Semantic Intensity: a measure of visual relevance

Gerardo Beni
University of California Riverside, Department of Electrical Engineering
900 University Ave
Riverside CA 92521
beni@ee.ucr.edu

Nicoletta Adamo-Villani
Purdue University, Department of Computer Graphics Technology
401 N. Grant St.
West Lafayette IN, 47907-2021
00 765 496 1297
Fax: 00 765 494 9267
nadamovi@purdue.edu

Abstract

This work deals with the quantification of the effects of discarding large amounts of irrelevant information in sets of images. In particular, the trade-off between the cost of losing information and the benefit of increasing perceptual speed is examined.

We look for a measure of such trade-off by a new measure, 'semantic intensity'. broadly defined as the ratio of amount of 'meaning' conveyed to the amount of 'effort' required to perceive it. The establishment of more precise definitions of these two quantities ('meaning' and 'effort of perception of meaning') is the pervasive theme of the paper.

A preliminary operational definition of semantic intensity is introduced and applied to a comparison between avatar and semantroid showing the advantage of the latter.

A more precise definition of semantic intensity is then given by relating 'meaning' to information and 'effort' to bit geometric density, compactness, contrast and homogeneity.

The relation of 'meaning' to information is mediated by an extension of Relevance Theory from the verbal to the visual domain, resulting in a new constraint which qualifies information as 'meaning'. The constraint is the requirement of direct representation, or pointing to a representation, without intermediate mental operations.

Application of this new definition of semantic intensity also shows the advantage of semantroids over avatars, in many practical cases.

This work may be regarded as a first step toward a quantitative extension of Relevance Theory to the visual domain.

Keywords:

Information, Meaning, Perception, Avatar, Relevance Theory, Cognitive Science, Image Processing, Semantroid

1. Introduction
In order to make efficient use of data they must be somehow reduced to manageable proportion. This need is general for any type of data and applies in particular to images. Examples of reductions of image data are image compression, feature extraction, and pattern recognition. The purpose of these reduction operations is to summarize the complexity of the image data into something that can be used efficiently.

For the examples mentioned, the methods are analytical (mathematical, computational). But the same need for summarizing image data exists also in visual perception. And the basic operations done analytically or by the human visual perceptual system are essentially the same two: (1) discovering regularities and (2) discarding irrelevant information. Both operations allow a simplified description of the image. In this paper we are not concerned with operation (1).

Discarding irrelevant information in an image requires a measure of relevance; and the benefits of the discarding operation must also be quantified. Lossy compression is an example in which a choice of what is relevant must be made; in this case, the benefit is the reduced amount of data to store or transmit.

The discarding of small amounts of irrelevant information (as in the case of lossy compression) has been quantified. But the drastic discarding of large amounts of irrelevant information has not been quantified yet. This is the focus of the paper.

An example of drastic discarding of irrelevant information is the abstraction of a photograph into a diagram. To a certain extent this is quantifiable by algorithms for edge detection and extraction. But more severe abstractions of images as in going from a photograph to a stylized character (see Figure1 [http://www.goofyfaces.com/gallery.php]) have not been quantified yet. Many questions, in fact, remain unanswered. In particular: (1) how do we compare the meaning conveyed by a movie with a human actor with the meaning conveyed by the same movie when the actor is replaced by a 2D cartoon character representing the same actor? (2) how do we make the same comparison when the 2D cartoon character is replaced by a 3D animated
avatar? And (3) what if we remove part of the scene or some of the objects? When is the meaning lost and when is clarity gained?

Such questions are relevant in deciding for example when to use a photograph or when to use a drawing to illustrate a scientific experiment. The former gives more details (some of them irrelevant); the latter gives fewer details but what it shows is less confusing.

Since answers to these questions are likely to become more important as information is increasingly conveyed visually, with this paper we begin an investigation of quantifying the effect of severely reducing the amount of visual data on the amount of 'meaning' conveyed and on the effort in perceiving it.

Figure 1. Stylized character derived from photograph.

2. Intuitive notion of semantic intensity

Recently the concept of *semantroid* (an avatar limited to head and hands) has been introduced (Adamo-Villani and Beni, 2004a, 2004b). The word *semantroid* is coined as the combination of 'semantic' and 'android' in order to convey the idea of an android (e.g., avatar, robot, etc.) highly efficient in conveying meaning. A measure of such efficiency is provided by the *semantic intensity* which is defined, broadly, as the ratio of the 'amount of meaning' conveyed to the 'amount of effort' required to perceive it. Every image, in so far as it conveys meaning, and in so far as this meaning requires some perceptual effort to be grasped, has a certain measure of semantic intensity.
The quantification of this intuitive concept has been sketched in (Adamo-Villani and Beni, 2004a) and in this paper we define the concept more precisely.

The notion of *semantic intensity* can be applied to images regarded as sets of component objects, and also, as we discuss later in this paper, to images in general. The former approach leads to a less rigorous but more intuitive illustration of the concept. Thus, we start with this approach. A segmentation of the image as a set of objects is assumed. Since an image can be segmented in different ways, the concept of semantic intensity, developed along these lines, is not a property of an image *per se* but of any one of its segmentations.

Given a segmentation of a 2D image in N objects $J_i$ ($i = 1, 2, .. N$), the intuitive idea of *semantic intensity* is based on the assumptions: (1) that each component object $J_i$ carries some meaning $M_i$ and that to perceive such meaning requires an effort $E_i$; (2) that measures for $M_i$ and $E_i$ can be found. With these two assumptions, the *semantic intensity* can be defined qualitatively as the ratio of the total meaning conveyed by the objects to the total effort made by the perceiver.

$$\Xi = \frac{\sum_i M_i}{\sum_i E_i}$$  \hspace{1cm} [1]

Assumption (2) requires establishing measures of 'meaning' and 'effort of perception of meaning'. The establishment of these two measures is the pervasive theme of the entire paper. In (Adamo-Villani and Beni, 2004a), in order to argue for semantroids as replacements for full-body avatars, simple, intuitive measures of $M_i$ and $E_i$ for animated characters were adopted as follows.

### 3. Operational definition of semantic intensity

For $M_i$ we do not consider the information contained in the $J_i$ object itself but only the information contained in its possible variations during the animation. The number of possible
variations scales with the number of degrees of freedom \( C_j \) of \( J_i \) for the motion of the avatar. Hence we take simply

\[
M_i = \gamma \ C_i \tag{2}
\]

where \( \gamma \) is a constant. (Alternatively, we could have chosen the information measure \( M_i = \gamma \ log_2 C_i \) but the qualitative results are not affected.)

More difficult is to devise a measure of \( E_i \) since perception 'effort' is not a well established concept. In analogy with the problem of measuring the difficulty of positioning a mouse on an object, we consider the effort of positioning the eye on the object as having a similar dependence on the geometry of the object and its relation to the image. Such dependence, in the case of the reaction time in positioning a computer mouse on an object is given, e.g., by Fitts' law (Fitts, 1954). We make then the assumption that the effort of perceiving the object \( J_i \) follows the law

\[
E_i = k_1 + k_2 \ log_2 (D_i/S_i + 1) \tag{3}
\]

where \( k_1 \) and \( k_2 \) are two empirical parameters.

We estimate \( k_1 \) and the ratio \( k_1/k_2 \) as follows. From eq.(3), the parameter \( k_1 \) measures the effort at distance 0 from the screen center (assumed to be the rest position of the eye). There must be an effort even at distance zero; we assume that this effort is the effort of visually scanning the object which we may take to be proportional to its area, which in turn, we can take to scale as the square of a length \( S_i \) characterizing its size, \( S_i^2 \). Note that this is not the case for Fitts' law (as applied to the time it takes the mouse to reach a target). In such a case there is no time cost in scanning the target.

Again from eq.(3) it can be seen that the parameter \( k_2 \) measures the effort at distance \( D_i = S_i \) from the center. The area to be scanned at this distance is (approximately) proportional to \((2S_i)^2\). Thus we may estimate the ratio

\[
k_2/k_1 = 3 \tag{4}
\]
'Rules of thumb' for these constants for the 'mouse on object' case also estimate the value $k_2/k_1 = 3$ (Raskin, 2000).

To estimate the *semantic intensity* $\Xi_s$ of a semantroid (represented in figure 2) vs. the *semantic intensity* $\Xi_a$ of the avatar from which the semantroid is derived, we refer to a standard character which approximates a typical avatar (Figure 3). The significant measures of this model are listed in Table 1. The number of degrees of freedom for the head and hand (26) are estimated from recent work on keyboard control of facial expressions and hand gestures. (see, e.g., Adamo-Villani and Beni, 2004c, 2004d). The other parts are assumed to have one dof, as an angle of rotation.

![Figure 2](image.png)

**Figure 2.** An example of “semantroid”. The numbers refer to the degrees of freedom of each component.
Figure 3. A typical avatar segmented into components. Distances of components from center of screen are indicated by red lines. White numbers indicate the dof=M of each component; yellow numbers indicate size S; and black numbers indicate distance D.

<table>
<thead>
<tr>
<th>COMPONENT</th>
<th>SIZE</th>
<th>DISTANCE</th>
<th>EFFORT</th>
<th>MEANING</th>
</tr>
</thead>
<tbody>
<tr>
<td>Head</td>
<td>4</td>
<td>10</td>
<td>102.8</td>
<td>26</td>
</tr>
<tr>
<td>Hand</td>
<td>3</td>
<td>17</td>
<td>82.9</td>
<td>26</td>
</tr>
<tr>
<td>Arm</td>
<td>5</td>
<td>9</td>
<td>136.4</td>
<td>1</td>
</tr>
<tr>
<td>Forearm</td>
<td>5</td>
<td>12</td>
<td>157.4</td>
<td>1</td>
</tr>
<tr>
<td>Trunk</td>
<td>10</td>
<td>0</td>
<td>100</td>
<td>1</td>
</tr>
<tr>
<td>Thigh</td>
<td>7</td>
<td>6</td>
<td>180.3</td>
<td>1</td>
</tr>
<tr>
<td>Leg</td>
<td>6</td>
<td>13</td>
<td>215.6</td>
<td>1</td>
</tr>
<tr>
<td>Foot</td>
<td>3</td>
<td>17</td>
<td>82.9</td>
<td>1</td>
</tr>
<tr>
<td>Avatar total</td>
<td>1913.8</td>
<td>89</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Semantroid</td>
<td>268.6</td>
<td>78</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Table 1. Parameters used for $E_i$ and $M_i$: $k_i = S_i^2$; $k_2/k_1 = 3$; $\gamma_i = 1$

From Table 1 and eq.(1), the ratio of semantic intensities turns out to be

$$\Xi/\Xi_s = 6.2$$

[5]

The analysis above is for 2D images. Since an avatar is typically a 3D model, in such a case its 3D measured lengths $D_i$ and $S_i$ should be averaged over the projections on the plane of the
screen. These averaging results in a constant scaling factor, hence it does not affect the ratio \( D_i / S_i \) and thus has no effect on \( E_i \).

The analysis above concludes, in so far as conveying meaning (apart from emotions) to the viewer, that semantroids are preferable to whole avatars. To see if this result makes sense it is interesting to look at intuitive notions. Figure 4 illustrates the intuitive notion that a semantroid contains most of the meaning in the message conveyed by a character.

![Figure 4](image)

**Figure 4.** (a) Pantomime dancer. Only head and hands are emphasized. (b) Semantic Intensity is concentrated in hands and face (public domain images).

4. **Toward a better measure of semantic intensity: bit density**

As developed so far, the measure of semantic intensity relies on perceived ocular motion effort to locate the various objects in the image and on the number of dof as measure of 'meaning'. More satisfactory measures of 'meaning' and 'effort' can be obtained by the following considerations. We first focus on 'meaning'. And, from now on, we consider images in general, regardless of segmentation into objects.

For our purposes an image is a set of \( N \) pixels each capable of taking \( C \) different colors; all the images characterized by the pair \( (N,C) \) form an image set \( \{N,C\} \). For the time being we do not consider the layout of the \( N \) pixels (e.g., we do not distinguish between a line of 64 pixels and a square array of 8x8 pixels). We consider this aspect later when we focus on 'effort'.

444
If we consider the set \{N,C\}, each of its images can be regarded as a possible outcome of an observation event, and we make the assumption that each outcome is equally likely. We also assume that each image has a *meaning* (a property as yet unspecified). Every image has one and only one meaning. We are excluding ambiguous images, i.e., images with more than one meaning. But several images may have the same meaning. So, meaning is a property of a subset of the set \{N,C\}. With this definition, images which are subjectively considered 'meaningless', also have meaning.

Note that 'meaning' here is not exactly information. In fact, we are not necessarily interested in the entropy (uncertainty) of the set \{N,C\}. Also, since all images are assumed to be equally likely to be presented to the observer, the entropy (uncertainty) of the probability distribution of the outcomes is maximized a priori.

Some of the pixels in an image set \{N,C\} may be such that they cannot change color. We call these pixels 'inactive' and the other pixels, 'active'. We assume that the inactive pixels are distinguishable by the observer as such. The inactive pixels can be referred to as the *background* of the image set. The background is a single image so it has one meaning, while the set of \(P = C^{N_A}\) (\(N_A\) is the number of active pixels) images formed by active pixels can have a maximum of \(P = C^{N_A}\) meanings.

Thus, the ratio of active pixels \(N_A\) to the total number of pixels \(N\), i.e., \(\rho = N_A/N\), is expected to be a factor in the measure of the semantic intensity of an image set. It measures the geometric density of 'meaning'.(See figure 5). Such notion is closely related to the measure of bit density, i.e., the number of bits per unit area. This is a common notion in describing bar codes, magnetic strip cards, etc. Below we return to this general measure of geometric bit density in order to relate it quantitatively to the semantic intensity.

Let us now refine the measure of density of meaning apart from the geometric density. Consider the set \{N,C\} consisting of \(C^{N_A}\) different images. Out of these images, a number \(P\)
\[ C^N \] will actually be perceivable (available for observation). And out of these P images, a subset of M images can be regarded (subjectively) as being significant to the observer.

A significant image is defined as an image whose meaning the observer will consider. An insignificant image is an image whose image will not be considered by the observer. Let \( m \leq M \) be the number of distinct meanings of the M significant images. Then \( \mu = \frac{m}{M/P} \) is another factor in the measure of the 'meaningfulness' of an image set. (In fact \( \mu \) is closely related to the information of the image set, \( \log_2 \mu \))

As an example, consider a deck of 52 cards. Only one card is visible at any time. Assume that the only significant images are the J Q K cards and that the meaning is irrespective of the suit. Then \( P = 52, M = 12, m = 3 \) and \( \mu = \frac{9}{13} \).

For a single, significant image, \( P = M = m = \mu = 1 \).

Thus, provisionally (we reconsider \( \rho \) later), the product \( g = \rho \mu \) can be taken as a measure of the overall 'density' of meaning of the set \( \{N,C\} \), before considering the details of the image set.

Apart from the number \( m \) of possible distinct meanings, we can regard \( \mu \rho \) as the overall probability that an observer looking at (i.e. processing the visual field of) the pixels of the set \( \{N,C\} \) will perceive a significant image. In fact, the probability \( \mu \) of perceiving a significant image is conditioned by the probability \( \rho \) of perceiving active pixels.

Since, in practice, the numbers tend to be large, it is convenient to take logarithmic measures. We use base 2 (as in measures of information). So 'meaningfulness' can be taken to be measured by

\[ g = \log [ \mu \rho ] \] [6]
Figure 5. Illustration of geometric density of meaning. The active pixels are represented as the colored ones.

Figure 6. Illustration of semantic density of meaning. The whole set has P members. Yellow set has P-M members and represents insignificant images. Other colors represent, in total, M significant images. The distinct meanings are represented by the distinct colors. Here m=4.

Example 1. An image set may consists of an array of 16x16 pixels of which only the top left 8 are active. Each pixel is capable of 2 different colors (black or white). All possible images are perceivable. Significant images are taken to be only those with a black leftmost pixel, and the meaning of two images is the same if they differ only by the rightmost pixel. Here \( N = 2^8 \), \( N_A = 2^3 \), \( C = 2 \), \( P = 2^8 \), \( M = 2^7 \), \( m = 2^6 \), and \( g = 6 + (7-8)) + (3-8) = 0 \).

Figure 7. The top left 8 pixels of example 1. Only the two lower images are significant, and they have the same meaning.

In this example the number of meanings compensates for the number of insignificant images and for the inactive pixel area.
Example 2. If we remove the constraint on the rightmost pixel, then each significant image has a distinct meaning, \( m = 2^7 \), and \( g = 1 \).

Example 3. On the other hand, if we restrict the significant images to only those with black leftmost two pixels, then \( M = 2^6 \) and assuming all these \( M \) image to have distinct meanings, we have \( g = 6 + (6-8) + (3-8) = -1 \).

These examples of \( g \) values are consistent with the intuitive notion of more or less 'meaningful' image sets and with the \( g=0 \) value of a single meaningful image. For the latter, we assume that all the pixels are active \( (N_A = N) \) but only one of the \( C_N \) possible images is perceivable. Since \( m = M = P = 1 \), then \( g = 0 \).

So far, we have not considered the content that determines the meaning of each image; we have only considered the geometric \( (N_A/N) \) and the semantic \( (M/P) \) densities of meaning, as well as the number of different meanings \( (m) \). But the content itself carries more or less meaning depending on other factors. To see what these factors are let us consider another example.

Example 4. This example has the same \( g \) as Example 1 but a different image set. Let the area be an array of 64 x 128 pixels of which only the top left square of 16x16 pixels is active. In this square the pixels can be black or white with the constraint that only one of the pixels is black. All possible images are perceivable. The images where the black pixel falls in the lower half of the array are meaningless. The meaning of an image does not change whether the black pixel is in a location or in its left-adjacent. With these constraints we have: \( N = 2^{13} \), \( N_A = 2^8 \), \( P = 2^8 \), \( M = 2^7 \), \( m = 2^6 \), and \( g = 6 + (7-8)) + (8-13) = 0 \).

Thus \( g \) is the same as in Example 1, but are the two cases equivalent?

If we look at the two examples from a quantity of information point of view, they appear identical. It takes the same number of bits to describe the two situations. What is then the difference?

To clarify this difference, let us reverse our point of view and ask the following. Suppose we want a set of \( M \) significant images each carrying a distinct meaning \( (m=M) \), what is the
minimum P so that we can maximize g? (We assume for now that all pixels are active, \( N_A = N \); this does not change the basic argument). Basically we are asking for the minimum number of bits required for the representation of the m meanings. This is clearly \( \log m \); and it is achieved by the standard representation of bits in a binary sequence. This can be seen by the following.

Example 5. If we want \( m = M = 2^6 \) we can obtain the most efficient representation by using 6 pixels each of which can be black or white. Here \( N = N_A = 6, P = M = m = 2^6 \) and \( g = 6 \). This case is thus more meaningful than Example 4 which has a \( g = 5 \) (excluding the geometric density loss of -5 which we are not considering here).

But intuitively we would expect the opposite. In fact, suppose one wanted to convey the meaning of values from 1 to 64. Clearly, Example 4 where the position of one black pixel identifies a meaning is an easier representation of value to grasp than the binary sequence case. In the former case it is a matter of counting; in the latter, the meaning is inferred after a calculation. We argue that this fact should be taken into account in an appropriate definition of meaningfulness.

This argument leads us to the need of introducing, along with 'quantity of meaning' \( g \), the concept of 'quantity of effort' in perceiving the meaning. If we do not take into account the perceptual effort, the two representations of the values of numbers in the examples above are equivalent. And the case of Example 4, being equivalent in number of meanings to Example 5, but more 'costly' in pixels needed to represent the meaning, would turn out to be less 'meaningful' (smaller \( g \)). But the advantage of Example 5 is so only because the two Examples (4 and 5) are assumed to be equivalent in the effort required to grasp the meaning. To appreciate the difference between the difference in the two ways of conveying the meaning and the relation of this difference to the 'effort' of perception we turn, below, to Relevance Theory.

Before doing so, however, we note that at the root of the difficulty is the notion (from information theory) that different representations are equivalent if they require the same number of bits for their description. And they are indeed, in information theory, because in information
theory one does not need to worry about the effort of translating, say, 1 0 1 0 0 into 20. But a perceiver of an image will take a certain mental effort to decode the bits represented as binary sequences into their meaning. This effort cannot be ignored in a measure of conveyed 'meaning'. This is the basic difference between bits of information and bits of 'meaning'. We elaborate further after comparing with Relevance Theory.

5. Relevance Theory

In order to obtain a measure of 'quantity of meaning', so far we have examined notions in the context of visual content; such notions have parallels in the context of verbal content. This parallel is evident from Relevance Theory (Sperber and Wilson, 1983), an area of cognitive science.

Relevance Theory is based on a definition of relevance and two general principles: a Cognitive and a Communicative Principle of Relevance. The theory is not quantitative. Relevance is defined in cost-benefits terms, as a property of inputs (e.g. sentences) to cognitive processes. The benefits are positive cognitive effects (e.g., understanding something), and the cost is the processing effort needed to achieve these effects. Thus, relevance is a concept closely related to semantic intensity insofar as 'amount of meaning' is a positive cognitive effect (a benefit) and the effort of perceiving the meaning is a processing effort (cost). Relevance Theory holds: (1) that, other things being equal, the greater the positive cognitive effects achieved, the greater the benefit of the input to the individual who processes it; and (2) that the processing of the input and the understanding of it, involves some effort of perception, memory and inference. The smaller the processing effort required, the greater the relevance of the input.

Paralleling these notions, Semantic Intensity Theory holds: (1) that, other things being equal, the greater the 'quantity of meaning' conveyed, the greater the benefit of the image to the individual who perceives it; and (2) that the processing of the image and the understanding of it, involves some effort of perception, memory and inference. The smaller the 'quantity of effort' required, the greater the semantic intensity of the input.
Given this close parallel between relevance and semantic intensity it is useful to recall the two principles of Relevance Theory. According to the Cognitive Principle of Relevance (Sperber and Wilson, 1983) the human cognitive system tends to allocate attention and processing resources in such a way as to maximise the relevance of the inputs it processes. This principle can be carried over to semantic intensity if lexical input is substituted with visual input. The observer of an image set will allocate attention and processing resources in such a way as to maximize the meaningfulness of the inputs it processes. Hence a highly meaningful image set (high g) will be attended to preferentially to an image set with low g. We make use of this fact in the application to the notion of semantroid (Adamo-Villani and Beni, 2004b)

The second principle of Relevance Theory, the Communicative principle, (Sperber & Wilson, 1983) states that utterances create general expectations of relevance. We could extend the principle from utterances to images and argue that images create general expectations of semantic intensity. But, in this paper, we are not concerned with what expectations images generate and so we restrict our attention to the first principle of Relevance Theory to which we now return.

It is instructive, in order to appreciate the importance of differences in visual representations of meaning (as in Examples 4 and 5), to consider an example from linguistic context (to which Relevance Theory applies). The following is taken almost verbatim from (Sperber and Wilson, 1983). Note that in this verbal example 'effort' corresponds to the idea of 'effort' in semantic intensity theory, while 'positive cognitive effect' corresponds to the idea of 'quantity of meaning' in semantic intensity theory.

*Here is a brief and artificial illustration of how the relevance of alternative inputs might be compared in terms of effort and effect. Mary, who dislikes most meat and is allergic to chicken, rings her dinner party host to find out what is on the menu. He could truly tell her any of three things:*

(a) We are serving meat.
(b) We are serving chicken.

(c) Either we are serving chicken or \((7^2 - 3)\) is not 46.

According to the characterisation of relevance, all three utterances would be relevant to Mary, but (b) would be more relevant than either (a) or (c). It would be more relevant than (a) for reasons of cognitive effect: (b) entails (a), and therefore yields all the conclusions derivable from (a), and more besides. It would be more relevant than (c) for reasons of processing effort: although (b) and (c) are logically equivalent, and therefore yield exactly the same cognitive effects, these effects are easier to derive from (b) than from (c), which requires an additional effort of parsing and inference (in order to work out that the second disjunct is false and the first is therefore true). Thus, (b) would be the most relevant utterance to Mary, for reasons of both effort and effect. More generally, when similar amounts of effort are required, the effect factor is decisive in determining degrees of relevance, and when similar amounts of effect are achievable, the effort factor is decisive.

Although Relevance theory is not quantitative, the qualitative correspondence is strikingly close when verbal context is replaced with visual context. What is most important for us is to focus on the comparison between case (c) and case (b) in the example. Both convey the same 'quantity of meaning' but the effort is different. So, what is the analogy with the visual case? The analogy is with the different levels of efforts for the same number of bits of information. For example, there is a higher level of efforts for bits represented by binary sequences, as in Example 5, than for bits represented by 'pointers' (An N-pointer can be defined as a (N,2) image with only one pixel differing in colour from the others) as in Example 4.
Since mental effort is not quantifiable (and this is the reason why Relevance Theory is not quantitative) we argue that:

(Constraint 1) any measure of perceptual effort must rule out any 'indirect inference' of the meaning conveyed by the image. Granted that 'indirect inference' is not precisely defined, the idea is to refer to mental operations using abstractions (i.e., mathematical, logical, analytical, and memory work). This constraint eliminates the use of binary sequences as representations of numeric values. Binary sequences require a calculation (except in trivial cases) to infer the value represented. Instead, according to Constraint 1, the meaning conveyed by an image must be directly perceivable (by the eye-brain system). While the 'pointer' satisfies this constraint, the binary sequence does not.

But what is then the meaning of a binary sequence set? Is it to be considered insignificant? The answer is in what the meaning of the binary sequence is taken to be. If it is numeric values, we cannot quantify its 'meaning' since the mental effort of inferring the numeric values cannot be quantified. Thus the binary sequence set is insignificant. On the other hand, consider Example 6.

Example 6. Suppose that each pixel of a binary sequence represents the status (on-black, off-white) of a different light bulb; then the representation of the meaning is direct and it is quantifiable: \( N=N_A \), \( P=M=m=2^N \). Thus \( g = N \), where \( N \) is the number of pixels. Consider a different case.

**Figure 8.** Two examples of \((N,2)\) pointers for \(N=20\) and \(N=64\) respectively.
Example 7. One single pixel capable of taking any of 8 colours. If each colour has a distinct meaning, \( g = \log 8 = 3 \), this case is quantifiable but only if the colours are meaningful in themselves and not as symbols. If the meaning of a red pixel is "Red" and of a blue pixel "Blue", there is no difficulty. But if the meaning of a red pixel is "anger" and the meaning of a blue pixel is "calm", then this is not quantifiable since it violates Constraint 1.

So, the concept of 'quantity of meaning' in semantic intensity is applicable to every image set but it is restricted by Constraint 1. Semantic intensity theory is intended for practical use and it is not intended to solve the very complex epistemological problem of the meaning of meaning (Ogden and Richards, 1923).

Does all this amount to a very limited application of the theory? Not really. For one thing, in most applications in which the theory is useful, image sets consists of direct representations of degrees of freedom, as in Example 6.

Consider now a typical case

![Image of L shaped object](image_url)

**Figure 9.** L shaped object can freely translate and rotate in 90° steps.

Example 8. (See Figure 9) On an area of 16 x 16 white pixels an L-shaped 4-pixel black object can translate anywhere and can orient in 4 directions : N-S-E-W. Here the image set
represents directly the 'meaning'. The 'quantity of meaning' g is calculated from eq.(6) as: \( N = N_A = 2^4 \times 2^4 \) (neglecting boundary effects); \( m = M = P = 4 \times 2^4 \times 2^4 = 2^{10} \), \( g = 10 \).

This is straightforward case and typical of common situations in which semantic intensity is calculated. In this typical case g is calculated simply by adding the bit-depth of each degree of freedom (4 bit x-translation; 4-bit y-translation; 2 bit orientation).

Returning to the issue of the 'effort' of perceiving the meaning we can now make clearer the significance of Constraint 1. An image has 'meaning' either directly, as representation of something (e.g., a chicken), or indirectly, as symbol of something (e.g., a checkmark or a flag). In the latter case, a symbol may point to another image (checkmark next to item in a catalogue) or to a concept (a flag pointing to the idea of a nation). The directly represented 'meaning' (chicken) satisfies Constraint 1 while the symbol (flag) pointing to a concept does not. Less clear is the situation with a symbol (checkmark) pointing to an image. In this case we argue that the meaning is carried directly whenever what is pointed to is part of the image set (and so the meaning does not require mental effort to be inferred from the symbol). Thus, we can quantify the amount of meaning of an image of the type shown in figure 10.

Example 9. In the case of figure 10, g can be calculated as follows. Assume that the image is composed of \( N = 70 \times 210 \) pixels. The only active pixels are those of the slider pointer \( N_A = 10 \times 190 \). The total number of images is (assuming the black pointer to be a 10x10 square) \( P = 190 \times 10 + 1 = 181 \). The meaningful images are those corresponding to positions of the black pointer which unambiguously identify one of the 4 objects. Thus, approximately, \( M = 150 \).
Finally $m = 5$ since only 4 unambiguous 'meanings' can be given with the image set and we add 1 meaning for the 'meaningless' cases. With these values we obtain $g = \log 5 + \log 150/181 + \log 1900/14700 = -0.9$. Thus, Example 9 shows that images of symbols pointing to images have an easily quantifiable amount of meaning. So, Constraint 1 eliminates from considerations only images which are symbols of non-visually represented concepts. Only in this case a non quantifiable mental effort is involved.

Note that the situation with images is more favourable than with words, which are always symbols and never direct representations. This is why there is always a mental effort involved in perceiving a meaning from a word, and this is why Relevance Theory cannot be made quantitative as long as it remains in the verbal domain. Semantic Intensity Theory may be regarded as a quantification of Relevance Theory made possible by the visual context in which it operates.

Essentially, the meaning must be accessible by any form of direct perception or by pointing to it, but, in any case, without any inference, calculation or other mental operation requiring a non-negligible time. Thus, the way an image conveys the meaning must be by way of representing or pointing to visual representations. In this way, the description of the image is also a description of its meaning. To repeat, the possible meanings that an image may convey must be represented directly, i.e., without requiring an intermediate mental operation. We may refer to these images as 'representations' whereas images that convey meanings via an intermediate mental operation can be referred to as 'symbols'. Semantic intensity deals only with representations. Clearly there are ambiguous cases in which it is not easy to tell whether an image is a 'representation' or a 'symbol' but it should also be clear that a distinction can be made and that this distinction can be operationally useful.

6. Compactness

So far we have arrived at the following conclusion. Since we are seeking a quantitative measure of perceptual effort, we must consider only representations of meaning which do not
involve intermediate mental operations. This is the gist of Constraint 1. Is then \( g \) a satisfactory measure of semantic intensity? Not quite. The reason is that we have so far succeeded in removing the role of mental effort in the perception of meaning but we have not included some important aspects of perceptual effort. To include these aspects we must revisit the geometric bit density \( \rho \).

In fact, so far, we have included perceptual effort only insofar as the geometric density factor \( N_A/N \) besides measuring information density also measures an effort of perception. The image of Fig. 5 would require more perceptual effort, and hence would have less semantic intensity, if the total area were doubled while the active (colored) area were to remain constant. The factor \( N_A/N \) broadly takes into account this effect.

![Figure 11](image)

**Figure 11.** The image has the same area as in Fig.5 but different shape.

However, it may be argued that a more accurate measure of effort is \( \sqrt{\rho} \) since it measures ratios of lengths; distances being the significant factor in visual perceptual effort (focusing of saccadic ocular movements).

A second point to consider is the effect of the shape of the area. The same area may require more or less perceptual effort depending on its shape. If the image of figure 5 were to be changed into that of figure 11, \( \rho \) would remain unchanged (\( N \) and \( N_A \) are the same) but the perceptual effort would change, figure 11 requiring more effort. Again it is the distance, or rather the ratio of the size to the distance, (hence \( \sqrt{\rho} \)) that comes into play.

It makes sense, therefore, to include a shape compactness factor for the total area and for the active area in the measure of perceptual effort. We may also argue that it is simpler and practically sufficient to consider only the compactness of the bounding box of the active areas. A measure of compactness is \( q = (\text{perimeter})^2/\text{area} \), which is maximal for a disk shape. Factoring
I/q with the geometric density term \( \rho (=N_A/N) \), we are finally led to replacing \( \rho \) in eq. (6) with

\[
\frac{\sqrt{N_A}}{c}
\]

and we get

\[
\xi = \log \left[ \frac{m}{M/P} \left( \frac{\sqrt{N_A}}{c} \right) \right]
\]

where \( c \) is the perimeter of the maximal bounding box of the active areas.

The factor \( \frac{\sqrt{N_A}}{c} \) is scale invariant but it is still not normalized. The normalization must necessarily be empirical since the perceptual effort is an experimental quantity and clearly cannot be predicted theoretically. It is possible, however, to make some plausibility arguments as to the value of the normalization factor. A straightforward normalization is to take (for \( g=0 \)) \( \xi = 0 \) if the image has maximum representational compactness, i.e., the active area fills a disk. For a disk of radius \( r \), \( N_A = \pi r^2 \), \( c = 2\pi r \), and \( \frac{\sqrt{N_A}}{c} = (2\sqrt{\pi})^{-1} \). Hence the normalization factor is \( (2\sqrt{\pi}) \), and equation (1a) becomes

\[
\xi = \log \left[ \frac{m}{M/P} \left( \frac{2\sqrt{\pi}N_A}{c} \right) \right]
\]

To check the plausibility of the normalization factor, we could compare the effect of a change in the geometric bit density with an equivalent change in 'meaning' bit density (log \( \mu \)). In other words, we compare a change in \( \log \left[ \frac{m}{M/P} \right] \) with a change in \( \log \left( \frac{2\sqrt{\pi}N_A}{c} \right) \). If we consider the same disk of maximum representational compactness, doubling the circumference turns out to be equivalent to a decrease of 1 bit (in \( \mu \)) while keeping the circumference unchanged, which is a plausible trade-off.

Equation [7] is closely related to the more empirical definition of \( \xi \) given in section 2 but it is of more general applicability and it is restricted by Constraint 1.
7. Separation of meanings and effort of perception

So far, we have constructed a measure of semantic identity for a set of images, by taking into account the density of meaningful representations and the geometric compactness of these representations. Another factor, however, also affects the effort of perceiving the meaning of an image. So far, we have assumed that each image has a clear and distinct 'meaning', and 'meaning' was not specified further. As noted, the meaning of 'meaning' is an epistemological problem (Ogden and Richards, 1923) beyond the scope of the paper. But some basic notions related to 'meaning' can be quantified. First of all, the meaning conveyed by an image must be related to some properties of the image. Such properties are investigated in image processing (Gonzalez and Woods, 2001) and related areas of research. Examples of some of the simplest properties are colour, area, shape, and topology. Many other can be considered. In giving meaning to an image, one or more of these image properties must necessarily be 'measured' by the perception process. For example, if the set of images represents a deck of cards, colours and shapes of the images are necessary properties to convey the meaning.

Even if we do not specify further the property, or combination of properties, necessary to specify the meaning, the following considerations apply. Without loss of generality, assume that there is only one property whose measure determines the meaning. The set of images, in general, will have a distribution of the values of this property. It is intuitive that the more the images of the set differ from one another in this property, the easier it is to distinguish the meaning of one image from another. A trivial example: any image in the set of three images A {one red, one green, one blue} is easier to distinguish than any image in the set B {one red, one crimson, one firebrick} . Various measures for this notion of effort of meaning discernment can be thought of. A straightforward choice is the standard deviation $\sigma_m$ of the distribution of the meaning-conveying property. The smaller $\sigma_m$, the more effort in perceiving the meaning.
So far, we have assumed only one meaning per image. But more than one image may have the same meaning (i.e., \( m < M \)). In such a case, the variation in the meaning-conveying property has an additional, opposite effect. In fact, in a set of images that should convey the same meaning, any variability of the meaning-conveying property requires an additional perception effort. For example, suppose that in order to convey the three meaning: red, green, and blue a set of six images (two for each color) is used. If the two 'red' images are 'firebrick' and 'crimson' it will be easy to perceive that both convey the common meaning 'red'. If the two 'blue' images are 'lightskyblue' and 'navy' it will take more perceptual effort to attribute the same meaning (blue) to the two. Even more effort would be required for conveying the general meaning 'green' if 'lime' an 'darkolive' are the colours of the 'green' pair. Thus, we are led to consider a measure for the effort of perception due to what we may call *nuances of meaning*. The larger the variability in the nuances, the larger is the effort. Again, a straightforward way to such a measure is the use of the standard deviation; more precisely, the standard deviation \( \sigma_n \) of the distribution of the meaning-conveying property for the same meaning.

Combining the two contrasting tendencies of variability of meaning and variability of the nuances of meaning, we are led to include the ratio \( \sigma_m / \sigma_n \) (where \( \sigma_m \) and \( \sigma_n \) are measured as percentages of the respective averages) in the measure of semantic intensity given in equation (2). But again, as in the case of the geometric compactness, this factor also must be adjusted to be 'commensurate' with the density of meaning. Taking as standard a 2-pointer (1 bit), sharp separation of the two meanings (black-white or white-black) amounts to \( \sigma_m \sim 50\% \) and \( \sigma_n = 0 \). For this case to give zero bit contribution to the semantic intensity, the ratio \( \sigma_m / \sigma_n \) is modified as

\[
\frac{\sigma_m}{\sigma_n} \rightarrow \frac{2\sigma_m}{1 + \sigma_n}.
\]

So we take as a measure of effort due to variations in the distribution of meaning the factor \( v \).
Combining eqs (7) and (8) we finally obtain a measure of semantic intensity:

\[ \xi = \log \left( (mv) \left( \frac{2\sqrt{\pi N_A}}{c} \right) \right) \]  

where (mv) can be regarded as an 'effective' number of meanings, i.e., the number of meanings reduced by the factor v due to the difficulty of distinguishing one meaning from another and to perceiving the same meaning in images with different nuances.

The factor v introduced to take into account the effort of separating meaning is a measure analogous to any of the measures used in problems of classification and clustering. In such cases, the competition between separating elements of a set in distinct classes while maintaining similar elements in the same class is at the root of many measures of clustering validity (Halkidi et al., 2002).

8. Segmentation into objects

In the first part of the paper (sect. 2), we have considered images as segmented into objects. This segmentation led to the practical result (eq. (1) to eq. (5.2)) favouring semantroids over full-body avatar. In the second part, we have developed the notion of semantic intensity independently from the operation of segmentation. Now, having developed a better measure of semantic intensity for not-segmented images, we reconsider segmented images.

We have previously considered the example of one set of cards. Consider now two such decks in separate locations. Two cards (one for each deck) are visible at any time. Assume that the only significant images are the J Q K cards and that the meaning is irrespective of the suit. Then, \( P = 52^2 \) and \( M = 12^2 \), but what is m? For one deck it is 3. For two decks it depends on how we specify the meaning property. Let the meaning be specified (as in the one deck example) by the JQK cards irrespective of the suit. If we specify the meaning to be given by a pair of such cards then \( m = 3^2 \). If the meaning is specified instead by, e.g., the highest value of such a pair,
then m = 3. The first case correspond to having segmented the image into two objects (the two decks). The second consider the two decks as a single object. For the first case, since we have defined \( \xi \) logarithmically, the semantic intensities from the two decks add up.

As in the example above, in general, segmentation of the image into objects yield

\[
\xi = \sum_i \xi_i \tag{10}
\]

where \( \xi_i \) is the semantic intensity of one object and the sum extends over all the objects.

9. Application to semantroid

Using eqs. [9,10] we can now recalculate the semantic intensity for the semantroid considered in sect.3. Making equivalent assumptions, we regard the avatar as segmented in the same components as in Table 1. Further, making the simplest assumptions, each degree of freedom of each component can be assumed to be represented by a N-pointer, where N varies with the resolution of the dof. For example, the rotation of one joint of a finger could have a range of 150 degrees resolved into 10 steps of 15 degrees; thus a 10-pointer would describe this degree of freedom. For an estimate, let us assume that each dof of the hands is represented by a 10-pointer, and for the other parts, by a 5-pointer. We also can assume sharp separations in the representations so that, \( \sigma_m = 50\% \) and \( \sigma_n = 0 \), hence \( v = 1 \). Thus, since \( M=P=m \),

\[
(mv) \cdot \frac{(M/P)}{\xi} \sim 10^{dof} \tag{11}
\]

Values are recorded in Table 2 in a form that corresponds as closely as possible to that of Table 1. We have used the terms 'effort' and 'meaning' in the 4\(^{th}\) and 5\(^{th}\) columns with no significance except to indicate a correspondence with the values of Table 1.

For the geometric factor, the size of Table 1 corresponds directly to \( \sqrt{N_A} \). But the new formulation, eq.(9), of the semantic intensity does not depend anymore on the distance from the center of the image (i.e. the center of the display). It depends instead on the range of spatial excursion of each segment, which determines \( c \), the perimeter of the bounding box of the active
area. These are estimated as in column 3 of Table 2. The bounding box is taken to be the entire screen for all the parts, since it may be assumed that any part may at some time be in any part of the screen. Again we label the column as in Table 1 but here ‘distance’ stands for the perimeter of the bounding box.

The estimated values are summarized in Table 2 and yield again a higher semantic intensity for the semantroid. The difference is however much smaller than in the estimate of sect. 3. It should be remarked that these results are obtained as very crude estimates and are meant to have only the significance of pointing out the notion that a semantroid can be a more efficient way of conveying meaning than a full avatar.

<table>
<thead>
<tr>
<th>COMPONENT</th>
<th>‘Size’</th>
<th>‘Distance’</th>
<th>‘Effort’</th>
<th>‘Meaning’</th>
<th>Semantic intensity</th>
</tr>
</thead>
<tbody>
<tr>
<td>$J_i$</td>
<td>$\sqrt{N_A}$</td>
<td>c</td>
<td>$\log(2c^{-1}\sqrt{\pi N_A})$</td>
<td>$\log(mvMP^{-1})$</td>
<td>$\xi$ (eq.(9))</td>
</tr>
<tr>
<td>Head</td>
<td>4</td>
<td>160</td>
<td>-3.49</td>
<td>26 log5</td>
<td>56.882</td>
</tr>
<tr>
<td>Hand</td>
<td>3</td>
<td>160</td>
<td>-3.92</td>
<td>26 log10</td>
<td>82.426</td>
</tr>
<tr>
<td>Arm</td>
<td>5</td>
<td>160</td>
<td>-3.17</td>
<td>1 log5</td>
<td>-0.848</td>
</tr>
<tr>
<td>Forearm</td>
<td>5</td>
<td>160</td>
<td>-3.17</td>
<td>1 log5</td>
<td>-0.848</td>
</tr>
<tr>
<td>Trunk</td>
<td>10</td>
<td>160</td>
<td>-2.17</td>
<td>1 log5</td>
<td>0.152</td>
</tr>
<tr>
<td>Thigh</td>
<td>7</td>
<td>160</td>
<td>-2.68</td>
<td>1 log5</td>
<td>-0.358</td>
</tr>
<tr>
<td>Leg</td>
<td>6</td>
<td>160</td>
<td>-2.91</td>
<td>1 log5</td>
<td>-0.588</td>
</tr>
<tr>
<td>Foot</td>
<td>3</td>
<td>160</td>
<td>-3.92</td>
<td>1 log5</td>
<td>-1.598</td>
</tr>
</tbody>
</table>

| Avatar total | 213.406 |
| Semantroid    | 221.734 |

Table 2. Semantic intensity calculations for semantroid and avatar as in Section 3.
10. Summary and conclusion

As stated in the introduction, the goal of the paper is to take a first step in investigating the effect of drastic data reduction. By 'drastic' we mean something more than what is done, e.g., in lossy compression or in edge extraction. In the latter two cases, the basic image segmentation into objects is preserved. Our focus is instead on the effect of discarding entire objects (regions) in images. When objects are conserved (as in lossy compression and edge extraction) the benefit is primarily a gain in less memory for data storage and less bandwidth for transmission. Although some gain maybe achieved also in scene understanding, this is not necessarily the case. On the contrary, the main benefit in the drastic reduction of visual data obtained by, e.g., discarding objects, is a potential gain in processing speed for scene understanding. In scene understanding, there is always a trade off between the amount of information presented and the amount of effort required to make sense of it. This has been the general area of this paper; the specific focus has been the development of a quantitative measure for this trade-off.

Our main result is the proposed new measure of semantic intensity. We have given a rationale for its formulation and have applied it to justify the use of a semantroid to replace full bodied avatars. The rationale has been based on measures of density of 'meaning' and of amount of perceptual effort. These measures are plausible but are not intended to be the final word since they are simple measures to cover a very broad group of concepts. It is not expected that details will be properly taken into account at this stage. The basic ideas incorporated in the formulation of semantic intensity fall into two groups: those related to quantity of 'meaning' and those related to quantity of 'effort'. These notions are summarized below.

A) Notions related to quantity of 'meaning'

--We assume that in a set of P perceivable images (e.g. a visual database, a segment of a movie) there are $M \leq P$ meaningful images with a total of $m \leq M$ distinct meanings.
Meaning is defined as the value of a specified property (or combination of properties) of each image. Meaningless images are those without the specified property (e.g., black and white images are meaningless if the specified property is 'hue').

All images in the set are assumed to have one and only one meaning (ambiguity is not considered) and to be equally likely to be perceived.

A fundamental constraint is imposed on the way meaning is conveyed: meaning must be conveyed without requiring analysis or inference, i.e., meaning must be directly represented. This constraint is imposed in order to make the effort of perceiving the meaning a measurable quantity. Mental effort involved in analysis or inference is not quantifiable.

B) Notions related to quantity of effort in perceiving the meaning.

Effort in perceiving the meaning is affected by two main factors:

1. The spatial distribution of the active pixels, i.e. of the parts of the image that change from one member of the image set to another. The intuitive notion that it takes more effort to perceive widely scattered elements than more compactly located elements has been quantified by a compactness measure for the active pixels, eq. (7).

2. The distribution of the meaning-conveying property. As a measure of this effort we have chosen, eq. (8), basically a ratio of the standard deviation of the meanings to the standard deviation of the nuances of meaning, i.e. the different values of the meaning-conveying property for a given meaning. Other measures are conceivable, but generally the effect of the distribution of meanings should be taken into account in any measure of perceptual effort.

The notions summarized above are the basis for a quantitative definition of semantic intensity. The formulation of semantic intensity as eq. (9) may be regarded as a first attempt at including the notions basic to the definition of semantic intensity in a practically usable form.

The constraint on direct (i.e., non inferential) representation of meaning is closely related to concepts from Relevance Theory. The latter applies to the verbal/linguistic context. As such,
relevance, unlike semantic intensity, cannot be quantified. Words are symbols and not direct representations of meaning. Hence a mental (therefore unquantifiable) effort is always associated with a meaning conveyed by a word. Images do not have such limitations. They can represent meaning directly. Thus a theory of visual relevance can be quantitative. This paper is a first step in this direction. While the theory of visual relevance is developed, experimental studies of semantic intensity applied to semantroids for various practical situations (e.g. sign language) is the next subject of our investigations (Adamo-Villani and Beni, in preparation).

Semantic Intensity is a quantification of the trade-off between loss of information and gain of perceptual speed which occurs in drastic reduction of image data. It can be applied to a wide range of representational decisions besides the case presented here (avatar vs. semantroid). Examples of representational decisions with practical impact are: photography vs. drawing, drawing vs. diagram, color vs. b&w, 3D vs. 2D, and so on. In all such cases, a large amount of information is lost in the second option. The gain in perceptual speed may or may not compensate for this drastic loss of information. Calculations of semantic intensity are useful to decide which type of representation is more efficient.

Situations of drastic loss of information are relevant to Human Computer Interface theory, design, and evaluation. In particular, any HCI using graphic interactions will face tradeoffs of the type mentioned above. The notions of this paper may be considered as prerequisites for studying the efficiency of human-computer interactivity via graphic interfaces. In fact, any interactivity is necessarily preceded by the perception of the information presented by the computer interface to the human user. The efficiency of this perception affects the response of the user. Familiar interface tradeoffs can be analyzed in the semantic intensity perspective. For example, the tradeoffs of thumbnails vs. icons, or large vs. small icons, or one menu with N items vs. two menus with N/2 items each, or a linear vs. square palette of buttons, and so on. Such tradeoffs can be analyzed with the notion of semantic intensity to determine the perceptual efficiency of the graphic representation before being analyzed with the known tools of HCI (e.g.
John, 1995) for interactivity efficiency. Thus, semantic intensity may be regarded also as tool in the theory, design and evaluation of HCI.

References


