Image representational model for predicting visual distinctness of objects

Xose R. Fdez-Vidal
Depto. Física Aplicada. Facultad de Física
Univ. de. Santiago de Compostela.
15706 Santiago de Compostela. Spain
faxose@usc.es

R. Rodriguez-Sánchez, J. Mtnez-Baena, J. Chamorro
Depto. Ciencias de la Computación e I.A.
E.T.S. de Ingeniería Informática.
Univ. de Granada. 18071 Granada. Spain
{rosa, jbaena, jesus}@decsai.ugr.es

Abstract

Here we show that a notion of congruence in statistical structure across 2D frequency bands produces a useful definition of visual patterns for perceiving target distinctness. In order to reach such a conclusion, firstly, the notion of congruence is used to induce a representational model for 2D images. Secondly, the visual-pattern based representational model is used to define a visual target distinctness metric that involves applying a simple decision rule over the distances between the visual patterns. Finally, a relation is established between the computational distinctness metric and psychophysical target distinctness estimates.

1. Introduction

Often implicit in the interpretation of visual search tasks is the assumption that the detection of targets is determined by the feature-coding properties of low-level visual processing [11]. Instead of assuming that perceived shapes are simple or statistical structure at a particular scale, we think it more appropriate to regard them as "visual patterns", distinguished at an object level. The "visual patterns" are simply defined as features which have the highest degree of congruence in statistical structure across different frequency bands.

Fig. 1 illustrates an example of the decomposition of a complex image containing a single military vehicle into its "visual patterns".

Firstly, the segregation of the clumps of energy in the amplitude spectrum induces the selection of a subset of activated filters (which are selectively sensitive to them) from a filter bank of logGabor functions centered at 12 orientations and 5 ranges. Due to conjugate symmetry, the filter design is only carried out on half the 2D frequency plane. Secondly, for any two activated filters, noted as ϕ_i and ϕ_j, their responses are compared based on the distance (a β-norm) between their statistical structure, computed over those pixels which form "fixation points" of the filters (local energy peaks on the filtered response).

Next, clustering on the basis of the distance between the filtered responses is performed in the set of activated filters, noted as *Active*, to highlight scale invariance of responses. As shown in the box entitled "Natural Clusters", the set of activated filters is then clustered into two natural groupings. Finally, for each grouping of activated filters, their filtered responses are summed for the automatic learned partitioning of the "visual patterns".

This paper analyzes whether this notion of congruence in statistical structure across frequency bands may be a useful definition of visual pattern for perceiving object distinctness. To this aim, the notion of congruence (Section 2) is firstly used to induce a representational model in which the image is decomposed into its visual patterns (Section 3). Fig. 2, shows several examples of the decomposition of images of a military vehicle (target) in a complex rural background.

The visual-pattern based representational model will be then used to define a distinctness measure between two images that involves applying a simple decision rule over the distances between their visual patterns. The distinctness metric will predict the target distinctness by the difference between the signal from the target-and-background scene and the signal from the background-with-no-target. A schematic overview of the metric is shown in Fig. 3. Several experiments are then performed to investigate the relation between this computational distinctness metric and the visual target distinctness measured by human observers (Section 4). Finally, a relation is established between the computational and the psychophysical target distinctness estimates.

2. The notion of "Visual Pattern"

A bank of filters should be employed to firstly decompose the original image into its most significant components (see Fig. 1). Here we consider a bank of logGabor filters which follows the description of visual spatial pattern analysis documented in [3, 4]. LogGabor functions are adopted as an appropriate method to construct filters of arbitrary bandwidth.

Let ϕ_i be a logGabor filter that can be represented as

0-7695-0750-6/00 $10.00 © 2000 IEEE

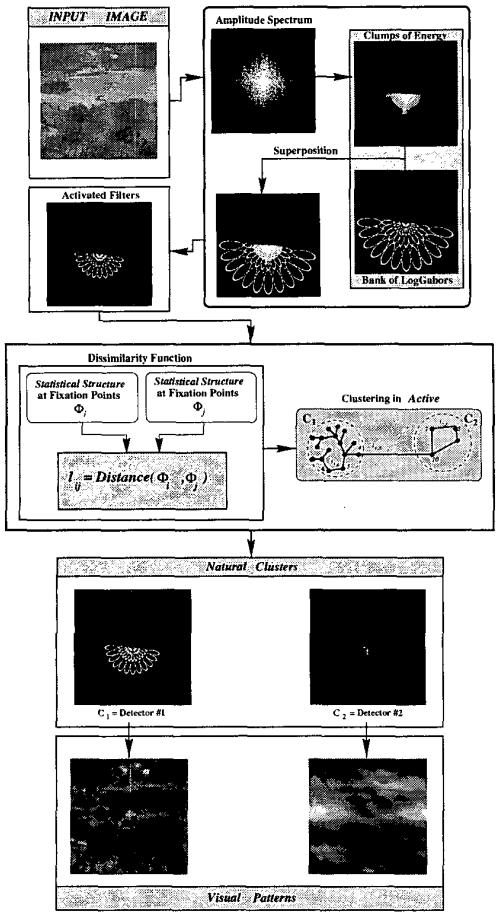

Figure 1. A general diagram describing the image representational model.

a Gaussian in the spatial frequency domain around some central frequency (r_i, θ_i), where θ_i is the orientation angle of the filter, and r_i is its central radial frequency:

$$\phi_i(r,\theta) = exp\left\{-\frac{\log^2 \frac{r}{r_i}}{2\log^2 \frac{\sigma_{r_i}}{r_i}}\right\} exp\left\{-\frac{(\theta-\theta_i)^2}{2\sigma_{\theta_i}^2}\right\} \quad (1)$$

with σ_{θ_i} and σ_{r_i} being the angular and radial sigma of the Gaussian around (r_i, θ_i), respectively. The bank of the filters should be designed so that it tiles the frequency plane uniformly (the transfer function must be a perfect bandpass function): The spatial frequency plane is divided into 12 different orientations, the radial axis is divided into 5 equal octave bands. In a band of width 1 octave, spatial frequency increases with a factor 2. The highest frequency filter (for each direction) is positioned near the Nyquist frequency to avoid ringing and noise. The wavelength of the five filters in each direction is set at 3, 6, 12, 24, and 48 pixels, respectively. The radial bandwidth is chosen as 1.2 octaves and the angular bandwidth is chosen as 15 degrees.

In order to decompose the image into its most significant components, strongly responding filters should be selected for the input image. Let *Active* be the set of filters in the bank that strongly respond to the spatial information content in the original image. They are units such that their amplitude spectrum and some clump of energy in the image amplitude spectrum overlap to some extent (see Fig. 1). [8] suggests a simple analysis for identifying this subset of activated filters from the filter bank.

Given a decomposition of the original image in its most significant components, only a further element is needed to define the concept of "visual pattern": a distance measure between the statistical structure at different orientations and scales.

Let $Distance(\phi_i, \phi_j)$ be a distance between the statistical structure of the filtered responses for ϕ_i and ϕ_j, computed over those pixels which form "fixation points" of the filters. For each activated filter, pixels whereupon the focus of attention should be shifted to measure visual distinctness, and which can therefore be regarded as interest or "fixation" points, are computed as local energy peaks on the corresponding filtered output of the original image [7]. In [8], *Distance* is given as a β-norm between the statistical structure at fixation points.

Given the metric $Distance(\phi_i, \phi_j)$, *a "visual pattern" is simply defined as congruence in statistical structure, as measured by* $Distance(\phi_i, \phi_j)$, *across a range of 2D spatial frequency bands.*

The individual filters spanning this particular range of bands will determine a natural cluster of units, noted as C_n, in the set of activated logGabors *Active*. By taking into account the statistical congruence across this range of frequency bands, a pair of filters ϕ_i and ϕ_j will belong to the same natural cluster C_n if there exists certainly continuity (i.e., there exists similarity in some statistics at the same spatial locations) across the filtered responses for an intermediate sequence of filters, between ϕ_i and ϕ_j, in C_n.

Therefore the definition of "visual pattern" induces a partition in *Active* into a number of natural clusters $C_1, C_2, \cdots, C_N$ such that:

$$Active = \bigcup_{n=1}^{N} C_n \ , \quad \text{and } C_p \cap C_q = \emptyset, \quad (2)$$

with $p \neq q$, $p, q = 1, 2, \cdots, N$ where, for each C_n, a pair of filters $\phi_i, \phi_j \in C_n$ if there exists a sequence of filters $\phi_{n_1}, \phi_{n_2}, \cdots, \phi_{n_l}$ in C_n such that

$$\begin{aligned} Distance(\phi_i, \phi_{n_1}) &\leq \varepsilon_n \\ Distance(\phi_{n_l}, \phi_j) &\leq \varepsilon_n \\ Distance(\phi_{n_k}, \phi_{n_{k+1}}) &\leq \varepsilon_n \ , \ k = 1, 2, \cdots, l-1 \end{aligned} \quad (3)$$

where ε_n denotes the degree of statistical congruence between a pair of filters in C_n and verifies that:

$$Distance(\phi_p, \phi_q) > \varepsilon_n \ , \ \forall \phi_p, \phi_q : \quad (4)$$

$\phi_p \in C_n$, $\phi_q \in Active - C_n$

The clustering of activated filters is performed as described in [8].

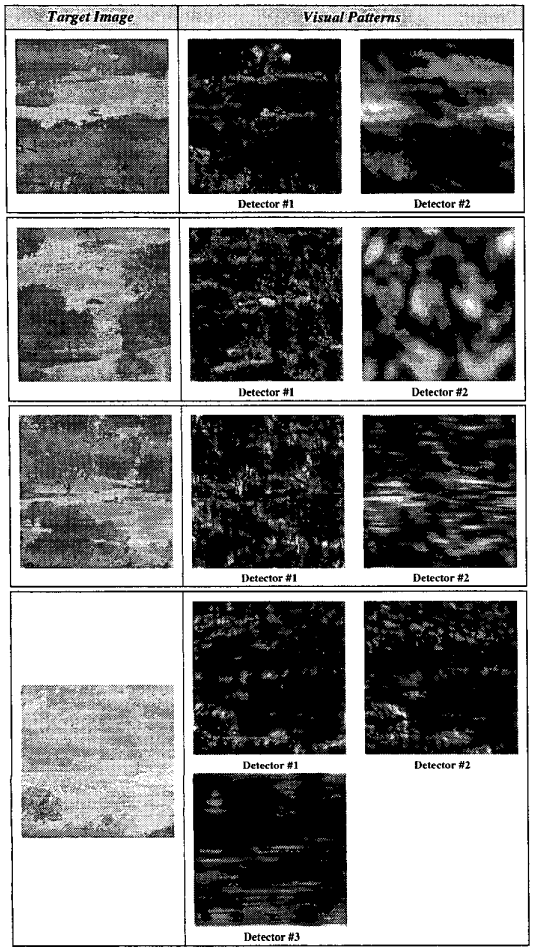

Figure 2. Decomposition performed on several images of a military vehicle (target) in a complex rural background. The original target images and their corresponding visual patterns.

3. Visual-Pattern based Image Representation

Given an input image $t(x,y)$, let $C_n = \{\phi_{n_j}\}$, $n = 1, 2, \cdots, N$, be the N natural clusters in $Active$, such that $Active = \bigcup_{n=1}^{N} C_n$.

Let t_n represent the visual pattern segregated on the input image $t(x,y)$ by pooling the responses of filters in the natural cluster $C_n = \{\phi_{n_j}\}$:

$$t_n = |\sum_j A_{n_j}| \tag{5}$$

where A_{n_j} denotes the original image $t(x,y)$ filtered through the logGabor ϕ_{n_j} in C_n and passed through a nonlinearity of the form:

$$\tanh(z,\tau) = \frac{1 - \exp\{-z\tau\}}{1 + \exp\{-z\tau\}} \tag{6}$$

where τ is a gain term [5]. This nonlinearity enables the system to respond to local contrast over several log units of illumination changes.

Therefore $t_1, t_2, \cdots, t_N$ represent a decomposition of the original image t into the set of its most significant visual patterns.

Fig. 2, shows the decomposition performed on several images of a military vehicle (target) in a complex rural background. The original target images and their corresponding visual patterns, as derived by the proposed model, are illustrated in this figure.

4. Predicting visual target distinctness

Rohaly [9] recently showed that image discrimination models, that quantify the visibility of the differences between a pair of images, can predict the visual distinctness of objects in natural backgrounds. The experiment reported in this section is performed to test if a computational metric that applies a simple decision rule to the distances between segregated visual patterns, also predicts visual target distinctness perceived by human observers.

The approach is as follows. First, a psychophysical experiment is performed in which observers estimate the visual distinctness of the target in each of 44 different test scenes (Section 4.2). Then, a computational metric is applied to quantify the visual distinctness of the targets in each one of two different datasets (Section 4.3). Finally, a relation is established between the computational and psychophysical target distinctness estimates (Section 4.4).

4.1. Images

The images used in the computational experiments are scenes representing a military vehicle in a complex rural background. Images are subsampled to 256 × 256 pixels. For each scene t, containing a target (vehicle), a corresponding empty scene e was created. The empty scene is everywhere equal to the target scene, except at the location of the target, where the target support is filled with the local background. The visibility of the targets varies throughout the entire stimulus set. This is mainly due to variations in the structure of the local background, the viewing distance, the luminance distribution over the target support (shadows), the orientation of the targets, and the degree of occlusion of the targets by vegetation.

The digital images used in the experiment were (see Fig. 4): (i) eleven target images that correspond to the scenes 16, 34, 28, 17, 26, 19, 8, 15, 3, 22, and 4, from the 44 slides, and (ii) the corresponding eleven empty scenes.

4.2. Psychophysical target distinctness

A psychophysical experiment was performed in which observers estimate the visual distinctness of the target. Search times and cumulative detection probabilities were measured for nine military targets in complex natural backgrounds. A total of 64 civilian observers, aged between 18 and 45 years, participate in the visual search experiment. The procedure of the search experiment is described in [6].

Search performance is usually expressed as the cumulative detection probability as function of time, and it can be approximated by [10]:

$$P_d(t) = \left\{ \begin{array}{ll} 0 & t < t_0 \\ 1 - \exp\left\{-\frac{t-t_0}{\rho}\right\}, & t \geq t_0 \end{array} \right\} \quad (7)$$

where, $P_d(t)$ is the fraction of correct detections at time t, t_0 is the minimum time required to response, and ρ is a time constant.

Fig. 5 shows the cumulative distribution functions corresponding to the search times measured for the target scenes used in the experiment here described. The overall difference between two of these functions can be measured by subtracting the area beneath their graphs. This operation corresponds to a Kolmogorov-Smirnov (K-S) test. To compare the relative distinctness of the targets in the different target scenes the curves are rank-ordered according to the area beneath their graphs. The resulting rank order for the target scenes in each of the two experiments is listed in the column with the header R_{P_d} in Table 1. These rank orders are adopted in each experiment as the reference standard for the evaluation of the computational metric.

Targets that give rise to closely spaced cumulative detection curves which are similar in accordance with a K-S test, have similar visual distinctness. Fig. 5 shows that the target images in each of the two experiments are clustered into a number of sets of targets with comparable visual distinctness: The dataset in the experiment is clustered into $\{16, 34\}$, $\{28, 17, 26\}$, $\{19, 8, 15, 3\}$, and $\{22, 4\}$. Hence, rank order permutations of elements of the same cluster are not very significant, whereas rank order permutations of elements of different clusters are therefore significant.

4.3. Computational Target Distinctness

Let $C_n = \{\phi_{n_j}\}$, with $n = 1, 2, \cdots, N$, be the N natural clusters in *Active* produced by the proposed model for the target image $t(x, y)$.

Let t_n represent the visual pattern segregated on the reference target image $t(x, y)$ by pooling the responses of filters in the natural cluster $C_n = \{\phi_{n_j}\}$ as given in equation (5). Therefore $t_1, t_2, \cdots, t_N$ represent a decomposition of the reference target image t into the set of its most significant visual patterns.

In order to compensate for the effect of image-to-image variations on the overall image light level, contrast normalization of each visual pattern is realized by dividing t_n by

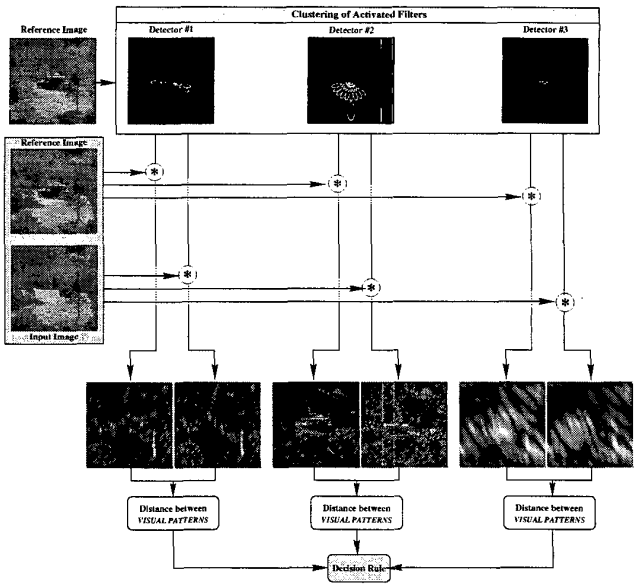

Figure 3. A schematic overview of the VP metric.

the sum of all filtered responses in *Active*, plus a saturation constant σ:

$$\frac{t_n}{\sigma^2 + \sum_i |A_i|} \quad (8)$$

where A_i denotes the original image $t(x, y)$ filtered through the logGabor ϕ_i in *Active* and passed through a non-linearity as given in equation (6).

Similarly passing the corresponding empty image $e(x, y)$ through the filters associated with each cluster C_n produced by the model on the reference image $t(x, y)$, results in a decomposition of e in $e_1, e_2, \cdots, e_N$.

Let $d_{VP}(t_n, e_n)$ be the difference between the visual patterns t_n and e_n, computed via the β-norm between their statistical structure over those pixels which form "fixation points" on t_n [1]: $d_{VP}(t_n, e_n) =$

$$\frac{1}{Card[FP(t_n)]} \left(\sum_{(x,y) \in FP(t_n)} \left| D[T^{t_n}(x, y), T^{e_n}(x, y)] \right|^{\beta} \right)^{\frac{1}{\beta}} \quad (9)$$

with $FP(t_n)$ being the set of fixation points for t_n; and $D[T^{t_n}(x, y), T^{e_n}(x, y)]$ defining a normalized distance measure between the vectors of statistics $T^{t_n}(x, y)$ and $T^{e_n}(x, y)$ at a fixation point (x, y). The default value of the exponent β in Equation (9) is 3.

The differences, $D_n = d_{VP}(t_n, e_n)$ with $n = 1, 2, \cdots, N$, between the visual patterns determine the overall distinctness between the reference target image t and the corresponding empty image e, by using a relatively simple decision rule. Two different decision rules can be considered here: (i) the average sum-of-differences rule, where the

system bases its response on the average sum of the differences $\frac{1}{N}\sum_{n=1}^{N} D_n$; and (ii) the maximum-difference rule, where the system bases its final response on the maximum of the differences $\max_{n=1}^{N} D_n$, rather than on the sum of differences. These simple rules are presumably adequate, because they are good descriptions of what all the complicated higher level stages of pattern vision actually contribute to visual target distinctness [4].

The visual pattern (VP) distinctness measure between reference target image t and empty image e can then be formulated as [2]:

$$VP(t,e) = \text{Decision-Rule}\{D_1, D_2, \cdots, D_N\} \qquad (10)$$

A schematic overview of the VP distinctness metric is given in Fig. 3.

4.4. Results

The digital images used in this experiment are eleven target images and the corresponding empty scenes as shown in Fig. 4.

The comparative results of the computational metric and those of both quantitative and qualitative measures are presented in Table 1.

At the bottom of each of the columns is shown the probability of correct classification of the metric in that column with respect to the reference rank order in column 2. The probability of correct classification P_{CC} is defined as:

$$P_{CC} = \frac{\text{Number of Correctly Classified Targets}}{\text{Number of Targets}},$$

where rank order permutations of targets of the same cluster are insignificant (i.e., they are correctly classified by the metric), whereas rank order permutations of elements of different clusters are significant.

The target distinctness values and the resulting rank order computed by the $RMSE$ metric are listed in column 3. $RMSE$ is defined as:

$$RMSE(R,I) = \sqrt{\frac{1}{N \times M}\sum_{x=1}^{N}\sum_{y=1}^{M}(R(x,y) - I(x,y))^2}. \qquad (11)$$

where $R(x,y)$ represents the target-and-background scene and $I(x,y)$ represents the background with no target. $RMSE$ performs poorly. Significant rank order permutations are displayed in boxes. Most rank orders computed by this metric are significantly out of order relative to the reference order induced by the psychophysical distinctness measure in column 2. The $RMSE$ yields a low probability of correct classification ($P_{CC} = 0.55$).

The target distinctness values and the resulting rank order computed by the VP metric are listed in the column with the header VP in Table 1. The VP metric induces a rank order with two significant order reversals: targets 26, and 3 are ordered incorrectly. The other targets have been attributed rank orders which do not differ significantly from the reference rank order based on the psychophysical measure. The VP metric yields the highest probability ($P_{CC} = 0.82$). These results show that the VP metric appears as well capable to rank order targets in the experiment with respect to their visual distinctness.

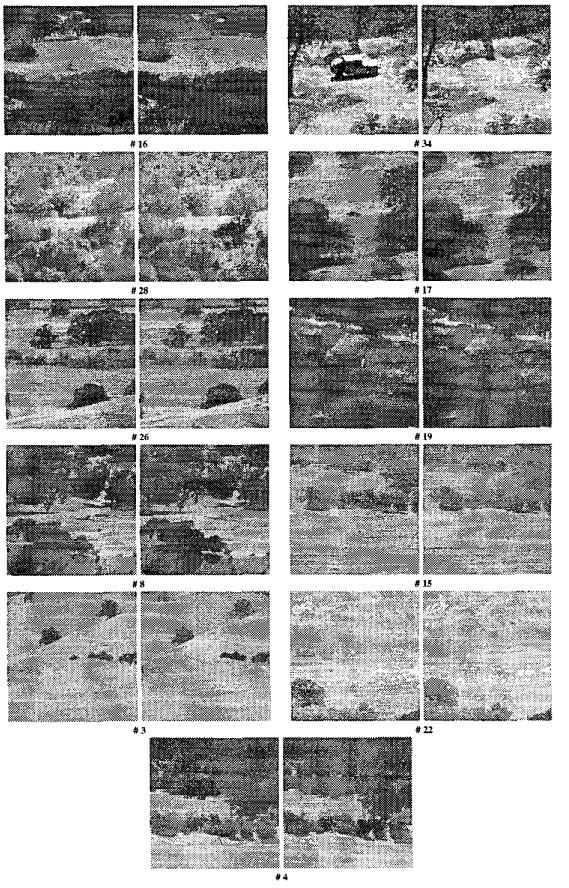

Figure 4. The 11 images and corresponding empty scenes used in the experiment.

5. Conclusion

The notion of visual pattern as congruence in statistical structure across 2D frequency bands was here used to induce a 2D image representational model. The induced representational model was used to define a distinctness measure between two images that involves applying a simple decision rule over the distances between their visual patterns. Several experiments were then performed to investigate the relation between this computational distinctness metric and the visual target distinctness measured by human observers. The results demonstrated that simple image metrics ($RMSE$) do not give good predictive results when applied to highly resolved targets in complex background scenes.

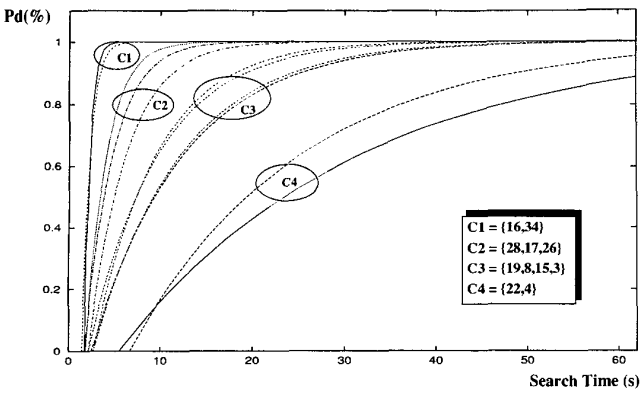

Figure 5. The target images divided into 4 clusters with comparable visual distinctness (in accordance with a K-S test).

Image Pair	R_{P_d}		*RMSE*		*VP*	
	Value	Rank	Value	Rank	Value	Rank
# 16	96.39	1	2.55	6	5.15	2
# 34	96.37	2	30.48	1	32.99	1
# 28	93.48	3	3.86	2	3.19	5
# 17	92.42	4	3.68	3	3.71	4
# 26	90.06	5	1.71	7	2.20	9
# 19	84.79	6	3.30	4	2.98	7
# 8	84.47	7	1.57	8	2.52	8
# 15	80.46	8	1.17	9	3.15	6
# 3	80.02	9	2.83	5	4.16	3
# 22	60.96	10	1.02	10	2.14	10
# 4	53.89	11	0.98	11	0.77	11
P_{cc}	-		**0.55**		**0.82**	

Table 1. On column 1: the dataset in the experiment; on column 2: the reference rank order; on columns 3 and 4: the target distinctness values and the resulting rank order computed by the computational metrics.

On the contrary, the distinctness metric that was computed from the images after they were decomposed into their "visual patterns" closely related to visual target distinctness as perceived by human observers. We conclude from this result that the notion of congruence in statistical structure across frequency bands may be a useful definition of visual patterns for perceiving target distinctness.

Acknowledgments. The authors thank Alexander Toet (TNO Human Factors Research Institute, The Netherlands) for providing us with the search times and cumulative detection probabilities from search experiments. This work was sponsored by the Dirección General de Enseñanza Superior (DGES) under grant PB98-1374 and the Spanish Board for Science and Technology (CICYT) under grant TIC97-1150.

References

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

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

[3] Field, D.J. "Relations between the statistics of natural images and the response properties of cortical cells." *Journal of The Optical Society of America A*, Vol. 4(12), pp. 2379-2394, (1987).

[4] Graham, Norma. *Visual Pattern Analyzers*, Oxford Psychology Series, No. 16, Oxford University Press. (1989).

[5] Malik, J., and Perona, P. "Preattentive texture discrimination with early vision mechanisms", *J. Optical Soc. America A.*, Vol. 7, pp. 923-932, (1990).

[6] Martinez-Baena, J., Toet, A., Fdez-Vidal, X.R., Garrido, A., Rodriguez-Sanchez., Rosa "Computational visual distinctness metric", *Optical Engineering*, Vol. 37, No. 7, pp. 1995-2005, (1998).

[7] Morrone, M.C. and Burr, D. C. "Feature detection in human vision: A phase-dependent energy model." *Proc. R. Soc. Lond. B*, Vol. 235, pp. 221-245, (1988).

[8] Rodriguez-Sanchez, R., Garcia, J.A., Fdez-Valdivia, J., Fdez-Vidal, X.R. "The RGFF representational model: a system for the automatically learned partitioning of "visual patterns" in digital images". *IEEE Trans. on Pattern Analysis and Machine Intelligence*, Vol. 21 (10), pp. 1044-1073, (1999).

[9] Rohaly, A.M., Ahumada, A.J. and Watson, A.B. "Object detection in natural backgrounds predicted by discrimination performance and models", *Vision Research* Vol. 37, pp. 3225-3235, (1997).

[10] Waldman, G., Wootton, J., and Hobson, G. "Visual detection with search: an empirical model." *IEEE Trans. on Systems, Man and Cyb.*, Vol. 21. 596-606, (1991).

[11] Zijiang, J. H., and Nakayama, K. "Surfaces versus features in visual search." *Nature*, Vol. 359, pp. 231-233, (1992).