Toward a Computational Theory of Perception

Rafael Grompone von Gioi
IIE, Universidad de la República, Uruguay
CMLA, ENS Cachan, France
jirafa@fing.edu.uy

ABSTRACT
This paper sketches a computational theory of perception. Perception is the process of acquisition of information about the environment by the senses. In this proposal, a model of the world and models of how information is projected into the senses are needed. Given these models, perception is the process of finding one configuration of the known world that corresponds well to the sensed information. The key point is how to validate a perception: a configuration of the environment is validated as a perception if the expected sensory information (the projection from the tested configuration into the senses) is similar to the observed one, up to the precision of the senses. Learning the models for the world and the senses is considered as different process from perception. A complete computational implementation is presented for a toy example of visual perception in a world of flat and superposed squares.

Author Keywords
perception, computational theory, computer vision

INTRODUCTION

The aim of perception is to obtain information about the environment. However, only information from the senses is available. How is it possible to integrate local tactile information and say that an object is round? How can visual and auditory information be merged? How is it possible for an animal to make global decisions based on the local information provided by the senses? This paper sketches an algorithmic theory of perception and shows a complete computational implementation for a toy example.

This paper states that perception is dependent upon both, a model of the external world and models for the senses. The perception process tries to find a configuration of the model of the world that is coherent with the sensed information. To validate a configuration, knowledge about the sensors is used to project the configuration under evaluation to obtain the expected measurements in the sensors. The configuration is accepted as a perception only if the measurements in the sensors correspond with those expected, up to the sensor's precision. Depending on the situation, one, many or no configuration could be accepted. An accepted configuration contains the perceived information about the environment expressed in terms of the model of the world.

A fundamental aspect is the validation step. The projection to the senses ensures that an accepted interpretation is indeed compatible with the sensed information (and not just the more probable one given the current knowledge) and will reject a configuration that is not. It is crucial for an agent acting on the environment to have reliable perception, even incomplete.

To give a simple example, the model of the world could be three dimensional Euclidean space where flat surfaces, spheres and cubes are possible. A configuration of that model could be a sphere of a given size placed over flat ground along with the relative position of the observing agent. Let say the agent has an eye (or camera). Then the light source(s) must be included in the model. If the configuration "a cube over the ground" is to be tested, then the corresponding retinal image is simulated and compared to the measured one. If the difference lies within the same order of magnitude as the noise level, then the configuration is accepted. On the other hand, if a flower is shown to this agent, no configuration will be coherent and thus the result will be no perception. Similar descriptions could be made for other sensors.

This does not implies that the configuration of the model has to be *viewed* by an homunculus (a little human) inside our brain. The configuration of the world is not a *re*-presentation as in a picture, but an *interpretation*. A perceived configuration, for example "two cubes and one sphere over the ground", is an interpretation in terms of objects and its relations of the near environment, and is expressed in the right way to be analyzed by the decision making procedures.

It is important to note that it is not possible to perceive something that is not included in the model of the world. For instance, to perceive that an object is round, "roundness" should be included in the model. The perception of new things is possible if they can be expressed in terms of existing concepts. For example, when I see Figure 1 I perceive an arrangement of a box, some tubes, some wires, maybe a microphone. My perception is made in term of concepts I know. But someone familiar with that object will see directly "an instrument to measure air quality". That kind of fractioned perception can eventually lead to learning a new concept. In the example, a new concept for that instrument could be added. Learning is a related but *different* process from perception.

Figure 1.

For some scenes multiple interpretations are possible. It may be useful to add selection criteria. One possibility is to select the simplest one. For example, given a valid configuration, adding an object in a hidden position will not change sensory information; in this case it is natural to prefer the minimal configuration compatible with the senses. In other cases, as in ambiguous figures, two or more interpretations are equally valid. With more context (that should be part of the configuration) it could be decidable, but in an isolated figure there is no way to select.

Sensing does not happen instantaneously, the information is obtained in the course of our actions. Also, we get information from many senses at the same time. The diversity of sensor types and timing is naturally handled here: for a configuration to be accepted, it needs to be validated by every sensor. The result is the integration of diverse information into one coherent configuration of the environment. The main reason to have a global model of the world is, precisely, to be able to integrate different information sources at different times into one coherent picture of the environment.

The information gained by perception is used to decide future actions. Part of the selected actions are aimed at direct vital needs: to escape, to eat, to reproduce. But actions can also be aimed at obtaining more information: turn the head toward some point of interest, walk to have a better point of view, move the hand to follow the shape of an object. All this is compatible with the proposition of this paper.

Perception and learning about the external world are different processes. Perception is a fast process that gets information about the configuration of the world now, while learning takes more time and gradually obtains information about persistent properties of the world. Part of the model of the world is common for many animals and was developed by natural selection over an extensive period of time. Another part is learned by each individual during its life. Good models of the world are crucial for useful perception. But this paper will focus on perception, assuming that the model of the world is known. We can think of it as analyzing an animal for just a few second or minutes, a time short enough so we can neglect any process of learning.

The proposal is unrealistic in its full form; it is computationally too expensive to be applied by animals or robots. It is more likely that animals use shortcuts to reduce the computational burden. For example, instead of projecting the entire configuration on each sensor, it could be performed part by part. Also, not every possible hypothesis can be tested, the search must be done by some *exploring* methods. Gestalt laws are examples of such methods in human perception. These heuristics should be regarded as approximations of ideal perception.

To illustrate the ideas, a complete implementation for a toy example in visual perception will be presented. The system's knowledge is restrained to superimposed squares of different sizes, positions and colors on a flat scene. The algorithm is able to find a correct interpretation of scenes made of squares, and finds a partial interpretation on a mixed scene. When no interpretation with squares is coherent (as for most images) the output is "no understanding".

RELATIONS WITH OTHER THEORIES OF PERCEPTIONS

The theory sketched in this paper is a variation of the "sense-data" theory [1]. When I see a tomato my perceptual system found that the configuration "a tomato in front of me" corresponds well to the sensed information. But I could not *directly* perceive the tomato, because there is *no tomato* in the physical world as the one I perceive. According to current physical theories, there is nothing like a red smooth surface. The "tomato" in front of me is almost empty and the concept of "red" does not exist in physics. Our model of the world and concepts like "tomato" are, strictly speaking, false; thus "tomato" can only be a mental concept. At the same time, our model is good enough to describe the world at the level of details that we need. As Arthur Eddington put it in a famous passage in the introduction to "The Nature of the Physical World":

> I have settled down to the task of writing these lectures and have drawn up my chairs to my two tables. Two tables! Yes; there are duplicates of every object about me – two tables, two chairs, two pens.
>
> [...]
>
> One of them has been familiar to me from earliest years. It is a commonplace object of that environment which I call the world. How shall I describe it? It has extension; it is comparatively permanent; it is coloured; above all it is *substantial*. [...]
>
> Table No. 2 is my scientific table. It is a more recent acquaintance and I do not feel so familiar with it. It does not belong to the world previously mentioned – that world which spontaneously appears around me when I open my eyes, though how much of it is objective and how much subjective I do not here consider. It is part of a world which in more devious ways have forced itself on my attention. My scientific table is mostly emptiness. Sparsely scattered in that emptiness are numerous electric charges rushing about with great speed; but their combined bulk amounts to less than a billionth of the bulk of the table itself. [...]

However, in a sense, perception is not indirect either. When we see a tomato there is no internal mental image of a tomato that should be then *seen* by an internal perceptual system. The result of perception is not a "picture" produced by a configuration of the world, but information about objects present in the near environment. There is no need of further steps, except for those that analyze possible actions and decide the best one.

Richard L. Gregory proposed that perception is essentially like the process of discovery by hypotheses in science [12, 5]. The proposal of this paper is roughly in this line of thoughts. Cognitive systems function essentially like science; but perception is only one part of it. Here, perception is just the process of finding the configuration of the world that matches the sensory information; the process of learning about the world – in the analogy a substantial part of the discovery by science – would be a different but related process. A second difference is that Gregory see in perception a Bayesian strategy, while, as will be discussed, this paper proposes a different validation mechanism.

Another important theory of perception was introduced by James J. Gibson, [4, 12]. He claims that perception is an exploratory process of the external environment, where information is sensed in the form of structural invariants. He pointed out important critics about other theories. For example, information is not only present on the sensory measurement; much is added when analyzed in conjunction with the configuration of the body and its actions. Also, the senses act more like sampling devices than like a photographic camera that gives a fixed image. However, the main point of his theory is the *direct* character of perception in the sense of not being mediated by visual sensations or sense data. This is barely achievable in the proposed framework: Imagine someone in complete darkness touching with one hand a large plate. The observer will never be in contact with the whole object; in spite of that, she or he would conclude that the object is round by integrating the information in time. The concept of "roundness" must be involved in this process.

Recently, J. Kevin O'Regan and Alva Noë put forward a "sensorimotor" theory of perception [13, 11]. According to them, perception is a way of acting, a way of exploring the environment. Perception happen when the organism masters the sensorimotor laws that relates how the actions of the organism affects the sensory input. The proposal of this paper is similar to the theory of O'Regan and Noë with a difference in point of view. Understanding the environment also imply an understanding of how actions affect the sensed information. But here, the sensorimotor laws are implicit in the model of the world and the models for the senses. When models are good enough, action leads, in the short term, to expected configurations, that should correspond well to expected sensory information. This is equivalent to having sensorimotor laws for the different perceived facts. However, that would only be important when it fails, as that would be a sign that learning is needed.

Another difference is that in the proposed framework learning is regarded as a different process from perception, while for O'Regan and Noë perception *is* a kind of learning, where the sensorimotor laws are deduced.[1] The reason to distinguish perception and learning is that most of the time an adult is perceiving, no learning process is involved. Moreover, perception is involved when one listen to recorded music or see a film, but no relevant action is involved, and no sensorimotor learning. The learning process involved in obtaining models of the world and for the senses require processes of deducing and adjusting models for space and sensors (for example as presented in [15]), as well as processes for learning new concepts. The learning process is fundamental, and far from trivial, but it will not be analyzed in this paper.

Bayesian theory is a widespread mathematical theory of perception, see [10]. It involves a probabilistic model of the world where some variables describe the configuration, x, and other variables represents observations, y. The model provides a probability distribution $P(x)$, and the law concerning observations, $P(y|x)$. Given all this, the perception process consist in finding the configuration x that correspond to an observation y. In Bayesian theory, knowledge about the environment is given by the so-called *posterior* probability $P(x|y)$, which can be computed using Bayes' rule:

$$P(x|y) = \frac{P(y|x) \cdot P(x)}{P(y)}.$$

The factor $P(y)$ does not depend on x and just implies a multiplicative factor. Finally the perception depends on the sensor's law $P(y|x)$ and in the *prior* probability $P(x)$.

The general schema of Bayesian theory is similar to the one presented in this paper. The details, however, are different. Both theories need a model of the world, but the nature is somewhat different. Bayesian theory ask for a *prior* distribution that gives to some configurations of the world "more probability" than others. The "sensing" or "projecting" rules are also different, even if related. However, the real difference is the nature of the result. In its pure form, Bayesian theory gives the perceived information as a probability distribution on the configurations of the world. A way of using this information is needed. Often the configuration that maximize $P(x|y)$ is selected as perception. This last criterion always produces a perception and only one. With insufficient knowledge, the best configuration may produce inconsistent perceptions, and in an ambiguous scene only one configuration would be perceived. As we will see, the validation criterion proposed here is different. The result could be one, many or no configuration of the environment. And an accepted configuration must be coherent with the sensed information. Finally, Bayesian theory implicitly assumes that all possibles configurations x will be evaluated (or that the problem is sufficiently regular to be able to use shortcuts). In this proposal it is explicitly assumed that it is not possible to be sure that the best configuration was found because its

[1] They mention the use of ready made sensorimotor patterns that avoid learning all every time; in a sense these patterns are like concepts in the model of the world.

space is huge and not regular. Thus, a configuration exploration process is needed and its quality affect the result.

A THEORY OF PERCEPTION

Lower animals possess limited capacity for action and simple sensors that give them some information about the environment. A set of special detectors triggering ready-made actions probably guide most of their behavior. On the other hand, higher animals' range of actions is large and possess sophisticate sensors. More important, the life of each being is a big investment; thus, it makes sense to spend energy in selecting good actions to try to preserve its life.

To be able to make useful global decisions, an animal must integrate all the information about the near environment into *one* coherent structure. Moreover, higher animals probably possess a mental mechanism that is isomorphic, at some level of details, to the near environment. That mechanism let them try in their mind actions before actually take them. Like a chess player, they look ahead some moves before deciding its action.[2] In this view, that mechanism includes a model of the world, and perception is the process that keeps track of sensory information and tries to have, at each time, a configuration of that model that is isomorphic to the near environment. The result of perception is not "a picture" of that configuration of model of the world; the result of perception is the set of objects and relations between them described in the accepted configuration of the world. The configuration is then used to test and evaluate possible actions.

Let us call S the sensory input. Depending on the animal or robot, and on the level of description, the information S can be expressed in different ways. For example, it could be a vector $S = (s_1, \ldots, s_n)$, where s_i is related to the activation of the neuron number i in the input system. Or s_i could be the value measured at position i in some sensor. Also, different sensors may have different timing. The particular coding or organization for S is not important here.

Given a model of the world $\mathcal{E}$, perception is a process by which a configuration of it E is selected, hopefully one that corresponds at some level of accuracy with the environment. This configuration should include the pose and position of the observer, a list of object with some properties, their position relative to the observer, etc. Again, the details of the model of the world and the particular coding for E or organization of $\mathcal{E}$ are not important. Only things and properties included in the model $\mathcal{E}$ will be able to be coded in E.

Perception also depends on models for the sensors. This information is needed to build a rule of projection to the senses,

$$\hat{S} = \text{Project}(E),$$

which, given a configuration of the model of the environment E, gives the expected measurements on the senses $\hat{S}$. For example, if E represents a cube over a plane, and the sensor is a camera, $\hat{S}$ should be the image formed by the projective transformation from the point of view indicated in E (the observer is considered to be part of the environment, so its position and pose is part of E).

Given the observed information S, the process of perception tries to find a configuration E for which $S \approx \hat{S}$. There are some reasons that make that it would be rarely the case that S and $\hat{S}$ are actually equal. The sensors, as any measurement system, are subject to some accuracy and precision. So S could be slightly different from the expected values even if E is the right one. Also, the model of world is surely not perfect.

To cope with this, the perception process also needs some information σ_s about the precision of the senses. Different senses have different precision; so σ_s could be a vector $(\sigma_1, \ldots, \sigma_n)$, σ_i being the precision of s_i. But again, the detail are not important here.[3] As a first approximation we will say that an interpretation E will be accepted if

$$|S - \hat{S}| < \sigma_s.$$

This validation method is, however, insufficient as the next section will discuss. The validation method is a fundamental point as it determines when a perception is valid and when not. It is not just the best interpretation that is perceived; an interpretation must *also* be good enough to match the sensed information.

What configurations E should be tried? A brute force mechanism would test every possible E. Nevertheless, that would require too much time (infinity?) to get a valid perception. Some shortcuts may be possible by exploiting restrictions of the environment or the senses. However, in the general case, an *exploration* mechanism based on shortcuts and heuristics is needed. This mechanism can be iterative, using previous (successful or not) tested configurations and the information of the error committed, to refine into a new guess. The exploring mechanism is the more difficult mechanism in perception.

Algorithm 1 suggest a perception loop that tries to analyze the sensory information S. This is an extremely simple version just to illustrate the idea. In a more realistic one the loop may also stop after a certain time or when new sensory information is available. The process of evaluating actions should also be considered here.

Algorithm 1: Perception

repeat
  $E_{old} \leftarrow E$
  $\hat{S}_{old} \leftarrow \hat{S}$
  $E \leftarrow \text{ExploreConfiguration}(S, E_{old}, \hat{S}_{old})$
  $\hat{S} \leftarrow \text{Project}(E)$
until($|S - \hat{S}| < \sigma_s$)

[2] What we call consciousness is probably a part of that mechanism in which the known information is evaluated, some "moves" are played in the mind, and action is decided.

[3] σ_s may be dependent on S. For example, in low light conditions visual information is much more noisy.

THE VALIDATION STEP

The validation step is crucial as it determines the result of perception. This section will show that the criterion presented in the last section is insufficient. The desired properties for a satisfactory one will be described and a way to a solution suggested.

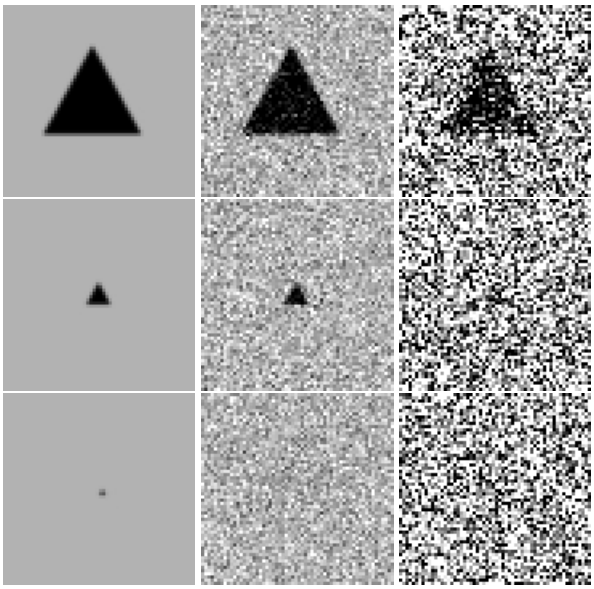

Figure 2. The validation step should guarantee that there is enough evidence to support a configuration. The nine images shown here were produced by drawing black triangles of different sizes over a gray background, and then adding of noise. In each case the observation is compatible with the presence of a triangle under noise, but only in some of them the evidence is enough to accept that interpretation.

As pointed in the last section, a configuration should be accepted when the sensed information is compatible with it, given the knowledge of the sensors used. But that is not enough. Figure 2 illustrates the problem. The nine images shown in the figure are compatible with the hypothesis of a black triangle over a gray background plus a certain quantity of noise (different on each column). Only in some of them, however, evidence is enough to accept that interpretation. On the first row we see the same triangle with increasing noise added. In the first case the triangle is clearly visible; in the second image the triangle is still visible, but we notice a strong noise. In the last case, the hypothesis of the triangle seems as good as many others. A similar effect but in another sense happen when we analyze the first column, where no noise is present but the triangles appear at decreasing sizes. In the last image (left-bottom) the "triangle" is so small that there is not enough *evidence* to support that hypothesis. A good validation criterion should accept a configuration only when the measurements are compatible *and* the evidence is enough to support the hypothesis.

In [2] Desolneux, Moisan and Morel introduced a theory of detection that satisfy the desired properties. Their method is based on a what they claim is a general perception principle, the Helmholtz principle, according to which an observed geometric structure is perceptually meaningful when its expectation in noise is small, see [3]. Detection is treated as a simplified hypothesis testing problem. In the classic decision framework, two probabilistic models are required: one for the background and one for the objects to be detected. In the *a contrario* approach (as it is called), however, the objects are directly detected as outliers of the noise model.[4] No probabilistic model is needed for the objects.

Figure 3 shows the result of line segment detection on the images of Figure 2 by LSD [7], an algorithm based on Desolneux, Moisan and Morel theory. The detection criterion did a good job giving detection only in cases where evidence is enough.

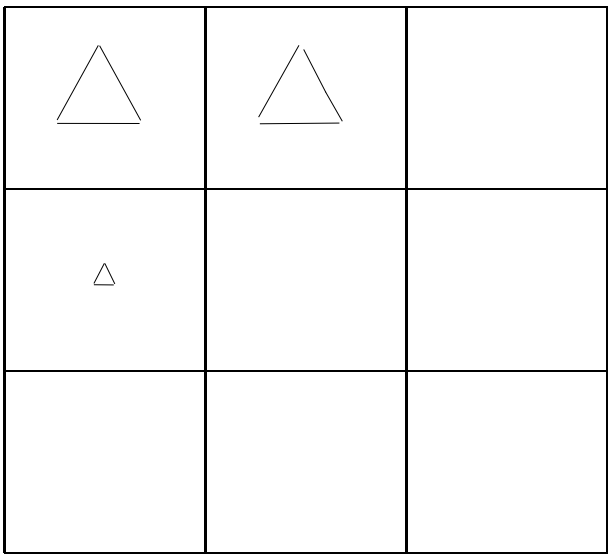

Figure 3. The result of applying LSD, a line segment detector, to the images of Figure 2. As desired, the algorithm only produces detection when there is enough evidence for that.

Desolneux, Moisan and Morel theory was successfully applied to many detection problems [3] and more are envisaged. However, the formulation of a general criterion for perceptual validation needs further development.

REMARKS

The theory sketched makes a clear-cut difference between *perception* and *learning*. Perception is the process by which the sensory information is interpreted in terms of an already present model of the environment; learning is the process by which the model of the environment is created or updated. Perception and learning are related. Perception clearly depends on learning for the model of environment. Part of the learning process happened by natural selection over huge periods of time. More learning and fine adjustments of the model also occurs on each individual since the beginning of its life.[5] Learning processes also use the perceived infor-

[4] The article "A statistical information theory of visual thresholds" by Violet Cane and R.L. Gregory in [6] suggests that perceptual thresholds are adapted according the noise levels in the senses. Desolneux, Moisan and Morel also performed psychophysical experiments that give support to their theory, [3].

[5] In [15], Philipona et al. presented an algorithm to learn information about the world by directly analyzing the sensory information resulting from random actions.

mation as raw material: a pattern of already known elements can be assembled into a new concept when repetitive appearance or inferred meaning suggest its utility. But learning is not analyzed in this paper.

Note that the model of the world needs to be accurate to some degree in order to be useful. However, a model do not need to be perfect to be useful. For example, Euclidean space is good enough for any animal needs. Yet, according to the general relativity theory it is only an approximation. A reasonable balance between the complexity and accuracy of the model of the world (including models of the senses) is needed. A model too accurate implies more energy spent in the perceptive system; an inaccurate one entails more prediction errors, thus poor decisions. The right balance is adjusted in each species by natural selection.

Different sensors provide different kind of information. The projection rules must deal with the particularities of each one of them. This includes different type of information projected, different timing, different dynamic, etc. The configuration perceived must match all of them, providing a way of fusion for the heterogeneous sensory information. Also, the environment is usually not static; accordingly, the configuration of the world has to include dynamic aspects. For example, if the scene is a sphere rolling, what is perceived is not a series of static configurations of the sphere at different position; the configuration perceived is a sphere *moving* in some direction and at some speed. The projection to the senses must consider the dynamic aspects.

The objective of perception is to get information about the environment to be able to decide actions. Among the possible actions are some that will improve the information of the environment. Thus, perception can guide action to be able perceive, to be able to act. For example, one can turn the head to point to a previously unseen point. Or one can walk to approach something in order to see details or be able to touch. Algorithm 2, suggest a variant of the perception loop including this kind of action. The process "ExplorativeAction" selects an action A to be done, based on the current perceived configuration E and evaluating the error between S and $\hat{S}$. The criterion for selecting A is not discussed here.

Algorithm 2: Perception with Explorative Action

do forever:
  $S \leftarrow$ New Sensory Information
  $E_{old} \leftarrow E$
  $\hat{S}_{old} \leftarrow \hat{S}$
  $E \leftarrow$ ExploreConfiguration$(S, E_{old}, \hat{S}_{old})$
  $\hat{S} \leftarrow$ Project(E)
  $A \leftarrow$ ExplorativeAction$(E, S, \hat{S})$

Perception is an *interpretation* process, and as such, it can makes mistakes. The perceived scene is at most *one* possible interpretation. Human perceptual system is very good, and the information sensed very redundant, hence mistakes are rare. We call it illusions or hallucinations when perception goes wrong. During an illusion, our perception system found an interpretation that is coherent with the sensory information, so it *could* be true, but it is not. Hallucinations can arise in different ways. They could be rooted in a bad or insufficient model of the world, or caused by inadequate sensor projection rules, as in Figure 4. Another possible reason is validation thresholds too permissive, that would accept any proposed configuration.

Figure 4. The pencil seems to be broken. Our projections rules does not consider the refraction of light at the water/air transitions. Thus, the perceptual system uses projections rules that would be well adapted for an empty glass; in that case the result would have been correct, as only a broke pencil could produce that observation.

Ambiguous illustrations, as in Figures 5, deserve a comment. These figures were conceived in purpose to have more than one interpretation. In contrast with illusions, here none of the interpretations is wrong. When people are shown this kind of pictures they usually see only one interpretation and stop there unless they are encouraged to continue searching for more. Normal environment is generally very complex and sensory information redundant enough that is hardly possible that a wrong configuration match; so it is reasonable to stop the exploration when a valid configuration is found.

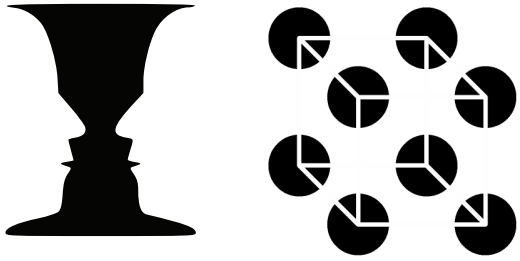

Figure 5. Two images with multiple interpretations.

In some cases no interpretation can be found for a scene. Even if rare, this happen to humans too; it produces a sensation of not being able to figure out what is being seen. Usually this situation ends when a clear interpretation appear, either spontaneously after a change of view, or suggested by someone.[6]

In a sense, every scene can have multiple interpretations: given a valid interpretation, a new one can be built by hiding

[6]This happen often in language perception, and particularly when exposed to a foreign language not perfectly mastered: sometimes we hear something that we fail to understand; often, the meaning appears clear after some minutes of reflection.

an element behind an object. This new interpretation would be as valid as the first one. However, it seems sound to apply Occam's razor and keep the more simpler interpretation, the one that involves less elements (and specially the smallest number of unobserved elements).

The most difficult part of perceptual systems is probably the mechanism for exploring the space of configurations in search for a good interpretation. Provided with the right interpretation the validation system will acknowledge it. But to spot it in limited time among a huge number of possibilities is a difficult task. Human mechanisms are extremely good. In challenging conditions, however, the search for configurations is less effective, as illustrates the famous picture shown in Figure 6. When people see this picture for the first time, usually fail to get a full understanding, or it takes considerable time. Yet, when signaled that a Dalmatian dog is present at the center, with the lowered head facing away and left, most people manage to see it well. This suggests that the problem is rooted in the exploration of configurations rather than in the validation.

Figure 6. What do you see in the picture?

The first stages of the visual system and its specialized mechanisms are probably part of the configuration exploring system, as well as the mechanisms described by Gestalt theory [8, 9]. The structures and patterns detected are used to restrain the search. Geometric features help to organize sensor information: occlusions are detected, projective geometry inferred, isolated information is grouped, etc. There are probably some specialized detectors, for example for faces. Gradually, the configuration of the world is unraveled.

Ideally, a configuration to be validated should be fully projected to the senses. This is probably unrealistic as it is too expensive to be done in animals or robots. Some shortcut are likely to be needed. The heuristics needed to handle the computational burden should be regarded as approximations, as good as possible, to ideal perception. For example, a configuration may be projected to the sensors, object by object in a modular way. Or the projection can be somewhat local. That could also help in the gradual process of discovering the configuration, by validating already found elements. Projection by parts would also make possible the perception of impossible objects [14] as the ones shown in Figure 7. There is no object that corresponds as a whole to these pictures. But parts of these figures do correspond to parts of objects. When we try to perceive these pictures as a whole the result is flat; when looked by parts, perception changes as we see different parts of them.

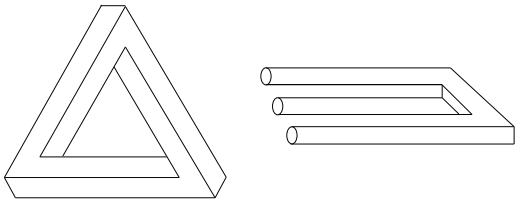

Figure 7. Impossible objects: the Penrose triangle and devil's tuning fork.

A TOY EXAMPLE: SQUARES

Perception of Fixed Images

The following experiments deals with fixed images. It is often argued that to analyze one fixed image is a far cry from the process of perception. Certainly, human or animal perception is much more rich. In natural life, animals never analyze *one* fixed image. To start with, animals usually have two eyes and process a flux of visual information. Also, the image on the retina is irregularly sampled, and needs constant eye movements to cover even part of a scene. Moreover, we are rarely exposed only to visual information. More often we can hear, touch, smell and taste at the same time. And more information is obtained by acting. True.

But this does not means that perception is *only* possible with all these capacities. A blind person can perceive even if visual information is missing; a deaf person can perceive without hearing; a paraplegic person can perceive with very limited capacity for moving. Reducing the information sources makes perception more difficult, less efficient and limited. But it is still perception. We are all capable of perceiving music by listening to records, while the only relevant stimulus is auditory.

The objective of computer vision is to understand and duplicate natural vision by computer algorithms. There has been some research using on-board cameras and computers on robots capable of action. But most of the work is done on fixed images; the reason is just the simplicity of the setting needed to work. It is clear that a lot of information is lost and many of the problems addressed by computer vision are much harder than they could be with more redundant information. However, there is still much to be learned by the analysis of fixed images. Humans have no problem understanding photos and we could expect that algorithms could understand at least a small fraction of what humans can with exactly the same information: the one contained in a fixed photo.[7]

[7] It is hard to believe that, for example, eye movements is fundamental to understand a photo and not just part of how human visual system works. Computer vision algorithms often use an equivalent mechanism to eye movements when they detect edges and spend more time analyzing the regions of the image near them.

A full-fledged perception system, useful for an animal, probably needs many sources of redundant information and the capacity for action. But we can use fixed image perception as a very simple case of study, as a test bed for perception. This somewhat isolated problem is to perception as the pendulum or the canon balls are to physics. The very same fundamental problems are involved as in general perception.

A World of Squares

To have a perceptual setting we need a model of the world and models for the sensors. Here we will set a very simple context so as to be able to show a full implementation. This will be, no doubt, a toy example, but hopefully one that will illustrate the ideas.

The "world" known to our system is composed of squares of different sizes and colors, superposed one to another on a flat scene. The sensor gets a digital image of the scene by sampling the flat world. It is assumed that the sensor introduces some blurring and noise.

The exploring mechanism starts by applying a line segment detector, LSD [7], to the image. Then, every combination of four line segments is tested looking for possible sets compatibles with a square. That is, two pairs of parallel line segments, both pairs roughly equally spaced, and one pair orthogonal to the other. The criteria used to decide if two line segments are parallel, orthogonal, or equally spaced are arbitrary but reasonable. A last condition required is that each of the sides is covered, in part, by one of the line segments. The result is a set of candidates for square.

The second step tries to build a configuration of candidates for square, with a certain order and color for each one, that would produce the sensed image. For this aim, a mask of occluded pixels is used. At the beginning the mask is empty as all the pixels are visible. Then, each one of the candidates is tested. If one of them has a more or less constant color, it is temporary accepted as visible. Thus, it is stored as the first square and its pixels are marked as occluded in the mask.

The same procedure is repeated in the search for a second square, then a third square, and so on. But, when checking if candidates present a constant color, only pixels that are marked as visible in the mask should be considered (the rest being occluded by previous squares). To be accepted, a candidate must also have a minimal number of visible pixels, otherwise we could accept completely occluded squares.

This process continues iteratively until there is no candidate left or there is no candidate that can be accepted. In each iteration of this process, more than one candidate may satisfy the criterion of visibility and uniformity. Each time a multiple option is possible a new branch is added to the tree of configuration to be tested; all the nodes must be tested for validation.

The expected sensory information (the image) can be synthesized for given configuration: an ordered list of squares with its position, size, orientation and color. For this toy example, the synthesis is trivially obtained by drawing the squares into an image. The validation step consists in comparing the synthesized image to the observed one. Pixels where the difference in color exceeds the noise level are considered as *not explained*. If all pixels are explained the configuration is valid. Inversely, if most pixels are not explained the configuration is rejected. Between this extremes, a configuration could be partially valid when some of its squares are fully covered by explained pixels (these are *explained squares*).

Among the valid or partially valid configurations, the ones with more explained pixels are selected. Among them, the ones with less unexplained squares are kept. And among them, the one with the smallest number of squares (in a sense, simpler) is considered the best interpretation. As the experiments of the next section will show, further research is needed to get a good criterion.

In this simple case of squares, the "evidence" problem mentioned when discussing the validation step is handled by the use of LSD, a line segment detector based on Desolneux, Moisan and Morel detection theory. Candidates for square are built based on the detected line segments; therefore a validation criterion with the right properties is implicitly used. Otherwise, unsatisfactory configurations could be accepted. For example, tiny squares could be validated, and any image could be interpreted as composed of tiny squares of the right color, as when images are represented by pixels. This is not possible when using LSD, the side of a pixel is too small to be detected and no candidate for it will be generated.

The selected configuration of explained squares is the perception.

Experiments

This section shows experiments made with the implementation described before. A first set of experiments is shown in Figure 8. For each row, the first column shows the input image, i.e., the sensory information. The middle column shows the synthesized image from the best configuration found, that is, the re-projection into the senses from the obtained configuration. The last column shows on black the pixels that were rejected.

On the experiment on the first row the scene can be interpreted as a supposition of squares, and this is what the algorithm found. There is some noise in the input image and some imperfections in the size and position of squares on the synthesized image. For that reason both images are not identical, but the difference is small enough to be accepted. No pixel was rejected. The perceived configuration is a reasonable interpretation of the scene. It is important to note that the result of perception *is not* the synthesized image but the information about the set of squares and its relation. For example, a described could be "a white square of size A and angle θ, at position (X, Y); behind it, a light gray square of size…". For this scene two interpretations are equally valid as in the world known to the system only one square is possible per depth level. In one interpretation the large dark gray square is behind the large light gray one, and in the other

interpretation is the opposite.

On the second row we see almost the same scene as before, except for some dark marks. The best configuration found is the same as before, but this time the algorithm found that there are some zones of the sensory input that were not explained. What we get is an incomplete perception of four squares and a zone yet to be explained. The scene of the third row is still the same with different dark marks. In this case the algorithm manage to understand two squares (that are detected without rejected pixels). The algorithm fails, however, to understand the light gray square with marks: due to the dark marks, it is not, indeed, a constant color square, the only element know to the system. As a consequence, the black square is not understood either. The algorithm was programmed to assume the background to be uniform, so it is set to the mean gray level of the remaining pixels. In this case we should say that the algorithm partially perceived the scene, understanding only two squares. For similar reasons as before, the algorithm fails to make sense of the scene on the last row. No pixel is validated and no perception results.

Figure 9 shows an experiment on a natural image. As expected, the algorithm do not manage to make sense of it: no configuration of squares corresponds to it. The synthesized image is uniform of the mean color of the image. Then, almost all pixels are rejected, except for some of them that have more or less the mean gray value. The scene is not understood and the result is no perception.

Even if the setting of squares is very simple, some interesting configurations are possibles. Some of the visual experiments developed by the Gestalt psychologist can be duplicated using only squares [8, 9]. Two of them can be seen in Figure 10. The first experiment was created to study how objects are ordered in the presence of subjective contours. Is the triangle in front or behind the square? One hypothesis is that vision choose, other conditions being equal, the configuration that requires less subjective contours. In the version made of squares, the algorithm gives both results equally valid. But the subjective contour criterion could be implemented and tested extensively and different variations compared.

Figure 11 shows two configurations found by the algorithm for second experiment: the Kanizsa square. The resulting squares were drawn in different grays to indicate the depth: light means near foreground, dark near background. The two configurations shown are equally valid, the only difference being the order of the two foremost squares. Both squares are white in the interpretation, thus the contours between them are subjective and both configurations produce the same image. In fact, there are many more configurations that are equally valid; for example, changing the relative depth of the four small squares will not change the synthesized image. In the original figure, the big horizontal square is perceived as a white square with a black border. Our algorithm only knows solid squares without border, so it can only make sense of it as two superposed squares, one black (that makes the border) and one white. For this reason

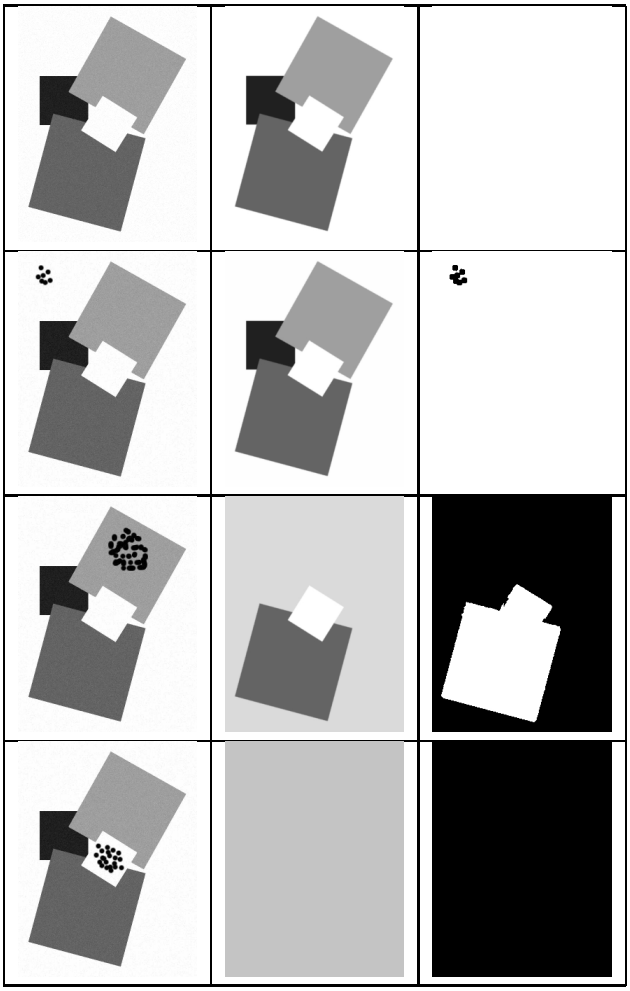

Figure 8. An experiment with pictures made with four squares and some dark spots. Left column: the observed image. Middle column: the synthesized image from the perceived configuration. Right column: validation mask, white pixels where accepted, black pixels where rejected.

Figure 9. House image experiment. As before, at the left is the original image; in the middle is the synthesized image from the perception; at right is the mask of validated pixels (white means validated; black means rejected). As expected for this image, no configuration of squares can reproduce the image. Thus, the synthesized image is just constant, and almost all pixels are rejected.

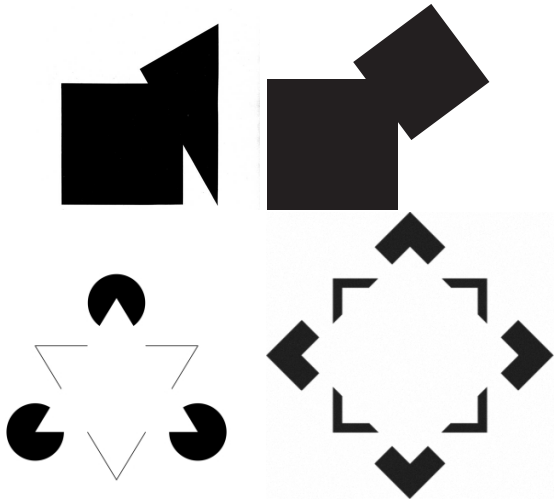

Figure 10. Two visual experiments of Gestalt theory. Left: original picture. Right: version using only squares.

too, the configuration on the right is valid, and equally valid to the other one.[8] According to this theory, the Kanizsa triangle is perceived because is the simplest configuration that makes sense of the figure. A detailed study of this kind of experiments could lead us to a better understanding of the factors involved in each one of the figures resulting from Gestalt theory.

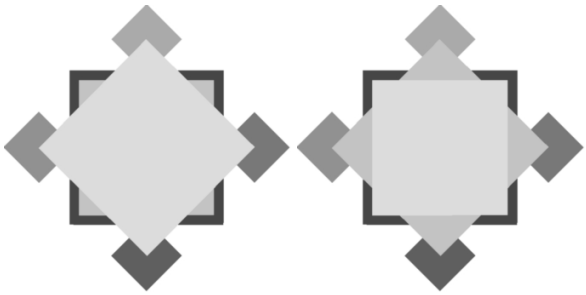

Figure 11. Two possibles configurations of squares that explain the Kanizsa square shown in Figure 10. Different shade of gray are used to represent the depth of each square: light is on the foreground, dark on the background. The difference between the two configurations is the order of the two foremost squares (the two squares that should be white). A priori there is no reason to prefer one configuration over the other.

CONCLUSION

This paper sketched a computational theory of perception. More work is needed to go from the outline presented here to a complete theory. A full implementation was presented for a toy example. Even if simple, the implementation can be used to study a subset of the rich family of experiments created by the Gestalt psychologist. Moreover, a simple extension of the model would allow to handle a larger subset of Gestaltic experiments. This setting provides a simple test bed for developing and experimenting, hopefully leading to a complete formal theory of perception.

[8] As a curious comment, the criterion of minimum subjective contours will give the standard interpretation.

ACKNOWLEDGMENTS

I would like to thank collaborators and friends for many remarks and suggestions, and more particularly to Pablo Arias, Gabriele Facciolo, Jérémie Jakubowicz, Gloria Haro, Stacey Levine, Maria Mavris, Enric Meinhardt-Llopis, Jean-Michel Morel, Pablo Musé and Gregory Randall.

REFERENCES

1. A.J. Ayer, *The Central Questions of Philosophy,* Penguin Books, 1973.
2. A. Desolneux, L. Moisan, and J.M. Morel, *Meaningful alignments*, International Journal of Computer Vision, 40(1):7–23, 2000.
3. A. Desolneux, L. Moisan, and J.M. Morel, *From Gestalt Theory to Image Analysis, a Probabilistic Approach*, Springer, Interdisciplinary Applied Mathematics series, vol. 34, 2008.
4. J.J. Gibson, *The Senses Considered as Perceptual Systems,* Houghton Mifflin Company, 1966.
5. R.L. Gregory, *Eye and Brain, the psychology of seeing,* World University Library, 2ed., 1972.
6. R.L. Gregory, *Concepts and Mechanisms of Perception,* Duckworth, 1974.
7. R. Grompone von Gioi, J. Jakubowicz, J.M. Morel, G. Randall, *LSD: A Fast Line Segment Detector with a False Detection Control*, IEEE Transactions on Pattern Analysis and Machine Intelligence, 19 Dec. 2008. More information and an on-line demo at `http://mw.cmla.ens-cachan.fr/megawave/algo/lsd`
8. G. Kanizsa, *Grammatica del vedere,* il Mulino, 1980.
9. G. Kanizsa, *Vedere e pensare,* il Mulino, 1991.
10. D. Mumford, *Pattern Theory: the Mathematics of Perception,* Proceedings of the International Congress of Mathematicians, Vol. I, 401–422, Beijing, 2002.
11. A. Noë, *Action in Perception,* The MIT Press, 2004.
12. A. Noë & E. Thompson (Eds.), *Vision and Mind, Selected Readings in the Philosophy of Perception,* The MIT Press, 2002.
13. J.K. O'Regan & A. Noë, *A sensorimotor account of vision and visual consciousness,* Behavioral and Brain Sciences, vol.24, pp.939-1031, 2001.
14. L.S. Penrose & R. Penrose, *Impossible Objects: A Special Type of Visual Illusion,* British Journal of Psychology, 49(1):31-3, February 1958.
15. D. Philipona, J.K. O'Regan, J.P. Nadal, & O.J.M.D. Coenen, *Perception of the structure of the physical world using unknown sensors and effectors*, Advances in Neural Information Processing Systems, vol.15, 2004.