# Image representational model for predicting visual distinctness of objects


XOSE R. FDEZ-VIDAL

Departamento de Física Aplicada

Facultad de Física

Universidad de Santiago de Compostela

15706 Santiago de Compostela. Spain.

ROSA RODRIGUEZ-SANCHEZ

Departamento de Informática.

Escuela Politécnica Superior.

Universidad de Jaen. 23071. Jaen. Spain.

J.A. GARCIA, J. FDEZ-VALDIVIA

Departamento de Ciencias de la Computación e I.A.

E.T.S. de Ingeniería Informática

Universidad de Granada

18071 Granada. Spain.

*Corresponding author:* J.A. Garcia

Departamento de Ciencias de la Computación e I.A.

E.T.S. de Ingeniería Informática.

Universidad de Granada. 18071 Granada. Spain.

*e-mail* : jags@decsai.ugr.es

*fax* : +34.958.243317

*URL* : http://decsai.ugr.es/~jags/

## 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, a connection between illusory percepts and statistical regularities across frequency bands is uncovered. Thirdly, 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.



## KEY WORDS

Visual distinctness of objects, representational model for images, integration of features, computational target distinctness metric, psychophysical target distinctness.


## 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.[1,2] 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. The logGabors are illustrated in the diagram by ellipses drawn, in the 2D spatial-frequency plane, at the point where their amplitude has decreased to the $(e^{-1/2})$ half width its maximum.

Secondly, for any two activated filters, noted as $\phi_i$ and $\phi_j$, their responses are compared based on the distance (a $\beta$-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 4). Figs. 2, and 3, show several examples of the decomposition of images of a military vehicle (target) in a complex rural background. Results are also provided (Section 5) to indicate the performance of this notion in the integration of image information for features that are separated in space (see Fig. 4).

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. 10. Several experiments are then performed to investigate the relation between this computational distinctness metric and the visual target distinctness measured by human observers (Section 6). Finally, a relation is established between the computational and the psychophysical target distinctness estimates (Section 7).

## 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 $\phi_i$ be a logGabor filter that can be represented as a Gaussian in the spatial frequency domain around some central frequency $(r_i, \theta_i)$, where $\theta_i$ is the orientation angle of the filter, and $r_i$ is its central radial frequency:

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

with $\sigma_{\theta_i}$ and $\sigma_{r_i}$ being the angular and radial sigma of the Gaussian around $(r_i, \theta_i)$, 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 sigma, $\sigma_{r_i}$, is chosen as 0.72 octaves.
- The angular sigma, $\sigma_{\theta_i}$, is chosen as 8.98 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). Appendix A 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 $\phi_i$ and $\phi_j$, 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.[5,6] In Appendix B, *Distance* is given as a $\beta$-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 $\phi_i$ and $\phi_j$ 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 $\phi_i$ and $\phi_j$, 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:

$$Active = \bigcup_{n=1}^{N} C_n \ , \quad \text{and } C_p \cap C_q = \emptyset, \text{ with } \ p \neq q, \ p, q = 1, 2, \cdots, N \tag{2}$$

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

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

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

$$Distance(\phi_p, \phi_q) > \varepsilon_n \ , \quad \forall \phi_p, \phi_q : \phi_p \in C_n, \ \phi_q \in Active - C_n \tag{4}$$

The clustering of activated filters is performed as follows.

## 3 CLUSTERING OF ACTIVATED FILTERS

We formulate the problem as the clustering of a dataset $X = \{i \mid \phi_i \in \mathit{Active}\}$ into a number $N$ of natural clusters $\mathcal{C}_0, \mathcal{C}_1, \cdots, \mathcal{C}_{N-1}$. We call clusters natural if the membership is determined fairly well in a natural way by the data.

This clustering is reduced to a sequence of stages of simpler partitioning. At each stage $j$, a subset $X^j$ of $X$ is divided into only two classes (for $j = 0$, $X^0 = X$):

(i) a natural cluster $\mathcal{C}_j$ which contain all the data points (filters) in $X^j$ which are assigned the same class of a seed point (filter) $seed_j$, with $seed_j$ being picked randomly from $X^j$, and

(ii) the data, $X^j - \mathcal{C}_j$, still not placed in any existing cluster, $\mathcal{C}_0, \mathcal{C}_1, \cdots, \mathcal{C}_{j-1}$.

The clarity of separation between clusters, as measured by a dissimilarity function, will be the criterion by which we derive a natural cluster $\mathcal{C}_j$ at stage $j$. The dissimilarity function is defined in Section 3.1. The criterion by which we define a natural cluster at stage $j$ is presented in Section 3.2.

The dynamic process of clustering is stopped at stage $j$, if the class $X^j - \mathcal{C}_j$ is the empty set. Otherwise, the process progresses, and the subset $X^{j+1}$ to be partitioned at the stage $j + 1$, it will be the one defined as $X^{j+1} = X^j - \mathcal{C}_j$.

Finally, the natural clusters in *Active* verifying equations (2)–(3) are induced as:

$$C_n = \{\phi_i \in \mathit{Active} \mid \ i \in \mathcal{C}_{n-1}\} \,, \quad \text{with} \quad n = 1, 2, \cdots, N, \tag{5}$$

and where $N$ denotes the number of clusters into which $X = \{i \mid \phi_i \in \mathit{Active}\}$ was partitioned, that is $\mathcal{C}_0, \mathcal{C}_1, \cdots, \mathcal{C}_{N-1}$.

See Fig. 1 for further illustration of this analysis.

### *3.1 Dissimilarity function*

Let $X^j$ be a subset of data not absorbed in any of the existing clusters $\mathcal{C}_0, \mathcal{C}_1, \cdots, \mathcal{C}_{j-1}$, at the stage $j$ of the dynamic processing; with $X^0$ being the given data set, $X^0 = X = \{i \mid \phi_i \in Active\}$. Next we define a graph $GRAPH^j = (X^j, U^j)$ corresponding to the data subset $X^j$, and with $U^j$ being the set of arcs $u = (k, l)$ between pairs of points in $X^j$. We associate with each arc $u \in U^j$ a real number $l(u) \geq 0$, and if $u = (k, l)$, we shall also use the notation $l_{kl}$ for $l(u)$. Let $l_{kl}$ be the distance from $k$ to $l$ defined as:

$$l_{kl} = Distance(\phi_k, \phi_l) \tag{6}$$

where $Distance(\phi_k, \phi_l)$ measures the distance between the filtered response of the two filters $\phi_k$ and $\phi_l$ as given in Appendix B (see equation (19)).

The cost of a path is defined as the greatest distance between two successive vertices on the path. Let $\mu(seed_j, k)$ be a set of arcs constituting a path between two points $seed_j$ and $k$ in $X^j$. And let $l(\mu)$ represent the cost of $\mu(seed_j, k)$ from $seed_j$ to $k$ defined as follows:

$$l(\mu) = \max\{l(u) \mid u \in \mu(seed_j, k)\} \tag{7}$$

Taking into account that two filters belong to the same cluster if there exists continuity (i.e., there exists similarity in their statistics at the same spatial locations) across the responses of filters in a path between them, the dissimilarity function is next defined as the cost of the optimum path from a seed point to each other on the graph. The optimum path between two data points $seed_j$ and $k$ is the path $\mu^*(seed_j, k)$ from $seed_j$ to $k$ whose maximum cost $l(\mu^*)$ is minimum:

$$\mu^*(seed_j, k) = \mathrm{Arg}_\mu \min\left[\max\{l(u) \mid u \in \mu(seed_j, k)\}\right] \tag{8}$$

Hence, the dissimilarity from the viewpoint of $seed_j$ to each $k$, is defined as the cost of the optimum path $\mu^*(seed_j, k)$ from $seed_j$ to each $k$:

$$d^{GRAPH^j}(seed_j, k) = l(\mu^*) = \max\{l(u) \mid u \in \mu^*(seed_j, k)\}, \tag{9}$$

with $\mu^*$ being the optimum path between $seed_j$ and $k$. The optimal path algorithm is given in [7].

### 3.2 *Clarity of separation at stage j*

Here we introduce the criterion by which we define the natural cluster $\mathcal{C}_j$ at stage $j$.

The set $\{d^{GRAPH^j}(seed_j, k) \mid$ with $k \in X^j\}$ is firstly ordered to obtain a new function:

$$d_j(i) = d^{GRAPH^j}(seed_j, k_i) \ , \quad \text{such that} \quad d_j(i) \leq d_j(i+1) \tag{10}$$

where $d_j(i)$ denotes the cost of the optimum path from $seed_j$ to $k_i$.

Let $\varepsilon_j$ represent the degree of closeness that is required between a pair of points that belong to the natural cluster of $seed_j$, noted as $\mathcal{C}_j$. Taking into account that $d_j(i)$ measures the closeness between $seed_j$ and $k_i$, with $k_i \in X^j$, we have that $\varepsilon_j$ can be defined as:

$$\varepsilon_j = d_j(i^*) \tag{11}$$

with $i^*$ being the location of the first significant rise in the value of $d_j(i)$ when $i$ increases. The value of $i^*$ is computed as the first zero crossing of the second derivative of $d_j$, as described in Appendix C.

A point $k_i$ from $X^j$ is then assigned the same cluster of $seed_j$ if the closeness between

$seed_j$ and $k_i$ is less than or equal to $\varepsilon_j$:

$$k_i \in \mathcal{C}_j \quad , \quad \text{if} \quad d_j(i) \leq \varepsilon_j \quad ; \quad \text{otherwise} \quad k_i \notin \mathcal{C}_j .$$

## 4 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 $C_n = \{\phi_{n_j}\}$:

$$t_n = |\sum_j A_{n_j}| \tag{12}$$

where $A_{n_j}$ denotes the original image $t(x, y)$ filtered through the logGabor $\phi_{n_j}$ in $C_n$ and passed through a non-linearity of the form:

$$\tanh(z, \tau) = \frac{1 - \exp\{-z\tau\}}{1 + \exp\{-z\tau\}} \tag{13}$$

where $\tau$ is a gain term.[8] 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.

Figs. 2, and 3, show 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 these figures.

## 5 INTEGRATION OF FEATURES

Much of object perception requires the integration of image information for features that are separated in space. Therefore a model capable of working with real-world images should perform tasks such as long-range figure completion, or phase-induced boundary detection.

The original pixelwise image description was here used to infer (through detectors automatically learned on the raw data and designed to extract information which is localized both spatially and in the frequency domain) the existence and location of certain classes of objects. These objects are defined as features which are statistically invariant across scale space or have the highest degree of alignment in statistical structure across different frequency bands.

This path leads directly to the uncertainty principle[9]: "Two-dimensional linear spatial filters are constrained by general uncertainty relations, which limit their attainable joint resolution in the 2D space domain and the 2D frequency domain." From these uncertainty relations, governing the employment of more than one coordinate system (2D frequency domain-2D spatial domain) in image processing, will come a locally uncertain class membership (e.g., which pixels are linear features and which are not) in the search for precise positional information.

To illustrate these uncertainty effects, we take the input image in Fig. 4 and perform filtering with four strongly responding logGabors. These four filters are activated units which are within the same natural cluster in *Active*. They are illustrated on the middle column by ellipses drawn, in the 2D spatial-frequency plane, at the point where their amplitude has decreased to the ($e^{-1/2}$) half width its maximum. The results of the spatial bandpass processing are shown on the right column. As can be seen from Fig. 4, the increase of resolution in the 2D frequency domain is made at the expense of loss of certainty of class membership.

The interesting point is that the inevitable loss of certainty of class membership leads to long-range boundary completion (and thus illusory contours) as well as other illusory percepts. In Fig. 4, as shown at the bottom, the pooling of filtered responses induces a sum of uncertainty effects that accounts for the long-range boundary completion.

Fig. 5 illustrates the result of this analysis on a Varin's line-end counterpart[10] to the Kanizsa square (as shown in Fig. 5.A). The Varin square converts standard edge type inducers of a Kanizsa square to a number of line-end inducers by replacing them with partial concentric circles. A natural cluster $C_n$ of activated logGabors derived by the clustering in *Active* is shown in Fig. 5.B. The filtered responses of the logGabors within cluster $C_n$, were summed to give a local estimate of the energy in the corresponding 2D frequency domain-2D space domain window, as shown in Fig. 5.C . Next, the position of the features shown in Fig. 5.D was estimated by the Shen-Castan edge detector[11] on Fig. 5.C . The long-range boundary completion is a consequence of the precision of the positional information (slow transitions appear identical to sharp edges), in accordance with the uncertainty principle.[12]

Fig. 6.A illustrates the Müller-Lyer illusion. This is the most famous size illusion. The center connecting line is seen as shorter in the top shape than in the lower shape. The two center lines are actually the same size. The proposed representational model produces an output, shown in Fig. 6.C, that contains a center fragment in the top shape that is shorter than that obtained in the lower shape (the corresponding cluster of filters is illustrated in Fig. 6.B). In fact, Fig. 6.D was obtained by applying the Shen-Castan detector on Fig. 6.C.

## 6 PREDICTING VISUAL TARGET DISTINCTNESS

Rohaly[13] 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 two experiments reported in this section are 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 6.2). Then, a computational metric is applied to quantify the visual distinctness of the targets in each one of two different datasets (Section 6.3). Finally, a relation is established between the computational and psychophysical target distinctness estimates (Section 6.4).

### *6.1 Images*

The images used in this study are slides made during the DISSTAF (Distributed Interactive Simulation, Search and Target Acquisition Fidelity) field test, that was designed and organized by NVESD (Night Vision & Electro-optic Sensors Directorate, Ft. Belvoir, VA, USA) and that was held in May and June 1995 in Fort Hunter Liggett, California, USA. These slides depict 44 different scenes. Each scene represents a military vehicle in a complex rural background. The 9 different vehicles that are deployed as search targets are respectively a BMP-1, a BTR-70, an HMMVV-Scout, a HMMVV-Tow, an M1A1, an M3-Bradley, an M60 , an M113, and a T72. 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 images used in the computational experiments are subsampled to $256 \times 256$ 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. This replacement is done by hand, using the rubber stamp tool in Photoshop 3.05. The result is judged by eye and is accepted if the variation in the background over the target support area does not appear to have an appreciable contrast with the natural variation in the local background.

Two different datasets, from the 44 slides made during the DISSTAF field test, were used in the visual target distinctness experiments:

(i) In the first experiment, the digital images were (see Fig. 7): (i) ten complex natural images containing a single target, that correspond to the scenes 16, 9, 37, 6, 30, 26, 29, 3, 21, and 11, from the 44 slides made during the DISSTAF field test, and (ii) ten empty images of the same rural backgrounds with no target, that were created using the rubber stamp tool in Photoshop 3.05.

(ii) The digital images used in the second experiment were (see Fig. 8): (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.

### *6.2 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 [14].

Search performance is usually expressed as the cumulative detection probability as function of time, and it can be approximated by:[15--18]

$$P_d(t) = \begin{cases} 0 & t < t_0 \\ 1 - \exp\left\{-\frac{t-t_0}{\rho}\right\} \ , & t \geq t_0 \end{cases} \tag{14}$$

where

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

Fig. 9 shows the cumulative distribution functions corresponding to the search times measured for the target scenes used in the two experiments 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, and 2, respectively. 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. 9 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 experiment 1 is clustered into $\{16, 9, 37\}$, $\{6, 30, 26, 29\}$, and $\{3, 21, 11\}$,
- the dataset in experiment 2 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.

### *6.3 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 (12). 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 the sum of all filtered responses in *Active*, plus a saturation constant $\sigma$:

$$\frac{t_n}{\sigma^2 + \sum_i |A_i|} \tag{15}$$

where $A_i$ denotes the original image $t(x, y)$ filtered through the logGabor $\phi_i$ in *Active* and passed through a non-linearity as given in equation (13).

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 $\beta$-norm between their statistical structure over those pixels which form "fixation points" on $t_n$:[19,20]

$$d_{VP}(t_n, e_n) = \frac{1}{Card[FP(t_n)]} \left( \sum_{(x,y) \in FP(t_n)} \left| D[T^{t_n}(x,y), T^{e_n}(x,y)] \right|^{\beta} \right)^{\frac{1}{\beta}} \tag{16}$$

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 $\beta$ in Equation (16) 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:[21]

$$VP(t, e) = \text{Decision-Rule}\{D_1, D_2, \cdots, D_N\} \tag{17}$$

A schematic overview of the VP distinctness metric is given in Fig. 10.

### 6.4 Results and discussion

#### 6.4.1 Experiment 1

The digital images used in this first experiment are ten target images and the corresponding empty scenes as shown in Fig. 7.

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 root mean square error $RMSE$ metric are listed in column 3. The $RMSE$ performs poorly, which is to be expected. Significant rank order permutations are displayed in boxes. The $RMSE$ metric produces a rank order with five significant order reversals: targets 16, 26, 29, 3, and 11 are significantly out of order relative to the reference order induced by the psychophysical distinctness measure in column 2. The other targets have been attributed rank orders which do not differ significantly from the reference rank order. The $RMSE$ yields a relatively low probability ($P_{CC} = 0.5$). These results show that the $RMSE$ metric appears not capable to rank order targets in Experiment 1 with respect to their visual distinctness.

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 29, 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. Among the computational metrics here tested, the $VP$ metric yields the highest probability ($P_{CC} = 0.8$). Summarizing, for the dataset in this first experiment, the $VP$ metric appears to compute a visual target distinctness rank ordering that correlates with human observer performance.

### *6.4.2 Experiment 2*

Fig. 8 show the digital images used in this second experiment: eleven images of a military vehicle in a rural background and the corresponding empty scenes.

The comparative results of the computational metric and those of both quantitative and qualitative measures are presented in Table 2. At the bottom of each of the columns is also shown the probability of correct classification of the metric in that column.

The target distinctness values and the resulting rank order computed by the $RMSE$ metric are listed in column 3. Again the $RMSE$ 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 ($P_{CC} = 0.27$).

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 2. 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 Experiment 2 with respect to their visual distinctness.

## 7 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. Results were provided to enlighten the connection between illusory percepts and regularities across frequency bands. The induced representational model was also 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. 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.

### ***Acknowledgement***

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 on 44 slides made during the DISSTAF field test.

This work was sponsored by the Spanish Board of Science and Technology (CICYT) under grant TIC97-1150.

## References

[1] Treisman, A.M., and Gelade, G. "A feature-integration theory of attention." *Cognitive Psychology*, vol. 12, pp. 97-136, (1980).

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

[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] Morrone, M.C., and Owens, R.A. "Feature detection from local energy." *Pattern Recognition Letters*, Vol. 6, pp. 303-313, (1987).

[6] 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).

[7] Garcia, J.A., Fdez-Valdivia, J., Cortijo, F.J., Molina, R. "A dynamic approach for clustering data." *Signal Processing*, Vol. 44, pp. 181-196, (1995).

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

[9] Daugman, John G. "Uncertainty relation for resolution in space, spatial frequency, and orientation optimized by two-dimensional visual cortical filters." *J. Opt. Soc. Am. A*, Vol. 2(7), pp. 1160-1169, (1985).

[10] Varin, D. "Fenomeni di contrasto e diffusione cromatica nell organizzazione spaziale del campo percettivo." *Rivista di Psicologia*, Vol. 65, pp. 101-128, (1971).

[11] Shen, J., and Castan, S. "An optimal linear operator for step edge detection." *CVGIP: Image Understanding*, Vol. 54(2), pp. 112-133, (1992).

[12] Wilson, Roland and Granlund, Goesta H. "The uncertainty principle in image processing." *IEEE Trans. on Pattern Analysis and Machine Intelligence*, Vol. 6(6), pp. 758-767, (1984).

[13] 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).

[14] 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).

[15] Krendel, E.S. and Wodinsky, J. "Visual search in an unstructured visual field." *Journal of the Optical Society of America*, Vol. 50, pp. 562-568, (1960).

[16] Rotman, S.R., Gordon,E.S., and Kowalczyk, M.L. "Modeling human search and target acquisition performance: I. First detection probability in a realisctic multitarget scenario." *Optical Engineering*, Vol. 28, pp. 1216-1222, (1989).

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

[18] Toet, A. "Computing visual target distinctness", TNO-report TM-97-A039, TNO Human Factors Research Institute, pp. 74, (1997).

[19] 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).

[20] Fdez-Vidal, X.R., Garcia, J.A., Fdez-Valdivia, J., and Rodriguez-Sanchez, R. "The role of integral features for perceiving image discriminability", *Pattern Recognition Letters*, Vol. 18, pp. 733-740, (1997).

[21] Fdez-Vidal, X.R., Toet, A., Garcia, J.A., Fdez-Valdivia, J. "Computing visual target distinctness through selective filtering, statistical features, and visual patterns." *DECSAI Technical Report*, DECSAI-980327. Also in $http://decsai.ugr.es/diata/tech\_rep.html$, electronic edition. (1998).

[22] Rodriguez-Sanchez, Rosa, Garcia, J.A., and Fdez-Valdivia, J. "The RGF pandemonium: a Low-Level Representational Model For Images." *Pattern Recognition*, Vol. 31(11), pp. 1797-1810, (1998).

[23] Quick, R.F. (1974). A vector-magnitude model of contrast detection. *Kybernetik*, 16, 65-67.

[24] Fdez-Valdivia, J., Garcia, J.A., and Garcia-Silvente, M. "An evaluation of the novel normalized-redundancy representation for planar curves." *International Journal of Pattern Recognition and Artificial Intelligence*, Vol. 10(7), 769-789, (1996).

## Appendix A: Activated units from the filter bank

Let an activated filter be a logGabor function $\phi_i$ from the filter bank selectively sensitive to patterns contained in the original reference image. These patterns produce clumps of energy upon the Fourier spectrum of the reference image. The clumps of energy can be captured into a collection of oriented spatial-frequency channels as described in [22]. Hence, the activated units from the bank can be obtained by identifying those filters whose amplitude spectrum and some clump of energy in the image amplitude spectrum overlap to some extent.

The filter $\phi_i$ in Equation (1) can be represented as a Gaussian in the spatial frequency domain around the central frequency $(r_i, \theta_i)$. Hence, the amplitude spectrum of a logGabor filter $\phi_i$ and some clump of energy in the image amplitude spectrum overlap to some extent if there exists a coordinate $(r^*, \theta^*)$ upon the oriented spatial-frequency channel capturing the clump of energy such that:

$$|r^* - r_i| \leq \sigma_{r_i}, \text{ and, } |\theta^* - \theta_i| \leq \sigma_{\theta_i}, \tag{18}$$

where $\theta_i$ is the orientation angle of the filter $\phi_i$, $r_i$ is its central radial frequency, and $\sigma_{\theta_i}$ and $\sigma_{r_i}$ are the angular and radial sigma of the Gaussian that determines $\phi_i$ in the 2D frequency domain.

**Appendix B: Distance between activated filters**

The image $r$ can be decomposed through the activated filters into the set of its most significant components $A_1, A_2, ..., A_m$, where $A_i$ represents the original image $r(x, y)$ filtered through a logGabor $\phi_i$ in *Active*.

Let $T^i(x, y) = (T^i_l(x, y))$ be a vector of statistics at $(x, y)$ on the local energy of $A_i$. The local energy of an image analyzed by a logGabor filter (hereafter, filter response) can be expressed as[3]:

$$\sqrt{O^2_{even}(x, y) + O^2_{odd}(x, y)},$$

where $O_{even}(x, y)$ is the image convolved with the even-symmetric logGabor filter and $O_{odd}(x, y)$ is the image convolved with the odd-symmetric logGabor filter at point $(x, y)$.

Let $Distance(\phi_i, \phi_j)$ be the distance between two components $A_i$ and $A_j$, computed via a $\beta$-norm between their statistical structure over those pixels which form "fixation points”:[23]

$$Distance(\phi_i, \phi_j) = Dist[i,j]^2 + Dist[j,i]^2 \tag{19}$$

where

$$Dist[p,q] = \frac{1}{Card[FP(p)]}\left(\sum_{(x,y)\in FP(p)} |D[T^p(x,y), T^q(x,y)]|^\beta\right)^{\frac{1}{\beta}} \quad , \tag{20}$$

with $p, q \in \{i, j\}$; where $FP(p)$ denotes the set of fixation points for the activated filter $\phi_p$; and with $D[T^p(x,y), T^q(x,y)]$ being a normalized distance measure between $T^p(x,y)$ and $T^q(x,y)$ as described in [20].

The default value of the exponent $\beta$ in equation (20) is 3. [4] discussed at some length several interpretations of the pooling formula and the selection of the exponent.

## Appendix C

Let $d_j''$ be the second derivative of $d_j$ computed as:

$$d_j''(i) = d_j(i) * \frac{d^2}{di^2}G_s(i),$$

with

$$G_s(i) = \frac{1}{\sqrt{2}s}\exp\{-\frac{i^2}{2s^2}\},$$

and where $d_j$ is convolved with the second derivative of the Gaussian at scale $s$, noted as $\frac{d^2}{di^2}G_s(i)$, to both smooth and differentiate the function. The zero crossings of $d_j''$ correspond to positions at which the dissimilarity $d_j$ undergoes a significant increment in its value. To locate the zero crossings marking a rise in $d_j$ due to inter-cluster differences, the unwanted detail from intra-cluster differences must be removed by smoothing. The question is: how much smoothing should be performed? The derivative should be processed at the scale that best describes the increments in $d_j$ due to inter-cluster differences, while removing spurious increments due to intra-cluster differences. [24] provides a suitable representation by extracting the most significant scales that best describe a noise signal. When $d_j''$ is computed at the smallest significant scale to describe the basic structure of $d_j$, the zero crossings from the inter-cluster differences can be extracted without confusion.

Table 1.- On column 1: the dataset in the experiment 1; on column 2: the reference rank order for the evaluation of the computational distinctness metrics; on columns 3 and 4: the target distinctness values and the resulting rank order computed by the $RMSE$, and $VP$ metrics, respectively. At the bottom of each column is shown the probability of correct classification of the metric rank order in that column with respect to the reference rank order in column 2.

Table 2.- On column 1: the dataset in the experiment 2; 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.

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

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.

Figure 3.- Decomposition into visual patterns performed on several images of a military vehicle (target) in a complex rural background.

Figure 4.- Uncertainty effects. Left column: input image. Middle column: activated log-Gabors. Right column: results of the spatial bandpass processing.

Figure 5.- (A) Varin square. (B) A natural cluster $C_n$ of activated logGabors derived by the clustering in $Active$. (C) Sum of the filtered responses over filters in $C_n$. (D) Position of the linear features estimated by the Shen-Castan edge detector.

Figure 6.- (A) The Müller-Lyer illusion. (B) Natural cluster $C_n$ of activated filters. (C) The filtered responses of the logGabors within cluster $C_n$, were summed to give a local estimate of the energy in the corresponding 2D frequency domain-2D space domain window. (D) Position of the linear features estimated by the Shen-Castan edge detector.

Figure 7.- The dataset used in the first experiment:(i) ten complex natural images containing a single target, that correspond to the scenes 16, 9, 37, 6, 30, 26, 29, 3, 21, and 11, from the 44 slides made during the DISSTAF field test, and (ii) ten

empty images of the same rural backgrounds with no target, that were created using the rubber stamp tool in Photoshop 3.05.

Figure 8.- The images used in the second experiment: (i) eleven target images that correspond to the scenes 16, 34, 28, 17, 26, 19, 8, 15, 3, 22, 4, and (ii) the corresponding eleven empty scenes.

Figure 9.- The target images in each of the two experiments are clustered into a number of sets of targets with comparable visual distinctness (in accordance with a Kolmogorov-Smirnov test). At the top, the dataset in experiment 1 is clustered into $\{16, 9, 37\}$, $\{6, 30, 26, 29\}$, and $\{3, 21, 11\}$. At the bottom, the dataset in experiment 2 is clustered into $\{16, 34\}$, $\{28, 17, 26\}$, $\{19, 8, 15, 3\}$, and $\{22, 4\}$.

Figure 10.- A schematic overview of the $VP$ metric.

| DATASET # 1 | | | | | | |
|---|---|---|---|---|---|---|
| *Image Pair* | $R_{P_d}$ | | *RMSE* | | *VP* | |
| | Value | Rank | Value | Rank | Value | Rank |
| **# 16** | 96.39 | 1 | 2.55 | $\boxed{4}$ | 7.718 | 3 |
| **# 9** | 96.27 | 2 | 16.90 | 1 | 14.893 | 1 |
| **# 37** | 96.19 | 3 | 16.12 | 2 | 11.519 | 2 |
| **# 6** | 90.95 | 4 | 1.94 | 7 | 5.386 | 4 |
| **# 30** | 90.66 | 5 | 2.37 | 5 | 5.281 | 5 |
| **# 26** | 90.06 | 6 | 1.71 | $\boxed{8}$ | 4.53 | 7 |
| **# 29** | 89.49 | 7 | 1.60 | $\boxed{9}$ | 3.078 | $\boxed{10}$ |
| **# 3** | 80.02 | 8 | 2.83 | $\boxed{3}$ | 5.122 | $\boxed{6}$ |
| **# 21** | 73.84 | 9 | 1.35 | 10 | 4.274 | 9 |
| **# 11** | 73.40 | 10 | 1.96 | $\boxed{6}$ | 4.311 | 8 |
| $P_{cc}$ | - | | **0.5** | | **0.8** | |

Table 1

| DATASET # 2 | | | | | | |
|---|---|---|---|---|---|---|
| *Image Pair* | $R_{P_d}$ | | *RMSE* | | *VP* | |
| | Value | Rank | Value | Rank | Value | Rank |
| **# 16** | 96.39 | 1 | 2.55 | $\boxed{7}$ | 5.15 | 2 |
| **# 34** | 96.37 | 2 | 30.48 | 1 | 32.99 | 1 |
| **# 28** | 93.48 | 3 | 3.86 | $\boxed{2}$ | 3.19 | 5 |
| **# 17** | 92.42 | 4 | 3.68 | 3 | 3.71 | 4 |
| **# 26** | 90.95 | 5 | 1.94 | $\boxed{8}$ | 2.20 | $\boxed{9}$ |
| **# 19** | 84.79 | 6 | 3.30 | $\boxed{4}$ | 2.98 | 7 |
| **# 8** | 84.47 | 7 | 1.57 | 9 | 2.52 | 8 |
| **# 15** | 80.46 | 8 | 1.17 | $\boxed{10}$ | 3.15 | 6 |
| **# 3** | 80.02 | 9 | 2.83 | $\boxed{5}$ | 4.16 | $\boxed{3}$ |
| **# 22** | 60.96 | 10 | 1.02 | 11 | 2.14 | 10 |
| **# 4** | 53.89 | 11 | 2.76 | $\boxed{6}$ | 0.77 | 11 |
| $P_{cc}$ | - | | **0.27** | | **0.82** | |

Table 2

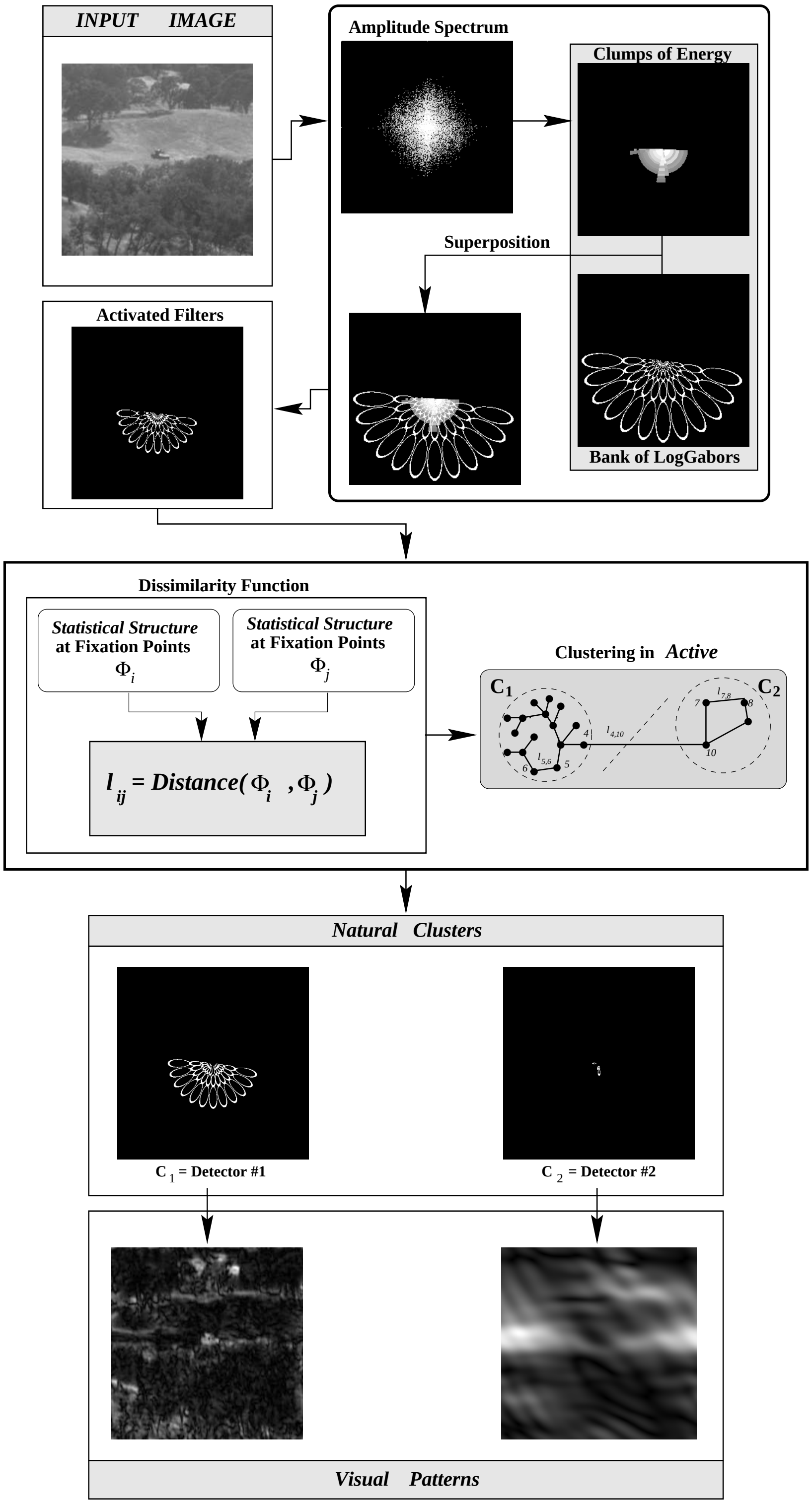

INPUT IMAGE
Amplitude Spectrum
Clumps of Energy
Superposition
Activated Filters
Bank of LogGabors
Dissimilarity Function
Statistical Structure
at Fixation Points
$\Phi_i$
Statistical Structure
at Fixation Points
$\Phi_j$
$l_{ij}$ = Distance($\Phi_i$ , $\Phi_j$ )
Clustering in Active
$C_1$
$C_2$
Natural Clusters
$C_1$ = Detector #1
$C_2$ = Detector #2
Visual Patterns

Figure 1

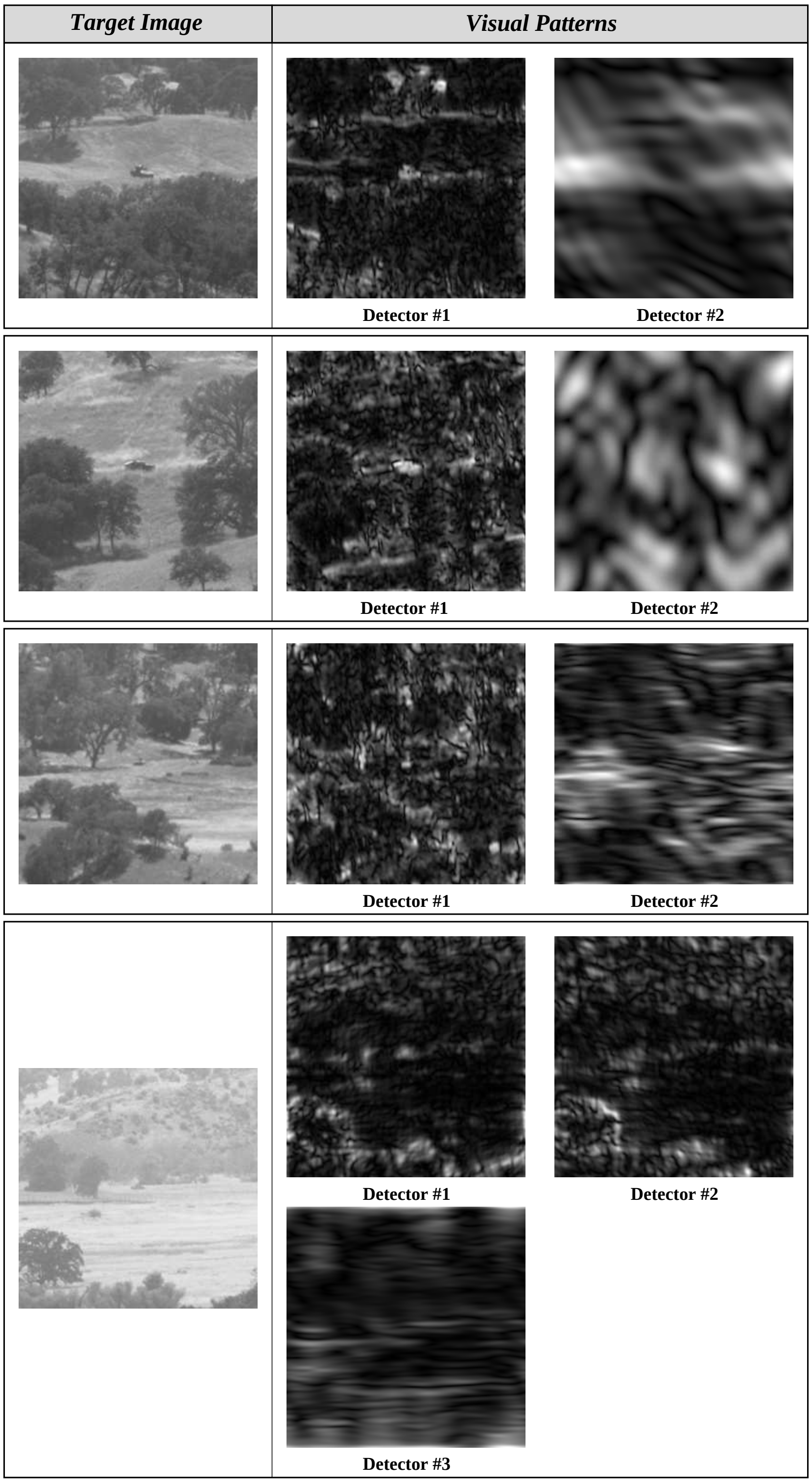

Target Image
Visual Patterns
Detector #1
Detector #2
Detector #1
Detector #2
Detector #1
Detector #2
Detector #1
Detector #2
Detector #3

Figure 2

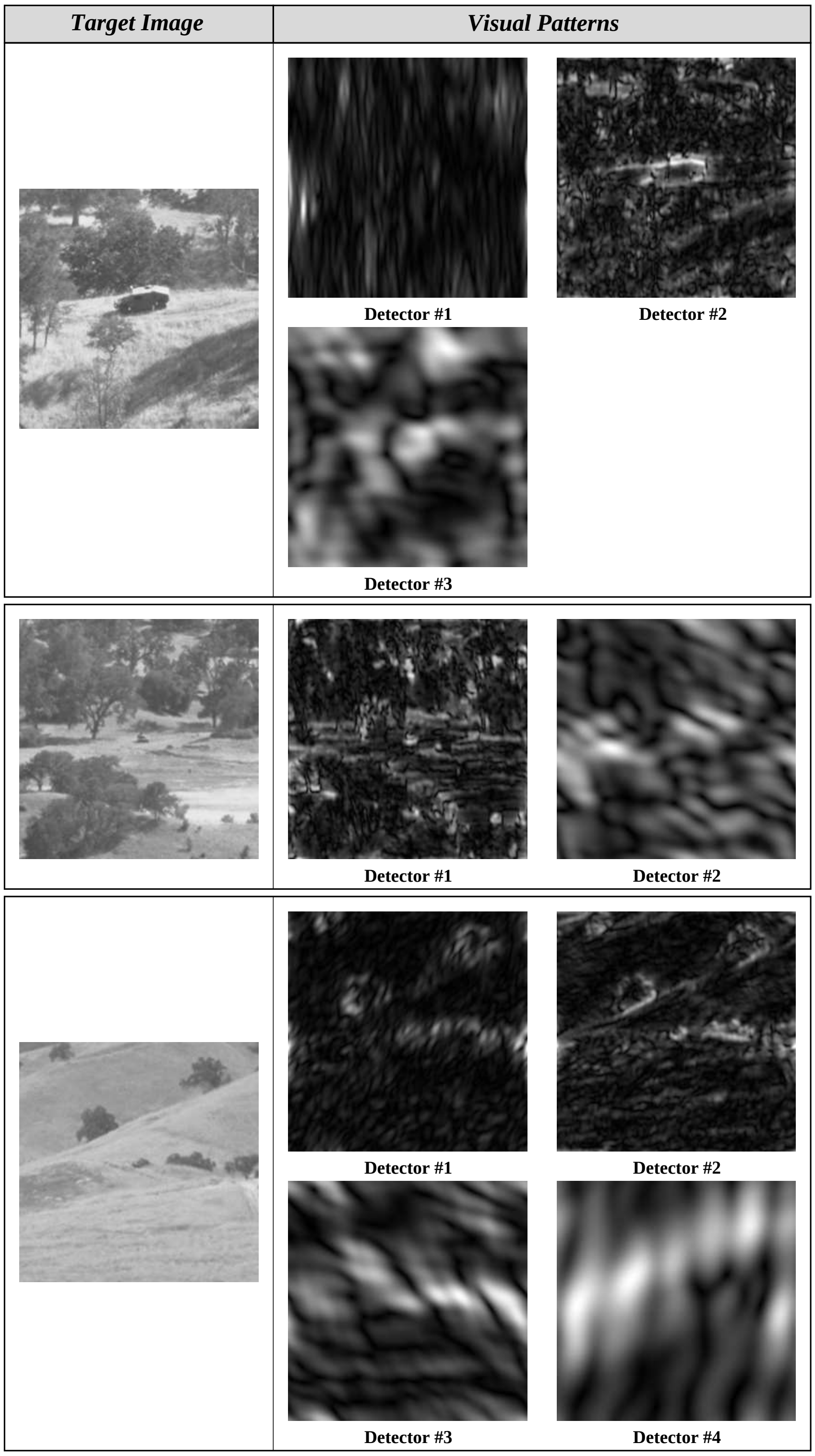

Target Image
Visual Patterns
Detector #1
Detector #2
Detector #3
Detector #1
Detector #2
Detector #1
Detector #2
Detector #3
Detector #4

Figure 3

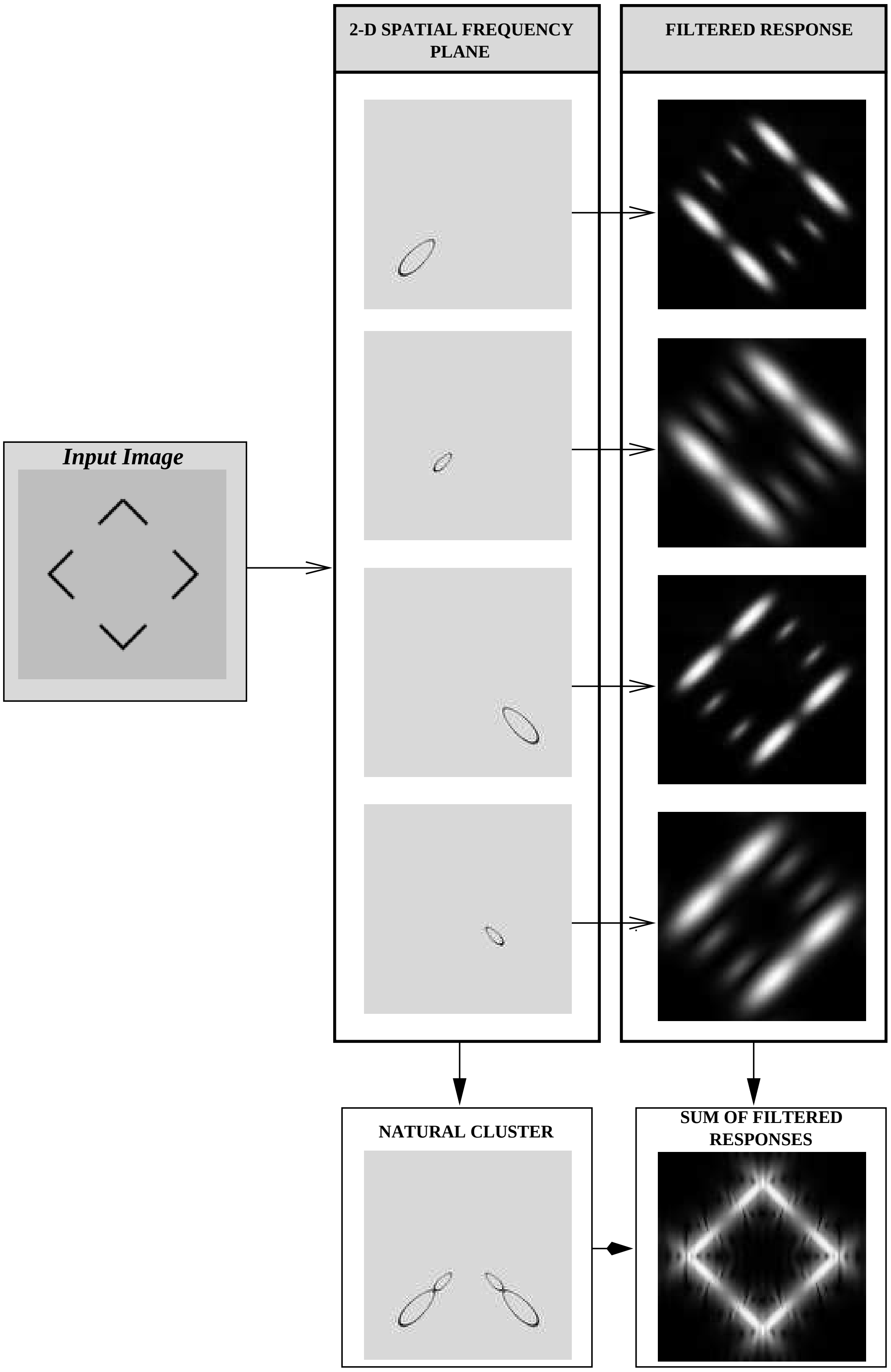

2-D SPATIAL FREQUENCY PLANE
FILTERED RESPONSE
Input Image
NATURAL CLUSTER
SUM OF FILTERED RESPONSES

Figure 4

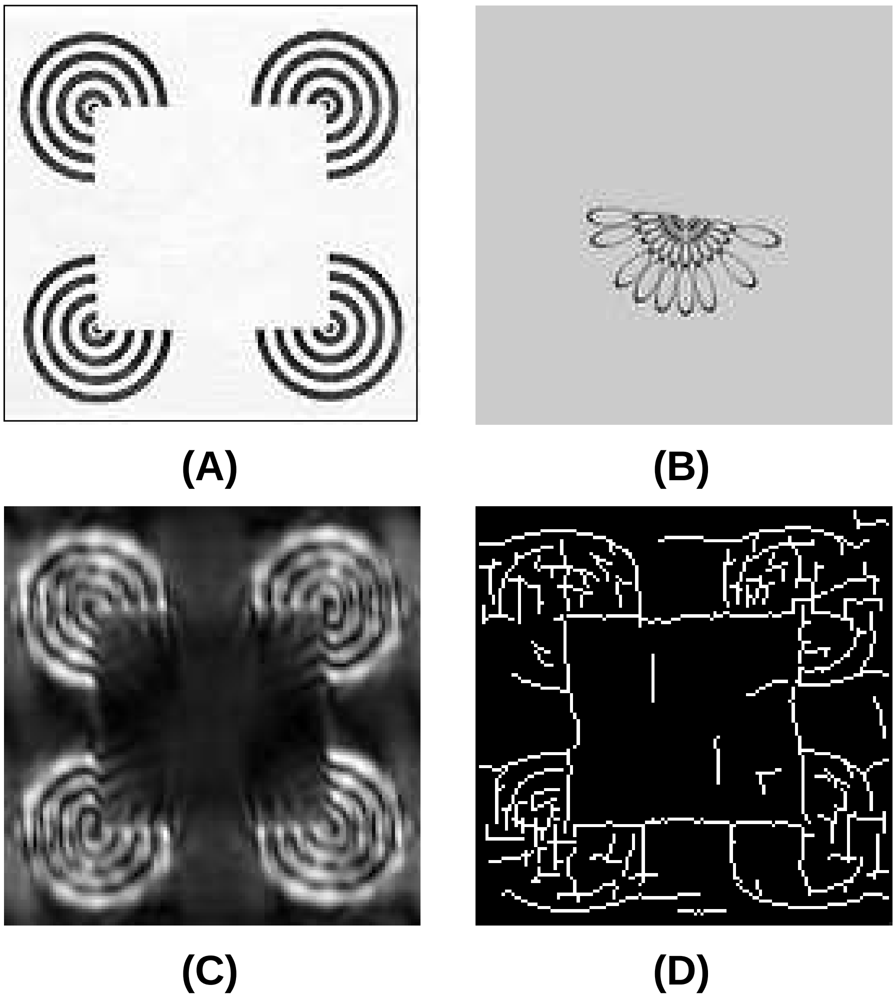



Figure 5

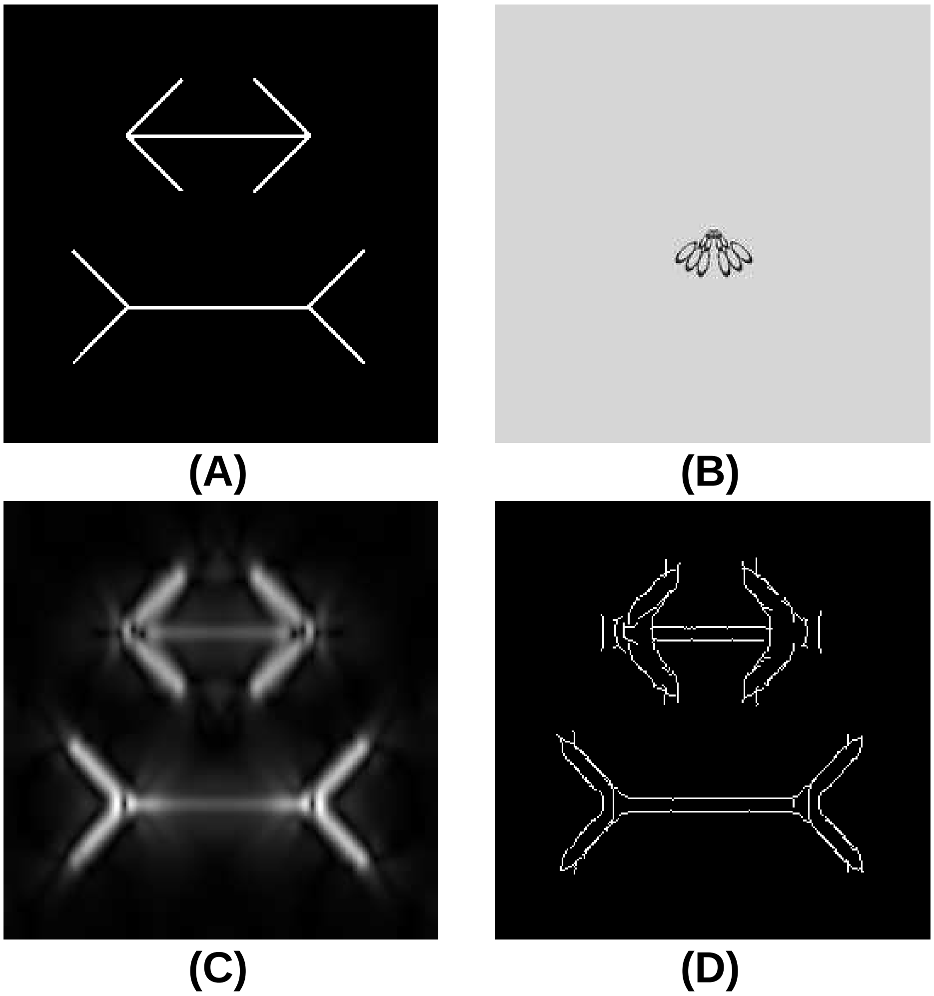


Figure 6

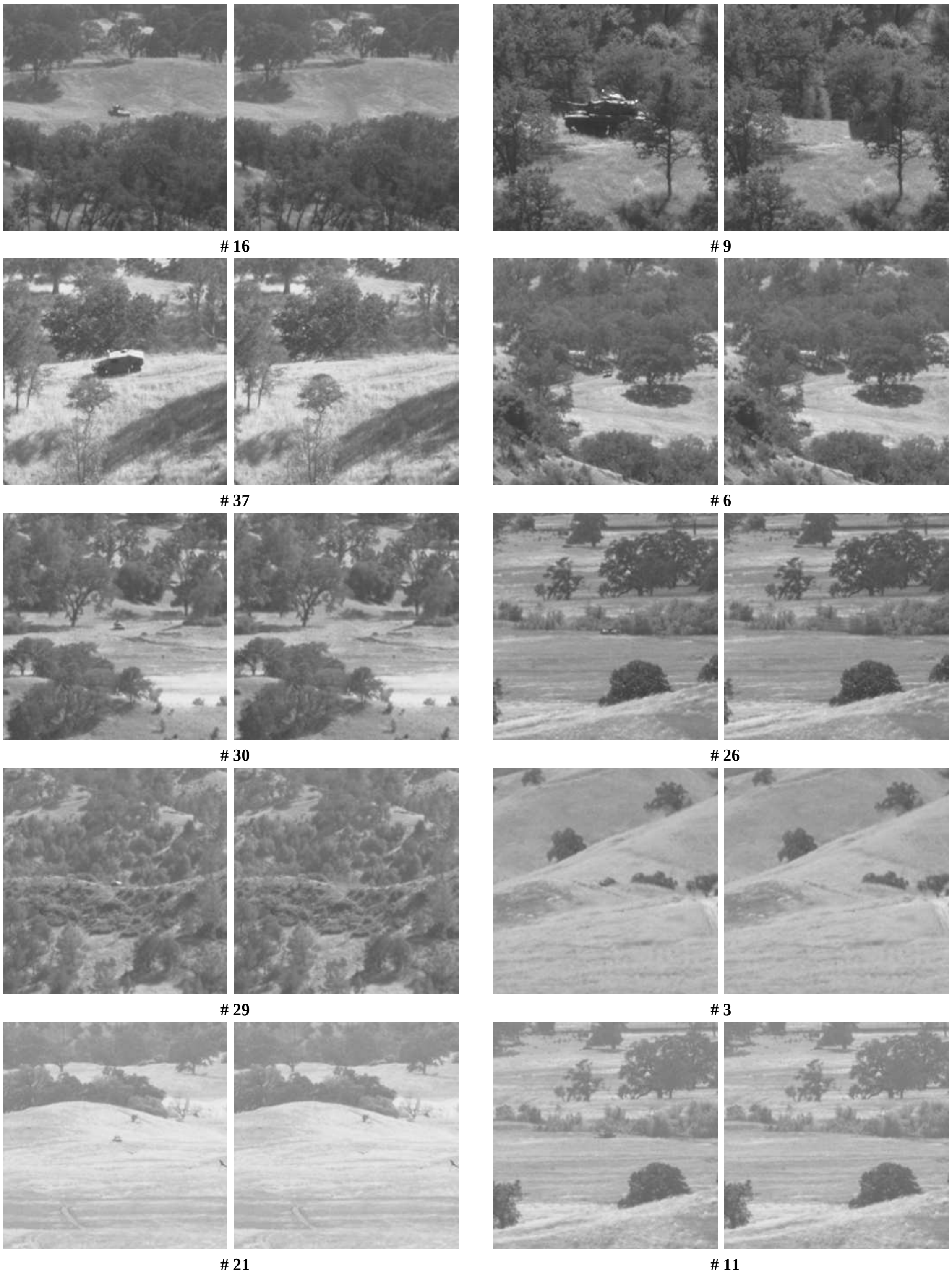



Figure 7

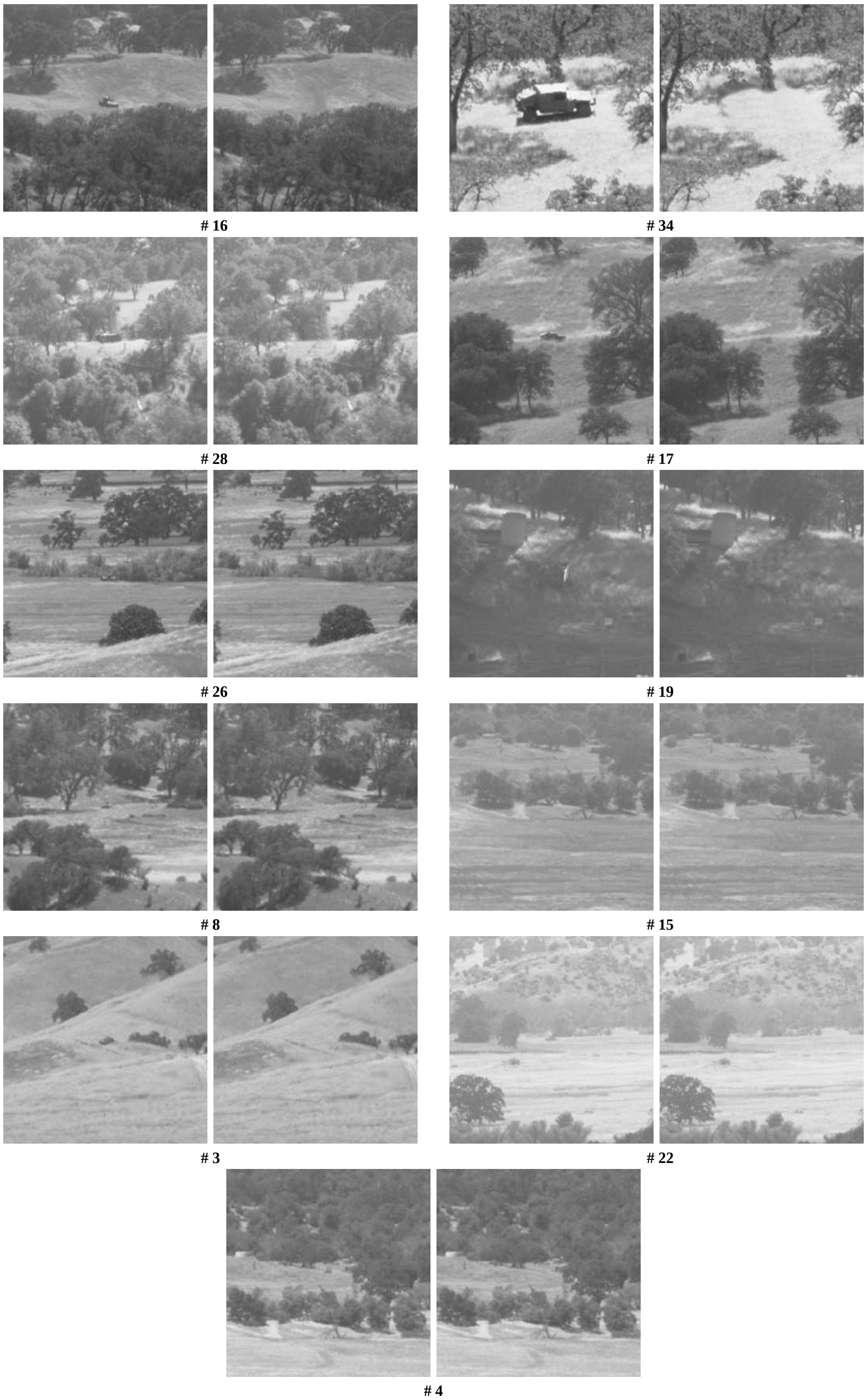
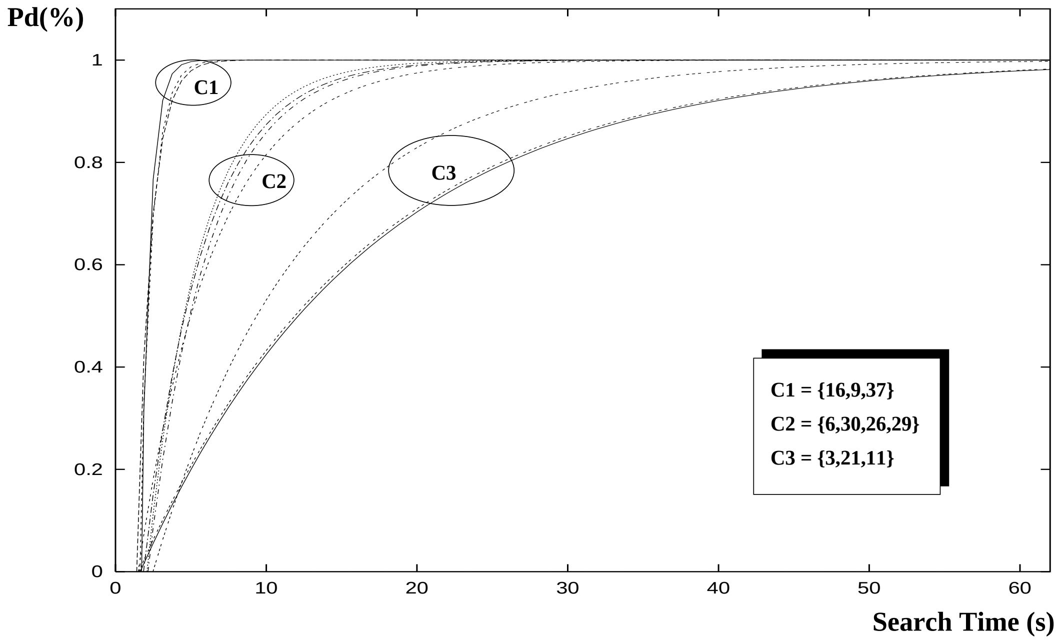
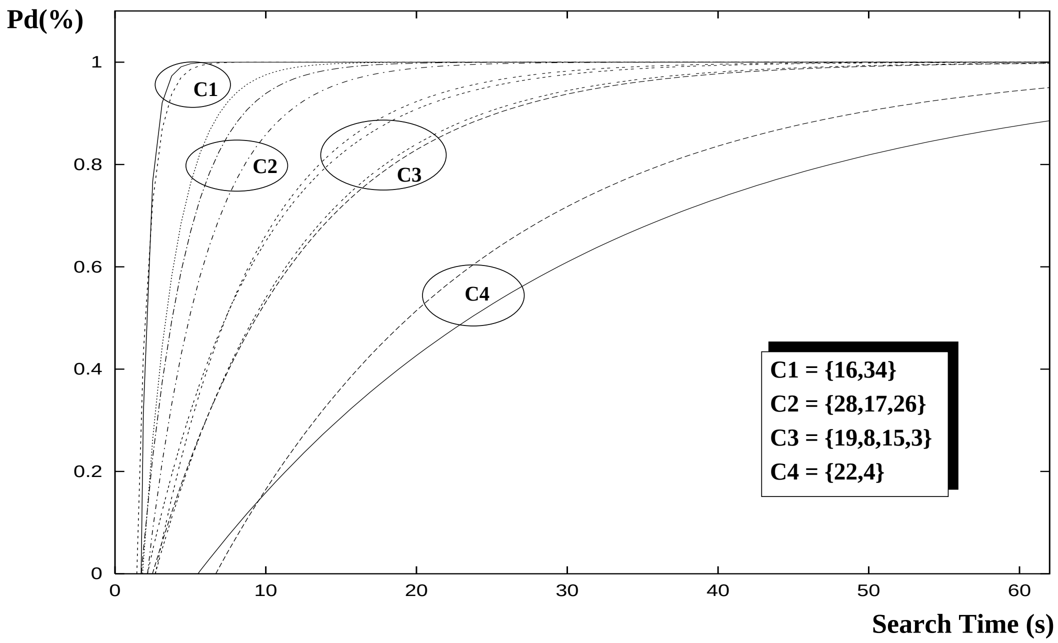



Figure 8




**DATASET # 1**




**DATASET # 2**

Figure 9

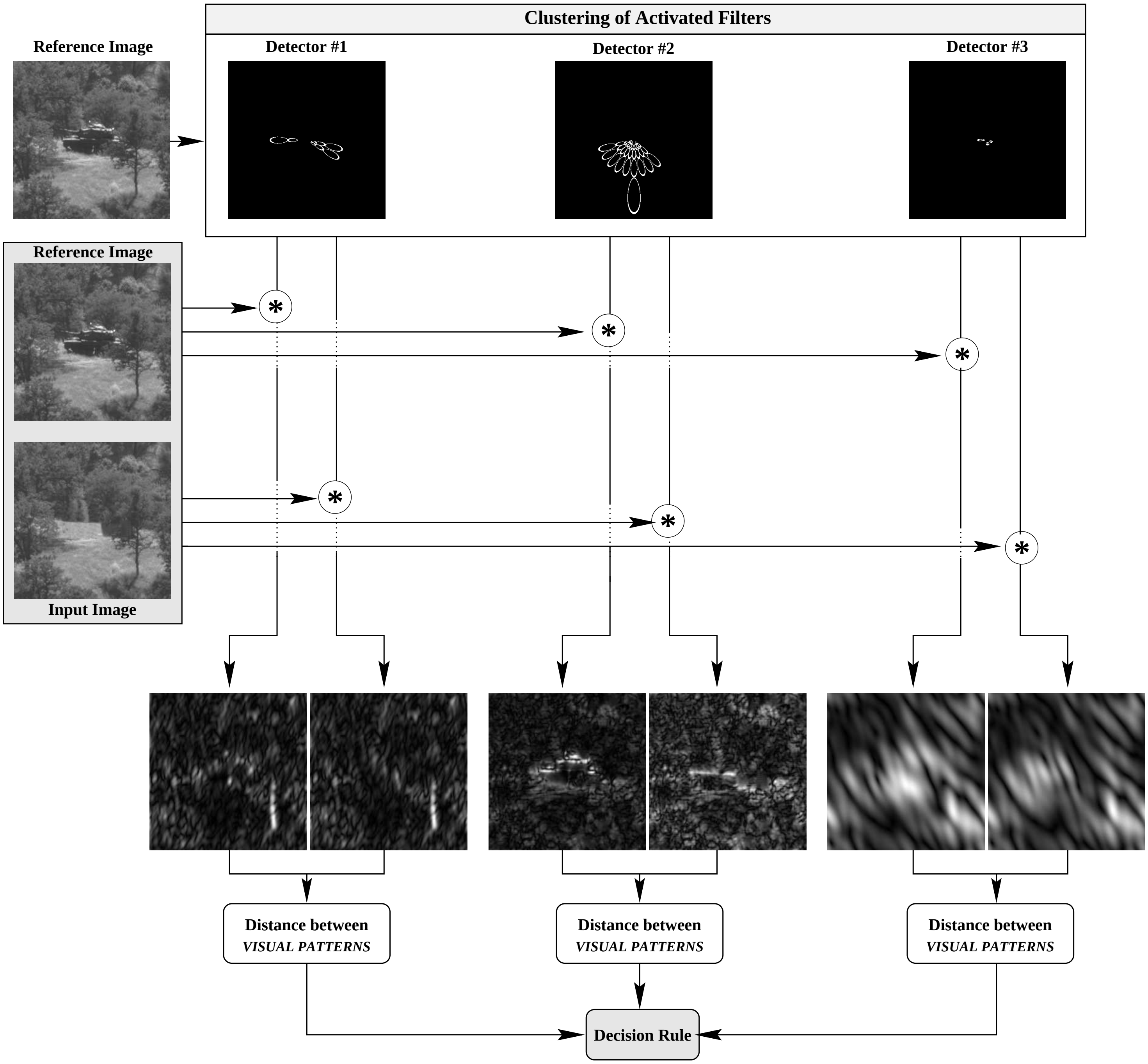

Clustering of Activated Filters
Reference Image
Detector #1
Detector #2
Detector #3
Reference Image
Input Image
Distance between
VISUAL PATTERNS
Distance between
VISUAL PATTERNS
Distance between
VISUAL PATTERNS
Decision Rule

Figure 10