is due to the essential three-dimensional structure of vegetation, which exposes to the observer deep dark pockets between the brightly illuminated leaves.

Another fundamental feature of the "green" notion is the size. To be an important part of the scene, the vegetation has to occupy a significant part of the scene. And that is one more option for differentiating "green" from many green artificial objects. To look ahead, the presence of such notions in the scene as sky or lake increases probability that a green spot is vegetation. Important features for "green" (like size, homogeneity, and position) can be defined only after the spot itself is defined. For that purpose a "green" channel was established by cutting out from the Hue channel the green interval (from H=50 to H=130). It solves the problem in a significant number of scenes but in many cases it extracts a mosaic of separated small spots – a tiny part of the vegetation visible on the image. The rest of the vegetation is closer to blue or red parts of the spectrum not to mention the parts of vegetation in deep shadows, which appear dark gray. If we try to expand the "green" area by expanding the interval extracted from the Hue channel, it picks out a lot of spots which are not part of vegetation.

We chose the following practical solution:

1) get spots from the "green" channel (from H=50 to H=130),
2) get spots from expanded channel (from H=30 to H=150),
3) add spots from the expanded channel, which have common borders to the "green" spots,
4) add gray spots, which fill holes in spots created in point 3.

Examples are shown in Fig. 10-18.

Trees

Between a broad variety of vegetation trees are most distinctive, particularly the stand-alone tree. The tree appears green in the center (around the stem), when covered with leaves, and exposes separated branches at the edges. Between the branches at the edge one can see the sky (or other background), which constitutes bays in the spot interpreted as sky. Between the sky bays and the green mass in the center of the tree there is a silent not interpreted zone. That zone contains small green spots isolated from the big central green spot, which because of their small size were excluded from further analysis. Similarly, that zone is occupied by small blue spots – sky visible through openings in the leaves – and consequently dropped from the analysis. An example of segmentation and interpretation resulting in "trees" extraction is shown in Fig. 15-18.

Water

Water is created from the same material as "sky", i.e. from blue spots and from gray spots minus spots recognized as sky or clouds. It was postulated that images that have "water" must have "sky" (evident restriction on the class of recognizable images). Each of the selected spots went through a number of tests analyzing geometrical and positional characteristics of the spot: width, height, touching the frame on left or right, flatness of the spot's top border. An example is shown in Figs 7b, 13 and 18.

In some cases there is no detectable border between sky and water (no visible horizon). We use two features to divide the combined spot: 1) mostly brightness of the sky increases from zenith to horizon (from top of the image toward the bottom); brightness of water mostly decreases from horizon down, and 2) when the spot combines into one sky and water, the border between them (the horizon) is often located in the narrowest part of the spot (see Fig. 7). In cases when water doesn't contact the sky another feature useful for identifying "water" appears – existence of shadows of objects located on the far banks of lakes or bays (see the same image Fig. 7). Shadows in water can be recognized by horizontal symmetry of contours of objects.

Ground

The Earth's surface is a general background for the vast majority of images. In some subclasses of images it is completely covered by other objects (like in indoor scenes), in the majority of outdoor scenes the Earth's surface is covered only partially (by trees, buildings, cars and so on). Visible surface can appear differently: as lawn, road, or plaza. We will call it "ground". It occupies a significant area of the image (greater than 4%) and touches the bottom of the image. Despite the simplicity, it works in many cases.

Another kind of ground is not gray but green. Spots that satisfy all positional and geometrical characteristics of ground and are qualified as "green" become "green ground". Various types of vegetation can appear as "green ground": it could be grass, plants, bushes or forest (as it is seen from mountains). It seems that differentiation of these classes could be done using texture characteristics (like autocorrelation function).

A particular kind of ground is the road in perspective. The color and texture characteristics of the road are the same as of the ground, but it has very specific geometrical characteristics. Because borders of the road in reality are parallel, the width of the road on the image will decrease gradually as the distance to the observer increases. When the road was a straight line the width became a linear function of distance, and the position of the horizon could be found.

Mountains and Snow

Take a look at Fig. 9. The regular description of this image would be "polyline". Let's modify the image – add a feature that would create perception of the "sky" – see Fig. 4 (b). Now mountains appear on the image. It shows that "sky" is a very creative ingredient of an image. The top, left, and right borders of sky in many cases are silent – they contain little or no information on the image. To the contrary, the bottom border is highly informative.

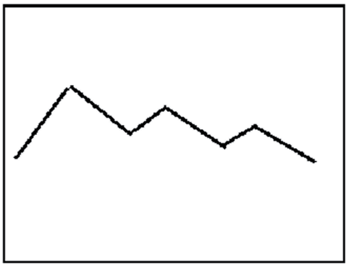

Fig. 9a

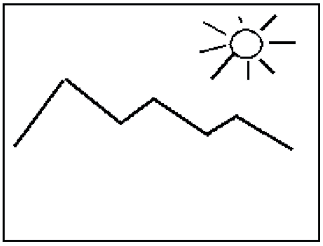

Fig. 9b

In the majority of landscapes the bottom border of sky is water, or green, or buildings, or mountains. Therefore the simple rule is: if the spot under sky was not recognized as water, or green, or buildings, it must be mountains. Of course, some restrictions on color, texture, size, and geometry have to be applied. Buildings that appear in front of sky can be identified by two procedures. One is the subroutine (detector), which finds buildings (see below), and the second one is a procedure that analyzes the bottom border of the sky spot looking for straight lines in general and vertical ones in particular.

If mountains will be covered with snow, spots that represent patches of snow are located between the mountains and the sky. If such spots exist and they satisfy some conditions then a statement is issued "mountains with snow" – Fig. 5 (c).

Buildings

The most characteristic features of images of buildings are straight lines – vertical and horizontal. Typically buildings are not represented on an image as a single spot because different walls have different luminosity, different parts of buildings have different colors (windows, walls, roof). All these parts appear as spots with linear borders. Vertical borders are an invariant of buildings. Lines, which are horizontal in nature in most cases, appear on images in perspective as a bunch of lines with a single imaginary apex. The majority of buildings have a periodical structure of windows. In urban scenes perspective lines are a valuable source of information: 1) they help to establish references between objects in the scene and its in-depth location, 2) if one object is recognized (like human, car or window), they allow estimation of the possible size of similar objects in other locations as shown in Fig. 14, 15.

Cars

In our list of basic notions, which we propose as the first step in building image understanding, "car" is the first notion of quite a different nature – it is defined by its geometrical form. The difficulty of recognizing cars is caused by the fact that cars look essentially different from different points of view. Despite all cars having similar main components, proportions are quite different and that increases difficulties of recognition.

There are programs for car recognition that are based on total search of cars in every point of an image and comparison of a given fragment with all possible appearances of a car (at different angles and different distances). We acknowledge that with gigantic speed and memory size of computers that problem can give satisfactory solutions in many applications. But from the history of AI we know that that kind of solution has very restricted areas of application. What we try to do is to find more intellectual tools that could deliver more general solutions. Let us deconstruct a modern vehicle to its ancient ancestor – keep only the wooden board and wheels – it is still a vehicle. There are some limitations that put restrictions of width of the base to wheels' diameter (roughly from 2 to 5). That means that the space under the vehicle is always in deep shadow. That feature doesn't depend on construction of the vehicle (with the exception of exotic cases). It is obvious that finding the darkest spots on the image will create a number of false alarms (deep shadows in trees, open doors in buildings, black paintings, etc). But it is also obvious that a number of natural restrictions can be applied. If the dark spot is inside the "green" area, or on the top of the image it is very unlikely to be an indicator of a car. There are also some general indications that there is a car over the dark spot: the car is mostly represented by a number of spots, and there are some geometrical restrictions on these spots (in size and relative locations). The majority of car images have straight lines. Some false alarms can be disaffirmed on the level of reinterpretation. Finding out if this approach would work could be done only by testing. An illustration is presented in Fig. 15.

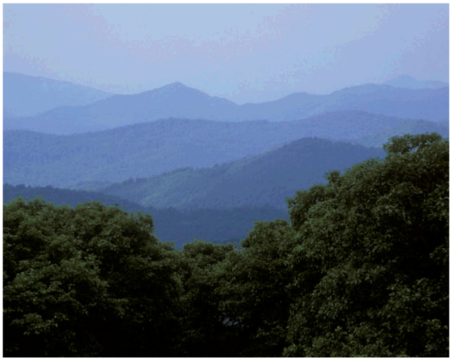

Fig. 10 Generated description: Blue Sky, Mountains, Green.

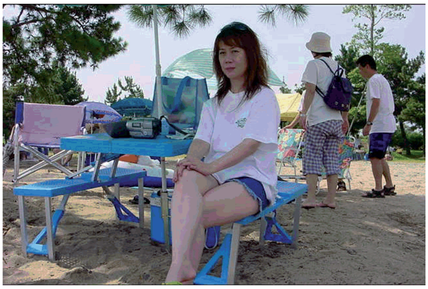

Fig. 11 Generated description: Sky, Green, Ground.

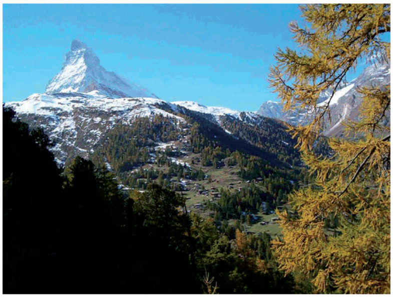

Fig. 12 Generated description: Blue Sky, Mountains, Snow, Green, One tree on the right, Ground.

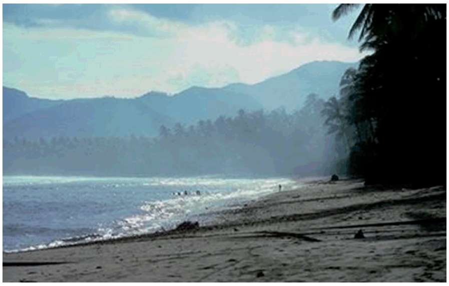

Fig. 13 Generated description: Green, Blue Sky, Ground, One tree on the right, Water, Mountains.

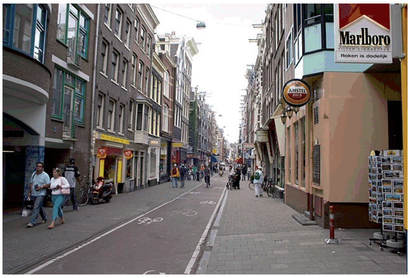

Fig. 14 Generated description: Gray Sky, Ground, City, Buildings.

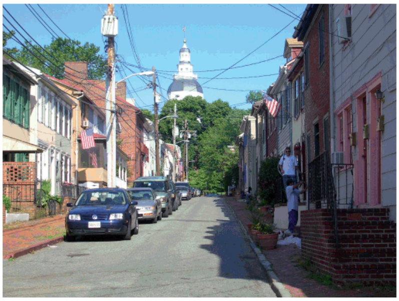

Fig. 15 Generated description: Green, Blue Sky, Ground, City, One tree on the left, one tree in the center, cars.

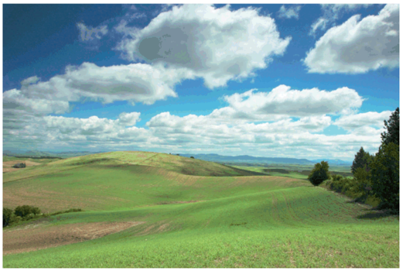

Fig. 16 Generated description: Green Ground, Blue Sky, Clouds, One tree on the right, Distant forest

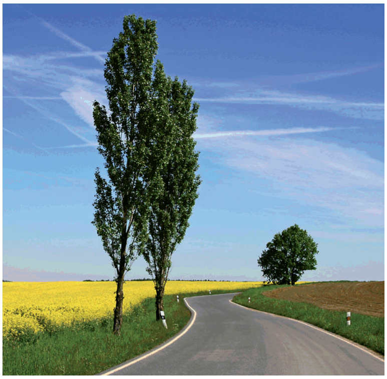

Fig. 17 Generated description: Green, Blue Sky, Ground, One tree in the center, trees on the left, Distant Forest

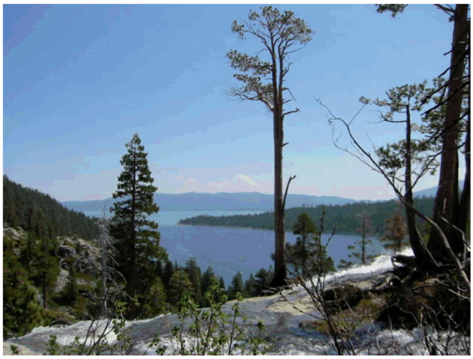

Fig.18 Generated description: Green, Blue Sky, One tree on the left, One tree on the right, One tree in the center, Trees in the centre, Distant Forest, Water, Mountains.

Reinterpretation

As Gestalt psychology claims, interpretation of a part of an image depends on interpretation of the rest of the image. All described above is mainly interpretation of each spot independently of other objects in the image. But even on that basic level of image understanding we were forced to reinterpret some notions. When searching for the "sky" we take in consideration whether the candidate for the sky is surrounded by trees. When looking for a car and finding one, we look around for more cars using fewer restrictions. If a car was found, we would look for ground with softer criteria. If sky is found, we would look on geometry of surrounding spots, and sometimes divide the sky spot into two spots: sky and water. If white spots are surrounded by sky, the spots are interpreted as clouds, but if white spots are surrounded by mountains, these spots are interpreted as snow.

Of course many other rules of reinterpretation must be (and will be) implemented at the first level of image understanding. According to Gestalt psychology, when dividing an image into parts two conditions have to be satisfied: 1) each part has to be meaningful, and 2) the whole has to be meaningful. On the first level we take care of the parts, on the next level we must take care of the meaning of the whole picture.

From the description of the DC algorithm it is clear that at first it outlines big spots. In the outdoor scenes it will be "sky", or "ground", or "green", or "water" and they could be immediately categorized as one of these notions. As was mentioned above, knowing one of these notions helps in segmentation and interpretation of other spots (completely in accordance with laws of Gestalt psychology). That is a big advantage of our "top-to-bottom" scheme of segmentation over "bottom-to-top" ones (like starting with gradients, or combining objects from pixels, which allows interpretation to start only after segmentation is finished).

Conclusion

1. Following basic ideas of Gestalt psychology and Linguistics we developed a working program that demonstrates image understanding in the domain of outdoor scenes by generating the first order description of the scene.

2. In that program we implemented a number of new ideas in image processing:

a) top-to-bottom segmentation by a Dual Clustering (DC) algorithm,

b) interpretation of an object depending on the nature of surrounding objects and their relative position (in complete accordance with Gestalt principle),

c) defining algorithms for generalized notions – concepts (like “sky”, “green”, “ground” etc.) that could not be defined in the tradition of classical pattern recognition based on a set of examples.

As a result we overcame two of three obstacles in image processing, which we mentioned in the introduction (dividing the image not in random pieces but in meaningful parts and generalizing complicated images through adequate short description). But we didn’t succeed in generalizing 3-D objects that look different at different angles (except “cars”).

3. The developed program opens two possibilities in constructing an image search engine – a program that searches the internet for images similar to one presented by the user. The weak version works as follows. Our program processes the given image and produces the description – a short set of words. That set of words is used as input to Google’s image search. In response Google returns a big set of images from the internet. The first 100 images are processed by our program that generates for each of the images a generalized description. Each of these descriptions is compared to the description of the initial image presented by the user and some measure of proximity is calculated. Finally, the 100 images are arranged according to the proximity and all images with proximity higher than some threshold are presented to the user.

The strong version of the proposed image search works as follows. Each image that appears on the internet is processed by our program and the generated description is stored with the image. So, a search for images similar to that given by user consists of straight comparison of the descriptions.

4. While working on that program we found that our program recognizes some concepts (like sky, clouds, mountains, ground, trees) on the gray images. It happens because our interpretation of objects is based more on structural and positional features and less on color. More than that: it turns out (for the same reasons) that some of the concepts we can recognize even on black-and-white images (like line drawings). All this means that we achieved deeper understanding of the nature of image perception.

5. The results demonstrate once more that the principles of Gestalt psychology developed by the founding fathers mainly on dotted and lined images are completely applicable to any images. More than that - the history of Artificial Intelligence proves that progress in imitating abilities of human vision couldn't be achieved without implementing principles of Gestalt psychology.

Summary

During the last 50 years there were many attempts in computer image recognition. At present the hot spot of Artificial Intelligence is image search on the internet. Still the results are far from being adequate. The main cause of stagnation in that field was the neglect of knowledge accumulated in psychology of perception in general and in Gestalt psychology in particular. In this paper we demonstrate that by using principles of Gestalt psychology combined with basics of Linguistics (concepts of "language of meaning" and "adequate language") it is possible to come up with a computer program which works humanlike in quite a big domain of real images of outdoor scenes. The program demonstrates images by generating the first order description of the scene.

In that program we implemented a number of new ideas in image processing: a) top-to-bottom segmentation by Dual Clustering (DC) algorithm, b) interpretation of an object depending on the nature of surrounding objects and their relative position (in complete accordance with Gestalt principle), c) defining algorithms for generalized notions – concepts (like "sky", "green", "ground" etc.) that could not be defined in the tradition of classical pattern recognition based on a set of examples. The developed program opens possibilities in constructing an image search engine – a program that searches the internet for images similar to one presented by the user.

The results demonstrate once more that the principles of Gestalt psychology developed by the founding fathers mainly on dotted and lined images are completely applicable to any images. More than that - the history of Artificial Intelligence proves that progress in imitating abilities of human vision couldn't be achieved without implementing principles of Gestalt psychology.

Keywords: Image understanding, generalization, adequate language, visual concepts.

Zusammenfassung

In den vergangenen 50 Jahren wurden viele Versuche auf dem Gebiet der automatisierten Bilderkennung gemacht. Der Schwerpunkt heutiger Forschung im Bereich der künstlichen Intelligenz ist die Bildsuche im Internet. Die Ergebnisse sind jedoch bei weitem nicht ausreichend. Der Hauptgrund für den Stillstand in diesem Bereich war die Vernachlässigung der Erkenntnisse, die allgemein von der Wahrnehmungspsychologie und speziell der Gestaltpsychologie zusammengetragen wurden. In dieser Arbeit zeigen wir, dass es durch die Anwendungen der Prinzipien der Gestaltpsychologie in Kombination mit Grundlagen der Linguistik (Konzepte „Sprache der Bedeutung" und „adäquate Sprache") möglich ist, ein Computerprogramm, das in einem durchaus großen Bereich von realen Bildern in freier Umgebung menschenähnlich funktioniert, zu entwickeln. Das Programm zeigt Bilder, indem es eine Beschreibung erster Ordnung der Szene entwickelt. Wir haben in dieses Programm eine Reihe neuer Ideen zur Bildverarbeitung implementiert: a) Top-down Segmentierung durch einen Dual-Clustering-Algorithmus (DC); b) Deutung eines Objekts abhängig von der Art der umgebenden Objekte und ihrer relativen Position zueinander (in vollständiger Übereinstimmung mit Gestalt-Prinzipien); c) Ableitung eines Algorithmus für generalisierte Begriffe (wie Himmel, Grün, Boden etc.), die nicht in der Tradition klassischer Mustererkennung, basierend auf Beispielen, definiert werden konnten. Das entwickelte Programm eröffnet Möglichkeiten zur Konstruktion einer Bilder-Suchmaschine – eines Programmes, das das Internet nach Bildern durchsucht, die einem vom Anwender vorgegebenen Bild ähneln.

Die Ergebnisse zeigen einmal mehr, dass die Grundsätze der Gestaltpsychologie, die durch die Gründungsväter entwickelt wurden und sich hauptsächlich auf punktierte und linierte Bilder bezogen, vollständig auf beliebige Bilder anwendbar sind. Mehr noch – die Geschichte der Künstlichen Intelligenz zeigt, dass Fortschritte in der Imitation menschlichen Sehens ohne die Anwendung von gestaltpsychologischen Prinzipien nicht erreicht werden konnten.

Schlüsselwörter: Bilderkennung, Verallgemeinerung, adäquate Sprache, visuelle Konzepte.

References

Andreevsky, E. & Guberman, S. (1996): From Language Pathology to Automatic Language Processing… and Return. *Cybernetics and Human Knowing,* vol. 3, No. 4, 41 – 53.

Bongard, M. (1970): *Pattern Recognition.* New York: Spartan Books.

Brett, K.D., Wertheimer, M., Keller, H. & Crochetiere, K. (1994): The legacy of Max Wertheimer and gestalt psychology. *Social Research,* 61, No. 4.

Ford, A. & Roberts, A. (1998): Colour Space Conversion. http://www.poynton.com/PDFs/coloureq.pdf

Guberman, S. & Wojtkowski, W. (2001): Reflections on Max Wertheimer's "productive Thinking:: Lesson for AI. *Gestalt Theory 23(2),* 132-142.

Guberman, S. (2007): Algorithmic Analysis of "Good Continuation" Principle. *Gestalt Theory, 29(2),* 148-168.

Guberman, S. & Woitkowski, W. (2002): Clustering Analysis as a Gestalt Problem. *Gestalt Theory*, *24 (2),* 143 - 158.

Guberman, S., Izvekova, M., Holin A. & Hurgin Y. (1964): Solving geophysical problems by mean of pattern recognition algorithm, *Proc of Acad.of Scies.of USSR*, Vol. 154, No. 5.

Köhler, W. (1975): *Gestalt Psychology. A mentor Book.* New York and Scarborough: Ontario.

Maximov, V. (1975): A system learning to classify the geometrical images. In: *Models of Learning and Behaviour (M.S. Smirnov ed.).* "Nauka", Moscow, pp. 29–120 (in Russian).

Mel'chuk, I. (2010): *Dependency Syntax: Theory and Practice.* State University of New York Press.

Metzger, W. (2006): *Laws of Seeing.* Cambridge (USA): MIT Press.

Rosenblatt, F. (1958): The Perceptron: a probabilistic model for information storage and organization in brain. *Psychol. Review*, Vol. 65, No. 6, 386–408.

Todorovic, D. (2008): Scholarpedia, 3(12):5345. http://www.scholarpedia.org/article/Gestalt_principles.

Vasiliev, J., Gelfand I., Guberman S. & Shik, M. (1969): Interaction in biological systems. *Priroda*, No. 6, (in Russian).

Wertheimer, M. (1923): Untersuchungen zur Lehre von der Gestalt II, in P*sychologische Forschung*, *4*, 301-350. English translation: http://psychclassics.yorku.ca/ Wertheimer/forms/forms.htm.

Yushi, J. & Shumeet, B. (2008): "PageRank for Product Image Search", WWW 2008, April 21–25, Beijing, China, 2008.

Shelia Guberman, PhD in Nuclear Physics, PhD in Computer Science. Former Chief Scientist at Institute of Applied Mathematics (Moscow). He is author of first application of Pattern Recognition to technical problems (oil exploration, medical diagnosis, earthquake prediction), of D-waves theory (seismicity), theory of similarity in Nuclear Physics, handwriting recognition technology. He published a number of papers on algorithmical analysis of Gestalt theory and its application to Artificial Intelligence.
Address: PO Box 2411, Cupertino, CA, 95015, USA.
E-mail: guboil@hotmail.com

Vadim V. Maximov is a Leading Scientific Researcher at the Institute for Information Transmission Problems, Russian Academy of Sciences, Moscow, Russia. He received his PhD in pattern recognition from the Institute for Information Transmission Problems in 1975. His research interests are in the fields of color vision and color processing in fish, frogs, toads, lizards, birds, cats, dogs, monkeys, humans and machines with special interest in the color constancy. He is the author of a book "Color Transformation with Changes of Illumination" (Moscow: Nauka 1984). He is member of the editorial board of the Journal of Integrative Neuroscience, Imperial College Press.
Address: Laboratory of Sensory Information Processing, Institute for Information Transmission Problems, Russian Academy of Sciences, Bolshoi Karetny 19, 127994, Moscow GSP-4, Russia.
E-mail: maximov@iitp.ru

Alex Pashintsev holds an MS degree from the Moscow Technical University of Radiotechnics, Electronics and Automation. With more than 20 years of expertise in advanced R&D and software development, he led the creation of three generations of handwriting recognition and OCR systems, digital ink technologies and products - including Apple Newton project, CalliGrapher, riteScript, AIR (Advanced Image Recognition) and other key technologies. He was Project Manager at ParaGraph Intl. and Silicon Graphics, Program Manager at Vadem and Vice President of R&D at Parascript and EverNote.
Address: Evernote Corp., 333 W. Evelyn Ave., Mountain View, CA, 94041, USA.
E-mail: avp2@msn.com