The Complexity of Perceptual Search Tasks

John K. Tsotsos

Department of Computer Science,
10 King's College Rd.,
University of Toronto,
Toronto, Ontario, Canada M5S 1A4

Abstract

This research was motivated by the following question: What is the inherent computational difficulty of the visual search experimental paradigm in psychology? This is an important issue since so many computational models have recently appeared that use visual search data as motivation. How can these results be properly used if the computational nature of the experiment itself is not understood? A computational definition of the visual search task is presented, and the bottom-up case is distinguished from the task-directed case. Then, a proof is given showing that the bottom-up case is NP-Complete in the size of the image, while the task-directed case has linear time complexity in the number of items in the display. The NP-Completeness of the bottom-up case is due solely to the inability to predict which pixels of a test image correspond to objects in a non-exponential manner. This provides the strongest possible evidence for the abandonment of purely bottom-up schemes that address the full generality of vision. It is thus necessary to sacrifice generality in order to re-shape the vision problem and to optimize the resources dedicated to visual information processing so that a tractable problem is addressed.

Introduction

In this paper, the general task of visual search will be shown to be inherently intractable in the formal sense. Given the ubiquity of visual search tasks in everyday perception, it may be true that visual perception in general is also intractable. Yet, human vision is an effortless and exquisitely precise sense. How can this be? Ancient philosophers were aware of the fact that humans could attend to the relevant and ignore the irrelevant. More recently, psychologists have studied attention, and have proposed that there must be some kind of processing limit in the brain to lead to such a phenomenon. Helmholtz claimed that a conscious or voluntary effort may focus attention on a particular spot in the visual field [Helmholtz 1925], and this led to the 'attentional spotlight' idea that is widespread in models of perception. Neisser first claimed that any model of vision that was based on spatial parallelism alone was doomed to failure, simply because the brain was not large enough [Neisser, 1967]. Stating that the brain is simply not large enough does not yield any useful constraints on tne architecture of the visual system. Arguments such as this, namely that a given fixed resource is not large enough to accommodate a specified problem lead naturally within the computational paradigm to consideration of computational complexity. Neisser's claim hinted at the difficult issues of computational complexity that must be addressed, even though when it was made complexity theory was barely in its infancy.

Complexity considerations are commonplace in the computational vision literature. Many researchers (for example, [Mackworth and Freuder, 1985], [Poggio, 1982], [Grimson, 1984] and others) routinely provide an analysis of the complexity of their proposed algorithms - this is simply good computer science. It is important to demonstrate that specific algorithms have tractable requirements in terms of computer size and execution time. But this is not the same as addressing the complexity issues of vision in general.

The task of visual search has not been defined in computational terms by the psychology community. According to the definition provided in Rabbitt [Rabbitt, 1978], a visual search task is a categorization task in which a subject must distinguish between at least two classes of signals, goal signals which must be located and reported and background signals which must be ignored. This definition does not specify how signals arc located, how signals are represented, nor how goal and background are distinguished. In most, but not quite all, experiments of this type, the subject knows the goal before viewing the stimulus. The goal is most often simply shown visually, although in some experiments verbal descriptions are given instead. Visual search

experiments typically measure response times for the categorization of signals. Categorization yields a yes/no answer; other types of experiments ask subjects to report on other aspects of the stimulus as well, but these are not considered in this paper. In experiments where the subject knows the goal before viewing the stimulus, observed response times give linear functions of the size of the display (or number of items in the display). Moreover, the search is a linear one because the slope of the negative case is twice that of the positive case. The literature documenting this is large (see [Nakayama & Silverman 86] or [Treisman 88] for example). The data on response time characteristics if the goal is unknown is less conclusive and rather sparse. Treisman and Sato report that unknown goal search produces overall increase in response times, but some show no changes in slope while others show large changes [personal communication]. It is as if features were inhibited in several successive passes, allowing unknown goals to emerge at different points in the search sequence. The former experimental setup corresponds to my definitions of Task-Directed Visual search while the latter is Bottom-Up Visual Search, both of which are presented later.

Response times are intimately connected to the speed of processing, the amount of processing machinery devoted to the task and the algorithm that is used to solve the task. The study of computational complexity measures the cost of solving problems, and is thus it is natural to ask whether the tool of complexity can be productively applied to the understanding of visual search performance in humans.

An earlier paper describes some of the possible approximations and optimizations that human vision may be using [Tsotsos 1988a] while a companion paper provides supporting neurophysiological and psychophysical evidence for this position [Tsotsos 1988b]. It should be noted that there is nothing inherent in the problem or the proof procedure that prohibits the applicability of the proof for signals of a different dimension or modality, for example, speech. It will be claimed therefore, without explicit proof that auditory or other perceptual search tasks are also NP-Complete for bottom-up strategies and have linear time complexity for task-directed strategies.

A Computational Definition of Visual Search

The general version of visual search seeks to find the subset of the test image that matches the goal, using some definition of 'match', and in its full generality includes the possibility of noisy or partial matches. The problem is viewed as a pure information processing task, with no assumptions made about how the data may be presented or organized. The description presented is an abstract one, and is not intended as an implementation level characterization. Further, the problem may be of arbitrary size and may use arbitrary stimulus qualities. Other characterizations of visual search may be possible; however, it is suggested that all useful ones will include the notions of images, measurements over images, and constraints that must be satisfied.

A test image containing an instance of the goal is created by translating, rotating, and/or scaling the goal, and then placing it in the test image. The test image may also contain confounding information, such as other items, noise, and occluding objects, or other processes may distort or corrupt the goal. Due to image discretization, there are a finite but large number of possible transforms for a given image: the goal may be translated to anywhere in the test image; the goal may be rotated about its origin by any angular amount; and, the goal may be scaled in two directions by arbitrary amounts. Since images are discretized, there are only a finite number of possibilities along each of these dimensions that would lead to distinct images. 2D spatial transforms are well-known.

The question posed by visual search has two variants: a bottom-up and a task-directed version. In the bottom-up case, the goals are either not known in advance or even if they are they are not used, except to determine when the search terminates. The task-directed case uses the goal to assist in optimizing the solution to the problem. Both will be addressed by the following analysis.

The solution to visual search involves solving a sub-problem, which we call Visual Match. Given a 2D spatial transformation of the goal, the Visual Match procedure measures the fit of goal to test image and makes a yes/no decision as to its suitability. Therefore, an algorithm for Visual Search may be the following:

1. For each goal:
   2. If transform hypotheses not exhausted then
      hypothesize a new 2D spatial transformation
   else exit this loop and return 'no'
   3. Apply transform to hypothesize location,
      orientation and scale of goal in test image
   4. Execute the Visual Match procedure
   5. If Visual Match returns 'yes' exit and
      return 'yes'
   else go back to step 2.

So, the input to the Visual Match procedure is two images (test and goal after a transformation) and a

method of measuring goodness of fit. In many experiments, no scale or rotation is present, so in such cases, the only parameter to hypothesize is location. The first theorem presented will be specific for bottom-up Visual Match. An instance of the Visual Match problem is specified as follows:

- A test image **I**
- A goal image **G**, modified using a 2D spatial transformation
- A difference function **diff(p)** for p ∈ **I**, **diff(p)** $\in R_\rho^0$

(R_ρ^0 is the set of non-negative real numbers of fixed precision ρ)

- A correlation function **corr(p)** for p ∈ **I**, **corr(p)** $\in R_\rho^0$
- Two thresholds, θ and ϕ, both positive integers

The task posed by Bottom-up Visual Match is:

Given a test image, a difference function, and a correlation function, is there a subset of pixels of the test image such that the difference between that subset and the corresponding subset of pixels in the goal image is less than a given threshold and such that the correlation between the two is at least as large as another specified threshold?

In other words, is there a set $\mathbf{I}' \subseteq \mathbf{I}$ such that it simultaneously satisfies

$$\sum_{\mathbf{p} \in \mathbf{I}'} \mathbf{diff(p)} \leq \theta \text{ and } \sum_{\mathbf{p} \in \mathbf{I}'} \mathbf{corr(p)} \geq \phi \, ?$$

One point regarding the above specification of visual matching must be emphasized. This definition is one which forces a bottom-up approach to the solution of visual matching. The goal image is not permitted to provide direction to any aspect of the computation other than in the computation of the diff and corr functions. The constraints given must be satisfied with subsets of the input image. It is not unexpected that this definition involves two constraints to be simultaneously satisfied and that the two constraints represent error and size satisfaction criteria. Note that this definition does not force acceptance of only the 'best' match, but accepts any 'sufficiently good' match. This is very similar to many other kinds of recognition definitions. For example, the standard definition of region growing involves the maximal contiguous subset of pixels that satisfies a given property [Ballard & Brown 82]. Moreover, this definition should not be interpreted as a template-matching operation. Although template matching may be posed in the above manner, there is nothing inherent in the definition to exclude other matching forms - the notions of image points, measurements, and constraints representing large enough size and low enough error are ubiquitous in visual matching definitions. Details on how this particular collection of data may represent the Visual Match problem follow.

1. A test image **I** is the set of pixel/measurement quadruples (x,y,j,m_j). x,y specify a location in a Euclidean coordinate system, with a given origin. Note that the locations are not necessarily contiguous. $\mathbf{M_i}$ is the set of measurement types in the image, such as colour, motion, depth, etc., each type coded as a distinct positive integer. m_j is a measurement token of type j, represents scene parameters, and is a non-negative real number of fixed precision, that is with positive error due to possible truncation of at most ρ. (Only a finite number of bits may be stored). $\mathbf{I}' \subseteq \mathbf{I}$ is a sub-image of **I**, i.e., an arbitrary subset of quadruples. It is not necessary that all pixel locations contain measurements of all types. Further, it is not necessary that the set of pixels be spatially contiguous. For ease of notation, $\mathbf{i_{x,y,j}}$ has value m_j. If $\mathbf{j} \notin \mathbf{M_i}$ or if the x,y values are outside the image array then $\mathbf{i_{x,y,j}} = 0$.

2. A goal image **G** is a set of pixel/measurement quadruples defined in the same way as **I**. The set of x,y locations is not necessarily contiguous. $\mathbf{M_g}$ is the set of measurement types in the goal image. The types correspond between **I** and **G**, i.e., type 3 in one image is the same type in the other. The two sets of measurement types, however, are not necessarily the same. The coordinate system of the goal image is the same as for the test image and the origin of the goal image coincides with the origin of the test image. $\mathbf{g_{x,y,j}}$ has value m_j. If $\mathbf{j} \notin \mathbf{M_g}$ or if the x,y values are outside the image array $\mathbf{g_{x,y,j}} = 0$.

3. For purposes of the proof, the diff function will be the sum of the absolute values of the point-wise differences of the measurements of a subset of the test image with the corresponding subset of the goal image. It is expressed as follows for an arbitrary subset **I'** of the test image:

$$\sum_{\mathbf{p} \in \mathbf{I}'} \mathbf{diff(p)} = \sum_{\mathbf{p} \in \mathbf{I}'} \left[\sum_{\mathbf{j} \in \mathbf{M_i}} | \, \mathbf{g_{x,y,j}} - \mathbf{i_{x,y,j}} \, | \right]$$

This sum of differences must be less than a given threshold θ in order for a match to be potentially acceptable. Note that other specific functions that could find small enough values of some other property may be as suitable. The threshold is a positive integer.

4. Since a null I' satisfies any threshold in the above constraint, we must enforce the constraint that as many figure matches must be included in I' as possible. 2D spatial transforms that do not align the goal properly with the test items must also be eliminated because they would lead to many background to background matches. One way to do this is to find large enough values of the point-wise product of the goal and image. This is also the cross-correlation commonly used in computer vision to measure similarity between a given signal and a template. A second threshold, Φ, provides a constraint on the acceptable size of the match. Therefore,

$$\sum_{p \in I'} \mathbf{corr(p)} = \sum_{p \in I'} \left[\sum_{J \in M_1} \mathbf{g_{x,y,j}} \times \mathbf{i_{x,y,j}} \right] \geq \phi$$

Note that there is no claim here that the algorithm necessarily corresponds to human performance. The function definitions are given primarily for purposes of the proof and are claimed to be reasonable ones. It is possible to provide other functions for difference and correlation and reconstruct similar proofs using them.

The Complexity of Bottom—Up Visual Matching

The bottom-up visual match problem as stated above has exactly the same structure as a known NP-Complete problem, namely the Knapsack Problem [Garey & Johnson 1979]. Therefore, it would seem that a direct reduction (by local replacement) of Knapsack to Visual Match is the appropriate proof procedure. The formal statement of Knapsack follows:

Knapsack
instance: Finite set U
for each u ∈ U there is a function s(u) ∈ Z^+
(the set of positive integers)
and a function v(u) ∈ Z^+
positive integers B, K
question: Find a subset $U' \subseteq U$ such that
$\sum_{u \in U'} \mathbf{s(u)} \leq \mathbf{B}$ and $\sum_{u \in U'} \mathbf{v(u)} \geq \mathbf{K}$

The following then is the main theorem:

Theorem 1:
Bottom–Up Visual Matching is NP–Complete.

It must be clear that this problem is NP-Complete because it shares with the Knapsack problem the following characteristics: an arbitrary subset of the input may be the correct solution; and, two constraints must be satisfied simultaneously. Other aspects of the problem statement such as the specific form of the functions or the fact that real numbers of fixed precision are used do not lead to the NP-Completeness.

One problem must first be solved before proceeding to the reduction. The statement of Visual Match involves images whose measurements may be non-negative real numbers (of finite precision *p* as stated in the original problem). By stating a precision *p,* we mean that the significant digits whose value is less than *p* are not represented. Therefore, a fixed number of bits are required to represent each value. This is easily solved by first proving that Knapsack with non-negative real numbers is NP-Complete. It should be stressed that the use of real numbers versus integers in no way leads to the NP-Completeness of the problem - the inherent structure of the problem is the same as that of Knapsack, regardless of the representation.

Knapsack-R_f^0
instance: Finite set A
for each a ∈ A there is a function w(a) ∈ R_f^0
(the set of non-negative real numbers of fixed precision *f*)
and a function z(a) ∈ R_f^0
positive integers C, D
question: Find a subset $A' \subseteq A$ such that
$\sum_{a \in A'} \mathbf{w(a)} \leq \mathbf{C}$ and $\sum_{a \in A'} \mathbf{z(a)} \geq \mathbf{D}$

The proof of NP-Completeness is trivial since the structure is identical. Now we are ready to prove the main theorem of this section.

Proof:
Let the set A, the functions w and z and the integers C and D specify an arbitrary instance of Knapsack-R_f^0. Set $\mathbf{M_i} = \mathbf{M_g} = \{1\}$. Define a test image and a goal image to be of size $\lceil |A|^{1/2} \rceil$ by $\lceil |A|^{1/2} \rceil$. Since |A| is less than or equal to $\lceil |A|^{1/2} \rceil^2$, $\lceil |A|^{1/2} \rceil^2 - |A|$ elements of each image will be of value 0, i.e, they join the background in the image. Therefore, $|\mathbf{I}| = \lceil |\mathbf{A}|^{1/2} \rceil^2$.

The elements of the set A will be used to determine values for the elements of the goal and test images in the following way. Set

$$\mathbf{diff(p)} = \sum_{M_1} |\mathbf{g_{x,y,j}} - \mathbf{i_{x,y,j}}| = |\mathbf{g_{x,y,1}} - \mathbf{i_{x,y,1}}| = \mathbf{w(a)}$$

and

$$\mathbf{corr(p)} = \sum_{M_1} \mathbf{g_{x,y,j}} \times \mathbf{i_{x,y,j}} = \mathbf{g_{x,y,1}} \times \mathbf{i_{x,y,1}}$$

z(a)

for each a £ A. p is the point with location (x, y) within both images, and the location is set arbitrarily, as long as each position is used and each unique position is associated with a unique element of A. It is clear from the above why the number 0 must be included as possible values for the functions w and z (it is not in Knapsack) because the difference function may have value zero, and because the background has value zero and thus the correlation function may be zero as well. A correspondence has now been set up between pixels in the test and goal image and elements of the set A. A subset of A has an associated subset of I, and the value of the difference and correlation functions correspond directly to values of the functions w and z for the corresponding element of A. Now one can solve for the values of g and i given the above pair of equations, for each spatial position. For ease of notation, g will be used for $g_{x,y,1}$ and similarly for i. Since i must be non-negative, if we wish g > i, then

$$\mathbf{i} = \frac{-\mathbf{w(a)} + \sqrt{\mathbf{w(a)}^2 + 4\mathbf{z(a)}}}{2} \quad \text{and } \mathbf{g} = \mathbf{w(a)} + \mathbf{i}$$

or if g < i,

$$\mathbf{i} = \frac{\mathbf{w(a)} + \sqrt{\mathbf{w(a)}^2 + 4\mathbf{z(a)}}}{2} \quad \text{and } \mathbf{g} = \mathbf{i} - \mathbf{w(a)}$$

It does not matter which assumption is made. The problem with this is the fact that the square root may be irrational and thus require an infinite number of bits to represent. This is solved by using the precision ρ, as given above, thus explaining the need to include this precision in the original definition. Thus if each value is truncated, then the error ϵ in each would be less than ρ. The value of ρ can be anywhere within the open interval:

$$0 < \rho < \frac{-\sum_{\mathbf{A}} \mathbf{w(a)} + \sqrt{\left(\sum_{\mathbf{A}} \mathbf{w(a)}\right)^2 + 4f|\mathbf{I}|}}{2|\mathbf{I}|}$$

This expression will be derived shortly. The error in the values can be stated explicitly as:

$$\mathbf{i}_{correct} = \mathbf{i}_{computed} + \epsilon, \quad \mathbf{g}_{correct} = \mathbf{g}_{computed} + \epsilon, \quad 0 \leq \epsilon < \rho$$

The error for g is the same as for i since no further approximations are necessary.

Next we must address the effect of this approximation on the diff and corr functions. It should be clear that if there were no truncation error, then the correspondence between the Visual Match problem and Knapsack - R_f is complete and the proof is also complete. In fact, the approximations do not affect the diff function at all, only the corr function. The diff function becomes

$$\begin{aligned}\mathbf{diff} &= |\,\mathbf{g}_{computed} - \mathbf{i}_{computed}\,| \\ &= |\,\mathbf{g}_{correct} - \epsilon - \mathbf{i}_{correct} + \epsilon\,| \\ &= |\,\mathbf{g}_{correct} - \mathbf{i}_{correct}|\end{aligned}$$

The errors cancel out since they are due to a single source. Therefore, the sums are the same, and if θ is set to C, the difference constraint will be satisfied by exactly those subsets that would satisfy it for the Knapsack problem. On the other hand the corr function becomes:

$$\begin{aligned}\mathbf{corr} &= \mathbf{g}_{computed} \times \mathbf{i}_{computed} \\ &= (\mathbf{g}_{correct} - \epsilon) \times (\mathbf{i}_{correct} - \epsilon) \\ &= \mathbf{g}_{correct} \times \mathbf{i}_{correct} + \epsilon \times (\epsilon - \mathbf{i}_{correct} - \mathbf{g}_{correct})\end{aligned}$$

There is an error term of $\epsilon \times (\epsilon - \mathbf{i}_{correct} - \mathbf{g}_{correct})$. In an exact representation, it is clear that the values of corr are exactly those of the function z for corresponding subsets. So, we need to show that as ϵ goes to zero, the error in the sum of the corr values also goes to zero. The largest possible value of ϵ is ρ. In the Knapsack-R_f^0 problem, the sums of the values of the function z(a) for the subsets of A are ordered and each is separated from the other by at least the value of f, by definition. It would then be sufficient to show that for ρ sufficiently small,

$$\sum_{\mathbf{I}'} \rho \times (\rho - \mathbf{i}_{correct} - \mathbf{g}_{correct}) < f$$

for all subsets I'.

i.e., the precision of values is higher than in Knapsack-R_f^0. The largest possible subset is the entire image, so if this inequality is true for the entire image, then it is also true for all other subsets. Since we do not know the correct values of i and g and we only have the computed values, substitute those into the above expression, giving:

$$\sum_{\mathbf{I}} \rho \times (\rho - (\mathbf{i}_{computed} + \rho) - (\mathbf{g}_{computed} + \rho))$$

$$= |\mathbf{I}| \times \rho^2 + \rho \times \sum_{\mathbf{I}} (\mathbf{i}_{computed} + \mathbf{g}_{computed}) < f$$

It remains to be shown that there exists some non-zero value of p that satisfies the above inequality. The value of p that satisfies the inequality is less than the root of:

$$|\mathbf{I}| \times \rho^2 + \rho \times \sum_{\mathbf{I}} (\mathbf{i}_{\mathbf{computed}} + \mathbf{g}_{\mathbf{computed}}) - f = 0$$

The values of $\mathbf{i}_{\mathbf{computed}}$ and $\mathbf{g}_{\mathbf{computed}}$ are not known before their computation, but the value of ρ is part of the problem definition and must be known before computation. Therefore, we can replace them with smaller known values and not affect the end result. Since w(a) = g - i, then w(a) is less than g + i, and

$$\sum_{\mathbf{A}} \mathbf{w(a)} \leq \sum_{\mathbf{I}} (\mathbf{i}_{\mathbf{computed}} + \mathbf{g}_{\mathbf{computed}})$$

Substituting this into the quadratic equation above and solving for ρ gives:

$$\rho' = \frac{-\sum_{\mathbf{A}} \mathbf{w(a)} + \sqrt{\left(\sum_{\mathbf{A}} \mathbf{w(a)}\right)^2 + 4f|\mathbf{I}|}}{2|\mathbf{I}|}$$

Since the solution must be positive there is only one possible root (and it is real). All variables have values that are known before the computation of the image elements. The value of precision for the problem can then be stated as any value of *p* such that 0 < *p* < *p'*. Given this precision, which is less than that of the original Knapsack-$R_f$roblem, and if Φ is set to D, it follows directly that the the subsets of the image that satisfy the second constraint are exactly those corresponding to the subsets of A that satisfy the second constraint of the Knapsack-$R_f$roblem. Therefore, set A' exists if and only if I' exists and Bottom-Up Visual Match is NP-Complete. But, since Visual Match is a subproblem of visual search, the Bottom-up Visual Search problem is also NP-Complete.

The Complexity of Task—Directed Visual Search

If we consider task-directed optimizations using the goal item, it is easy to show that the problem has linear time complexity. The key is to direct the computation of the difference and correlation functions using the goal rather than the test image. However, we still seek the appropriate subset of the test image. If there is a match that satisfies the constraints, then its extent can be predicted in the test image; all locations are possible. The Task-Directed Visual Match task is stated as follows:

Given a test image, a goal image, a difference function, and a correlation function, is there a subset of pixels of the test image such that the difference between that subset and the corresponding subset of pixels in the goal image is less than a given threshold and such that the correlation between the two is as large as possible? In other words, is it true that

$$\sum_{\mathbf{p} \in \mathbf{G}} \mathbf{diff(p)} \leq \theta \text{ and } \sum_{\mathbf{p} \in \mathbf{G}} \mathbf{corr(p)} \geq \phi \ ?$$

where

$$\sum_{\mathbf{p} \in \mathbf{G}} \mathbf{diff(p)} = \sum_{\mathbf{p} \in \mathbf{G}} \left[\sum_{\mathbf{J} \in \mathbf{M_g}} | \mathbf{g_{x,y,J}} - \mathbf{i_{x,y,J}} | \right]$$

and

$$\sum_{\mathbf{p} \in \mathbf{G}} \mathbf{corr(p)} = \sum_{\mathbf{p} \in \mathbf{G}} \left[\sum_{\mathbf{J} \in \mathbf{M_g}} \mathbf{g_{x,y,J}} \times \mathbf{i_{x,y,J}} \right]$$

Note that the definition of diff and corr have changed slightly in that the pixel locations and measurement set used are that of the goal rather than the test image. The computation of the diff and confunctions is driven by the goal image and the measurements present in the goal. A simple algorithm is apparent. First, center the goal item over each pixel of the test image; compute the diff and corr measures between the test and goal image at that position; among all the positions possible, choose the solution that satisfies the constraints. The resulting time complexity function for visual match would be $O\{|G| \times |M_g| \times |I'|$ In other words, the worst case number of computations of the diff and corr functions is determined by the product of the size of the goal image in pixels and the number of measurements in the goal image. If the complexity of visual search (within which visual match is embedded) is considered, this would add only a multiplicative term |T| to the above function, where this represents the total number of possible rotations, translations and scalings. If display items can be localized, (via an attentional spotlight), the complexity is linear in the number of items in the display, but all rotations and scalings must still be considered. Since at least one linear algorithm exists, this leads to the second theorem:

Theorem 2:
Task—Directed Visual Search has Linear Time Complexity

Note that no sacrifice of generality is necessary to deal with the task-directed problem. However, this is true only for the decision problem as stated involving perfect matches. If the problem is stated as one required to 'find' the matching image subset, the task-directed version is constrained to find subsets

that are the same size and shape as the goal, whereas the bottom-up version can find subsets of arbitrary sizes and shapes. The bottom-up case is still NP-Complete, and the task-directed case is still linear, but they are not as directly comparable. This is true even if partial matches are permitted.

This provides a strong hypothesis: since visual search experimentation in psychology presents a view of search performance as having linear time complexity, and not exponential, the inherent computational nature of the problem strongly suggests that task-directed influences play an important role in human perception.

Conclusions

Complexity theory has not been previously applied to try and uncover the inherent difficulty of behavioral experimental paradigms (for a comprehensive overview of methods that have been applied to uncover the limits of perception see [van Doom et al. 1984]), even though several researchers are attempting to incorporate the results of such experiments into computational theories. Is it possible that the consideration of the computational difficulty of the experimental task alone could lend insight into the interpretation of response times of human subjects? This paper demonstrated that indeed additional insight is possible. The results argue very strongly against purely bottom-up approaches to the general vision problem and to computational modelling of human perception. It is claimed without proof that the same results hold for perceptual search tasks in stimulus modalities other than vision.

Visual search is a common if not ubiquitous sub-task of machine vision algorithms. For example, purely bottom-up versions of region growing, shape matching, structure from motion, the general alignment problem, and connectionist recognition procedures, etc., all are specialized versions of visual search in that the algorithms must determine which subset of pixels is the correct match to a given prototype or description. The problems they attempt to solve are therefore NP-Complete. It must be stressed that the NP-Completeness does not depend on the specific functions used for the two constraints nor on the representation of images. NP-Completeness results from the facts that a subset of pixels must be chosen whose extent cannot be predicted a priori, that simultaneously satisfies (as opposed to optimizes) two constraints.

Acknowledgements

The following assisted with the development and verification of the proof: Steve Cook, Hector Levesque, Charles Rackoff, Raymond Reiter, and especially Bart Selman and Gilbert Verghese. The author is a Fellow of the Canadian Institute for Advanced Research. This research was conducted with the financial support of the Natural Sciences and Engineering Research Council of Canada and the Information Technology Research Center, a Province of Ontario Center of Excellence.

References

Garey, M. & Johnson, D. (1979). *Computers and intractability: A guide to the theory of NP-completeness,* W.H. Freeman and Co., New York.

Grimson, W.E.L., (1986). The combinatorics of local constraints in model-based recognition and localization from sparse data, *Journal of the Association for Computing Machinery 33-4,* 658 - 686.

Helmholtz, H. von, (1925). Hanbuch der physiologischen detik, English translation Southall, J.

Mackworth, A. & Freuder, E. (1985). The complexity of some polynomial network consistency algorithms for constraint satisfaction problems, *Artificial Intelligence 25,* 65 - 74.

Nakayama, K., Silverman, G., "Serial and Parallel processing of Visual feature Conjunctions", *Nature 320-6059,* pp. 264 - 265, 1986.

Neisser, U., (1967). *Cognitive psychology,* Appleton-Century-Crofts, New York.

Poggio, T. (1982). Visual algorithms, AI Memo 683, MIT.

Rabbitt, P. (1978). Sorting, categorization and visual search, in *The Handbook of Perception: Perceptual processing, Vol. IX,* edited by E. Carterette and M. Freidman, Academic Press, New York.

Treisman, A. (1988). Features and objects, *The Quarterly Journal of Experimental Psychology 40A-2,* 201 - 237.

Tsotsos, J., (1988a). A 'complexity level' analysis of immediate vision, *International Journal of Computer Vision 1-4,* 303 - 320.

Tsotsos, J., (1988b). Analyzing vision at the complexity level, submitted for publication.

van Doom, A., van de Grind, W. & Koenderink, J. (1984) (editors). *Limits in perception,* VNU Science Press, Utrecht, The Netherlands.