Computational Models for Predicting Visual Distinctness

Computer Vision Group[1*]

[1]Department of Computer Science and Artificial Intelligence, University of Granada, CITIC-UGR, 18071 Granada, Spain

[*]To whom correspondence should be addressed; E-mail: jags@decsai.ugr.es

Here we deal with two different situations in predicting visual target distinctness by means of a computer vision model. First, it is assumed that the structure of the target-and-background scene and the image without a target can be determined exactly. In this case the main problem is how to select relevant information into a limited attentional bottleneck. In the second case, it may happen that the structure of the target and non-target scenes cannot be determined exactly. Therefore, the structure of the images should be characterized statistically by discrete probability distributions.

Introduction

Measuring target acquisition performance in field situations is usually impractical and often very costly or even dangerous. Thus it is of great benefit to have advance knowledge of human visual target acquisition performance for targets or other relevant objects. However, search performance inherently shows a large variance, and depends strongly on prior knowledge of the perceived scene. Therefore, a typical search experiment requires a large number of observers to

obtain statistically reliable data.

Visual target acquisition is a complex process, and many factors involved are not yet fully understood. One thing is evident: the more a target stands out from its background the easier it will be to detect it, and the quicker it will be found. It is therefore likely that visual target distinctness is an important determinant of visual search performance.

Target saliency for humans performing visual search and detection tasks can be estimated by means of the difference between the image from the target-and-background scene and the image from the same background with no target. Thus, relevant computational models of early human vision typically process an input image through various spatial and temporal bandpass filters and analyze first-order statistical properties of the filtered images to compute a target distinctness metric. If they give good predictors of target saliency for humans performing visual search and detection tasks, they may be used to compute visual distinctness of image subregions (target areas) from digital imagery.

Here we deal with two different situations in predicting visual target distinctness by means of a computer vision model. First, it is assumed that the structure of the target-and-background scene and the image without a target can be determined exactly. In this case the main problem is how to select relevant information into a limited attentional bottleneck. Thus, References (*1–4*) introduce various computational vision models for selecting significant information for perceiving target distinctness in this first case. In the second case, it may happen that the structure of the target and non-target scenes cannot be determined exactly. Therefore, the structure of the images should be characterized statistically by discrete probability distributions. Due to the availability of a large number of measures, we have to know what postulates and properties should be satisfied by an information measure and then what is the amount of relative information between the respective distributions of the target image and the image with no target. Reference (*5*) analyzes these points.

Models of feature perception in distortion measure guidance

References (*1,2*) analyze the relation between two different problems in computer vision: what a proper model for identifying significant stimulus locations in an image is, and the comparative performance of selective measures and pixel-by-pixel error metrics. The natural relation between both problems arises from the fact that a proper selection of significant locations in the target image might be used to guide its comparison with another image through any reasonable metric.

In Reference (*1*), we study an approach (to improve the correlation between the subjective rating and the the mean square error —MSE— metric) in which the differences of the images to be compared are computed upon locations at which humans might perceive features in the reference image —for example, line features or step discontinuities. A visual model for feature perception is used to measure distortion between the target image and the image without the target. The actual success of the resulting distortion measure would then depend on both the validity of the vision model and which error metric was used in the perceptual domain. The problem is then to select a metric for image discriminability which corresponds to the human observer's evaluation. How conjunctions of features can be incorporated into such a metric is the next subject of Reference (*2*).

Computational measures based on space-frequency analysis

Reference (*3*) studies a different approach to improve the correlation between subjective rating and MSE. In this approach the non-target image to be compared with that of the reference is passed through an operator designated to compare the excitation levels of the non-target image to those of the target picture, with excitation levels being given by a set of active units tuned to particular orientation and spatial-frequency components. This paper investigates the

relationship between visual target distinctness in complex natural scenes measured by human observers, and two different computational visual distinctness measures computed from image representational models based on selectively filtered images and statistical features.

The *first measure* computes the structural dissimilarity between two related images filtered by a bank of spatial frequency and orientation selective (*log*-Gabor) filters. The *second measure* may be described in terms of two different stages: A "preattentive" stage, in which the image is selectively filtered by a bank of 2D *log*-Gabor filters, and an "integration" stage in which we integrate and compare the separable representations (i.e., statistical structure) at attentional locations.

Defining the notion of visual pattern

Up to this point in the interpretation of visual search tasks was the assumption that the detection of targets is determined by the feature-coding properties of low-level visual processing. Reference (*4*) presents a new distinctness measure that is applied at a much higher level of image representation than feature detection at the level of perceived shapes or surfaces. Instead of assuming that such forms are simple or integral features (i.e., statistical structure at a particular scale), we think it more appropriate to regard them as "visual patterns" distinguished at an object or surface level. To make such a distinction, a system for the automatically learned partitioning of "visual patterns" in the original reference image is given, based on a sophisticated, band-pass filtering operation, with fixed scale and orientation sensitivity. In this scheme, the "visual patterns" are defined as the features that have the highest degree of alignment in the statistical structure across different frequency bands. The analysis reorganizes the reference image according to a constraint of invariance in the statistical structure and consists of three stages: (i) Pre-attentive stage; (ii) integration stage; and (iii) learning stage. The first stage takes the reference image and performs filtering with a set of *log*-Gabor filters. Based on their re-

sponses, activated filters which are selectively sensitive to patterns in the image are short-listed. In the integration stage, common grounds between several activated sensors are explored. The filtered responses are analyzed through a family of statistics. For any given two activated filters, the distance between them is derived from distances between their statistics. The third stage, the learning stage, performs cluster partitioning as a mechanism for learning the subspace of *log*-Gabor filters needed to partition the image data.

A computational distinctness measure can then be computed from the images after they have been transformed into a new perceptual domain in which they are decomposed into their "visual patterns." The resultant model has perceptual access to "visual patterns" but not to filtered images or statistical features at a particular level of resolution. A main result will be finding that this computational measure that applies a simple decision rule between segregated visual patterns relates to visual target distinctness as perceived by human observers.

Information theoretic measures

It often happens that the structure of a certain scene cannot be determined exactly due to various reasons. Under such circumstances, the structure of the reference image and the input image can be characterized statistically by discrete probability distributions. Here we ask the following question: What is the amount of relative information between these discrete probability distributions?

Due to the availability of a large number of measures for this purpose, a question naturally arises about the criteria for the choice of the measure to be used in a particular investigation. For this goal, we have to know what postulates and properties should be satisfied by the information theoretic measure. It is therefore of great value to develop an axiomatic characterization of relative information for predicting visual target distinctness from 2D digital images. Reference (*5*) addresses exactly this point.

Experimental Results

In Reference (*6*), several experiments are performed to investigate the relationship between the different target distinctness measures and the visual target distinctness measured by human observers. First, a psychophysical experiment is performed in which human observers estimate the visual distinctness of targets in a database. The subjective ranking induced by the psychophysical target distinctness is adopted as the reference rank order. Second, the visual target distinctness is estimated by using the different measures between the original scene and an image of the same scene in which the target support has been artificially filled in with the local background. A relation is then established between the computational and the psychophysical target distinctness estimate, for each one of the measures.

There is a central claim that can be made based on References (*1–5*), but first we should elaborate a bit more about it.

We have performed an additional experiment in which all the measures from References (*1–5*) were compared to predict visual target distinctness on the dataset described in Figs. 1 and 2. The dataset in this experiment is composed of fifteen complex natural images containing a single target, $\{2, 4, 5, 9, 11, 12, 15, 18, 19, 24, 26, 27, 29, 32, 36\}$, and the corresponding fifteen empty images of the same rural backgrounds with no target. The value of the evaluation function for each computational measure (i.e., the corresponding fraction of correctly classified targets P_{CC}) is presented in Table 1. Also, the BC_a intervals with 90% confidence are given for each measure in Table 2.

From Tables 1 and 2 we conclude the obvious result that both the visual-pattern-based measure VP and the information theoretic measure $\mathcal{E}_C^{Z_1,\cdots,Z_n}$ show the best overall performance in the experiment. The statistical accuracy of this result is demonstrated based on the BC_a confidence intervals for the corresponding bootstrap sampling distributions of P_{CC}, as described

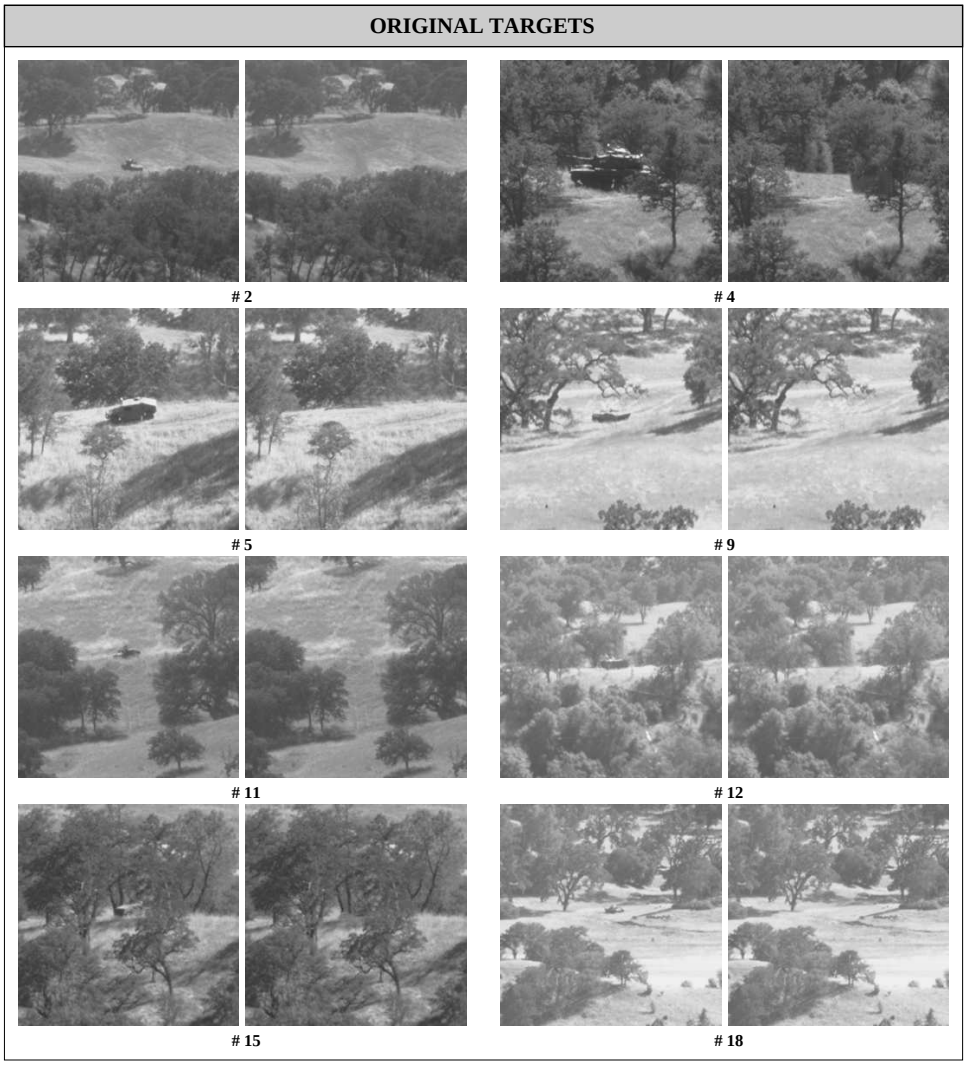

Figure 1: Original dataset of target and non-target scenes.

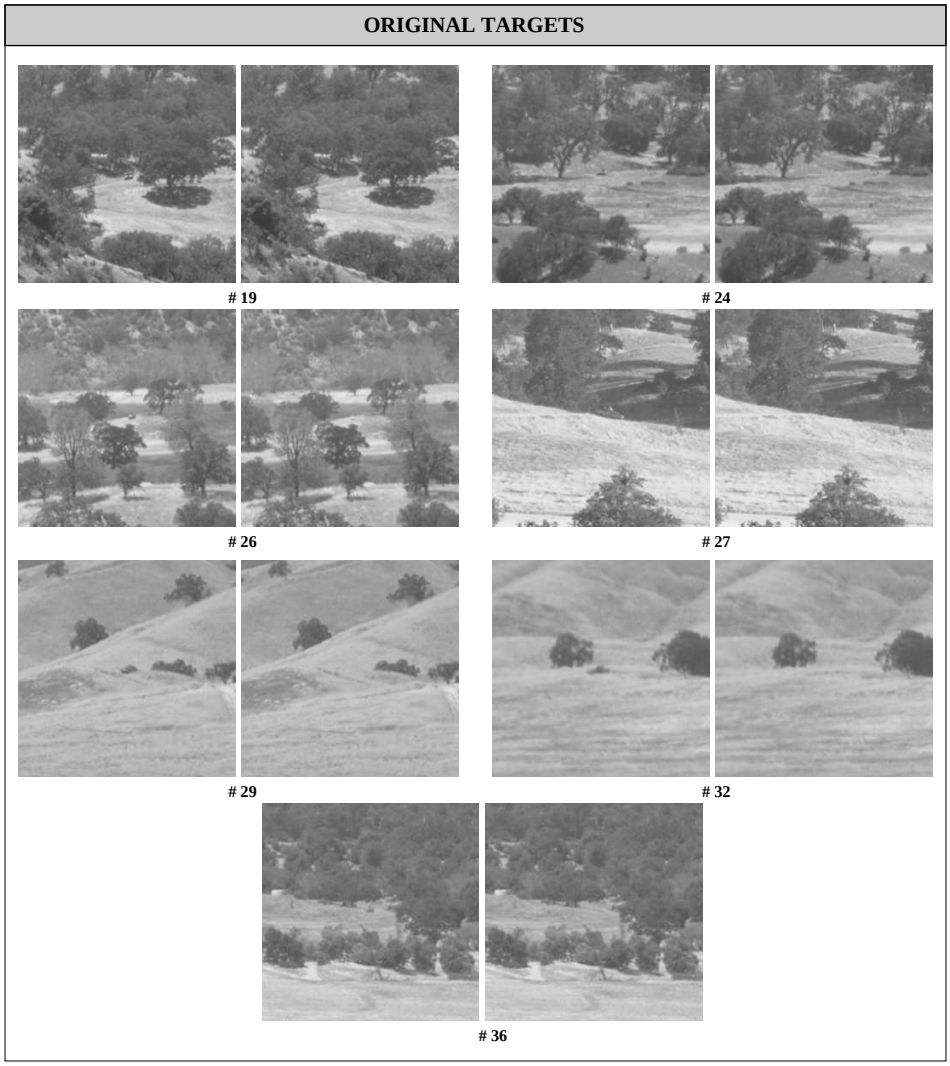

Figure 2: Original dataset of target and non-target scenes.

Table 1: Evaluation function for each computational measure

Measure	P_{CC}
FR	0.67
IF	0.53
VP	**0.73**
$\mathcal{E}$	0.47
$\mathcal{D}$	0.47
$\mathcal{E}^{Z_1,\cdots,Z_n}$	0.47
$\mathcal{D}_S$	0.40
$\mathcal{E}_C^{Z_1,\cdots,Z_n}$	**0.80**
$\mathcal{D}_C$	0.73
$\mathcal{D}_C^{Z_1,\cdots,Z_n}$	0.73
MAE	0.47
RMSE	0.53
SNR_{log}	0.60
SNR_{peak}	0.53

Table 2: BC_a intervals with 90% confidence for each measure

Measure	*BCa INTERVALS WITH 90% CONFIDENCE*
FR	$(0.006, 0.822)$
IF	$(0.038, 0.833)$
VP	$(0.400, 0.733)$
$\mathcal{E}$	$(0.067, 0.600)$
$\mathcal{D}$	$(0.067, 0.600)$
$\mathcal{E}^{Z_1,\cdots,Z_n}$	$(0.067, 0.600)$
$\mathcal{D}_S$	$(0.067, 0.467)$
$\mathcal{E}_C^{Z_1,\cdots,Z_n}$	$(0.533, 1.000)$
$\mathcal{D}_C$	$(0.200, 0.733)$
$\mathcal{D}_C^{Z_1,\cdots,Z_n}$	$(0.200, 0.733)$
MAE	$(0.200, 0.467)$
RMSE	$(0.133, 0.667)$
SNR_{log}	$(0.200, 0.600)$
SNR_{peak}	$(0.133, 0.667)$

in Table 2. But, given that the computational measures VP and $\mathcal{E}_C^{Z_1,\cdots,Z_n}$ come from two distinct approaches to predict visual distinctness, a natural question is: what is the meaning of this result?

As described in Reference (*4*), the VP measure is computed from the images after they have been transformed into a new perceptual domain in which they are decomposed into certain features (or "visual patterns") that have the highest degree of alignment in their statistical structure across a number of scales and orientations. These features are likely to be invariant over a particular range of scales and orientations and can be judged unlikely to be accidental in origin even in the absence of specific information regarding which objects may be present. In this sense visual patterns are said to be partially invariant features.

For the partitioning of a digital image into its visual patterns, Reference (*4*) proposed a frequency-based separation according to a constraint of congruency in statistical structure across frequency bands. Based on this definition of visual pattern as congruency in statistical structure, the differences between visual patterns from the target image and the corresponding visual patterns from the nontarget image determine the overall distinctness between the reference target image and the corresponding empty image by using a relatively simple decision rule, as described in Reference (*4*). Hence, the computational measure VP is a function of the discrepancy between congruency in statistical structure across scales and orientations.

On the other hand, the compound gain $\mathcal{E}_C^{Z_1,\cdots,Z_n}$ is a measure of information gain between target and non-target scenes such that Postulates 1–6 as stated in Reference (*5*) are satisfied. The interesting point is that the other information theoretic measures that are based on Postulates 1–5 only (i.e., $\mathcal{E}$, $\mathcal{D}$, $\mathcal{E}^{Z_1,\cdots,Z_n}$, and, $\mathcal{D}_S$) yield a relatively low probability in the last experiment ($P_{CC} < 0.5$), and hence they appear incapable of rank ordering targets in this experiment with respect to their visual distinctness. From this result we claim the essential role of Postulate 6 to improve the correlation with the visual target distinctness as predicted by humans.

Postulate 6 (i.e., the significance conservation constraint) states that if interest points of the target image and their significance is some given constraint on the nontarget image, then the selective information gain is also a function of the discrepancy between the significance of interest points in the target image and their significance in the non-target scene. Thus the compound gain $\mathcal{E}_C^{Z_1,\cdots,Z_n}$ is a function of the discrepancy between the significance of interest points in the target image and their significance in the nontarget scene.

In Reference (*5*) we have selected a local energy model for the perception of low-level features, and consequently, we assumed that features are perceived at points where the Fourier components are maximally in phase. Hence, in the implementation of the computational measure $\mathcal{E}_C^{Z_1,\cdots,Z_n}$, the significance of an interest point in any image was simply computed as the phase congruency across scales at this point. This means that the essential point in the definition of the information theoretic measure $\mathcal{E}_C^{Z_1,\cdots,Z_n}$ was the fact that it is a function of the discrepancy between the phase congruency across scales at interesting points (i.e., points where the Fourier components are maximally in phase) in the target image and the corresponding phase congruency in the nontarget image.

We now have the conditions to respond to the question: What is the meaning of the fact that both the visual-pattern-based measure VP and the information theoretic measure $\mathcal{E}_C^{Z_1,\cdots,Z_n}$ show the best overall performance in the target distinctness experiments —even though they come from distinct approaches? Both in the information gain $\mathcal{E}_C^{Z_1,\cdots,Z_n}$ and in the visual-pattern-based measure VP, we are just computing the discrepancy between certain types of congruency (i.e., partially invariant features). Hence we can now claim the central role of the comparison between these features in predicting visual target distinctness.

It should be pointed out that there are other pieces of evidence in this same line of reasoning. Thus, it is well known that the perceptual organization capabilities of human vision seem to exhibit the properties of detecting viewpoint-invariant structures and calculating varying degrees

of significance for individual instances (*7*). Following these ideas, Lowe (*8*) proposed that the structures to be detected in the image should be formed bottom-up using perceptual grouping operations that exhibit exactly these properties in the absence of domain knowledge, yet must be of sufficient specificity to serve as indexing terms into a database of objects. Given that we often have no prior knowledge of viewpoint for the objects in a database, these indexing features that are detected in the image must reflect properties of the objects that are at least partially invariant over a wide range of viewpoints of some corresponding three-dimensional structure. This means that it is useless to look for features with particular sizes or orientations, or other properties that are highly dependent upon viewpoint.

References and Notes

1. Xose R. Fdez-Vidal, J.A. Garcia, J. Fdez-Valdivia, and A. Garrido, "Using models of feature perception in distortion measure guidance," *Pattern Recognition Letters*, vol. 19 (1), pp. 77-88, (1998).

2. Xose R. Fdez-Vidal, J.A. Garcia, J. Fdez-Valdivia, Rosa Rodriguez-Sanchez, "The role of integral features for perceiving image discriminability," *Pattern Recognition Letters*, vol. 18, pp. 733-740, (1997).

3. Xose R. Fdez-Vidal, A. Toet, J.A. Garcia, J. Fdez-Valdivia, "Computing visual target distinctness through selective filtering, statistical features, and visual patterns," *Optical Engineering*, vol. 39 (1), pp. 267-281, (2000).

4. Xose R. Fdez-Vidal, J.A. Garcia, J. Fdez-Valdivia, and R. Rodriguez-Sanchez, "Defining the notion of visual pattern for predicting visual target distinctness in a complex rural background," *Optical Engineering*, vol. 39 (2), pp. 415-429, (2000)

5. J.A. García, J. Fdez-Valdivia, X. R. Fdez-Vidal, and R. Rodriguez-Sánchez, "Information theoretic measure for visual target distinctness," *IEEE Trans. Patt. Anal. Mach. Intell.,* Vol. 23(4), pp. 362-383 (2001).

6. J.A. Garcia, J. Fdez-Valdivia, X.R. Fdez-Vidal, and R. Rodriguez-Sanchez, *Computational models for predicting visual target distinctness*, SPIE Optical Engineering Press, PM-95, Bellingham, Washington USA, (2001).

7. M. Wertheimer, "Principles of perceptual organization," In *Readings in perception*, pp. 115-135, Van Nostrand, Princeton, NJ, (1958).

8. D.G. Lowe, "Three-dimensional object recognition from single two-dimensional images," *Artificial Intelligence*, vol. 31, pp. 355-395, (1987).

Acknowledgments. This paper was sponsored by the Spanish Board for Science and Technology (MICINN) under grant TIN2010-15157. It is a pleasure to acknowledge the significant and pervasive contribution of Dr. Alexander Toet to the discipline of target distinctness. His papers and technical reports at TNO Human Factors Research Institute in Soesterberg contained the basis of many of the experimental design used in our work. We are enormously grateful for the many hours Dr. Toet spent on reading our manuscripts and his suggestions for improvements. There is no way to thank him enough for such generous help.