MEASURES OF SIMILARITY BETWEEN TWO IMAGES

Charles C. Taylor
Department of Statistics
University of Strathclyde
Glasgow, U.K.

SUMMARY

Psychology and physiology give us some insight into the way in which humans derive information from images. We give a brief introduction to some of these theories in a general setting and consider the application to automatic procedures. Distance measures are discussed, with examples, to illustrate some of the difficulties.

1. Introduction

Suppose that the true scene t is corrupted by noise to a distorted image f, and that we have a restoration algorithm a which produces a restored image r.

$$a : f \mapsto r$$

We would like to measure the distance from the true image to the restored image, $d(t, r)$, possibly taking into account the starting point $d(t, r \mid f)$, although it is not clear how to use this information at present. Baddeley (1987) has distinguished 3 roles for error measures:

(i) a theoretical framework for deriving a new reconstruction algorithm.

(ii) a benchmark for comparing different reconstruction algorithms in a computer experiment.

(iii) a measure of achieved quality without reference to the true image.

A fourth reason for considering such a distance measure could be to give further insight into the performance of an algorithm, perhaps as a preliminary to (i).

A commonly used error measure involves a pixel-by-pixel comparison,

$$\frac{1}{N}\sum_{i=1}^{N}|t_i - r_i|$$

or

$$\left(\frac{1}{N}\sum_{i=1}^{N}(t_i - r_i)^2\right)^{1/2}$$

where N is the number of pixels and t_i is the class (or grey level) of pixel i in the true scene. Although widely used, these measures are also widely recognized as unsatisfactory; they ignore the spatial context and are inadequate in expressing human perceptions of similarity. Our objective in this paper is to derive a measure which in some way reflects the human assessment of errors, and comparison of images. We begin by describing some psychometric models, and attempt to use these principles in deriving appropriate measures.

2. Psychometric models

In the literature there are many models to describe the way in which humans perceive, and process information from, images. So far, no one seems to have proposed a satisfactory answer; without exception, each model has some flaw or limitation. We give a brief introduction to some of these models which may be relevant to our objective.

2.1 Template models

This proposes that the human mind retains a library of specific images. It then recognizes an image by choosing the "closest" match to one of those in the memory. For example, two such images in store may be A and R (see Figure 1). When presented with an alternative image (Figure 1) this mind would then decide this was an R, based on the fact that this is the closest. Clearly, this approach may be satisfactory for reading a typewritten post code, but it is not flexible enough to reflect human perception.

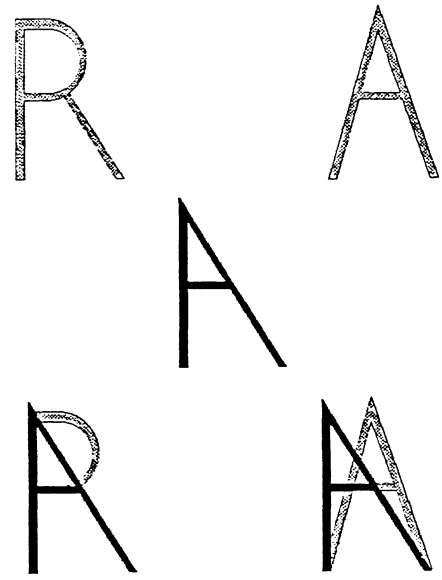


Figure 1. Template matching.

2.2 Prototype models

This relaxes the rigidity of the template model. It involves extraction and abstraction of information. The description of the model itself seems a bit abstract, but there is some experimental support. Psychology experiments show that the greater the number of "transformations" from the prototype, the less confident the subjects were in being sure they had seen the image before. The model thus assumes that stimuli are labeled as a prototype with a distortion; a search process then locates the "best fit". This idea is further developed in the next section, though without some further specification it would be hard to implement in an automatic procedure.

2.3 Feature models

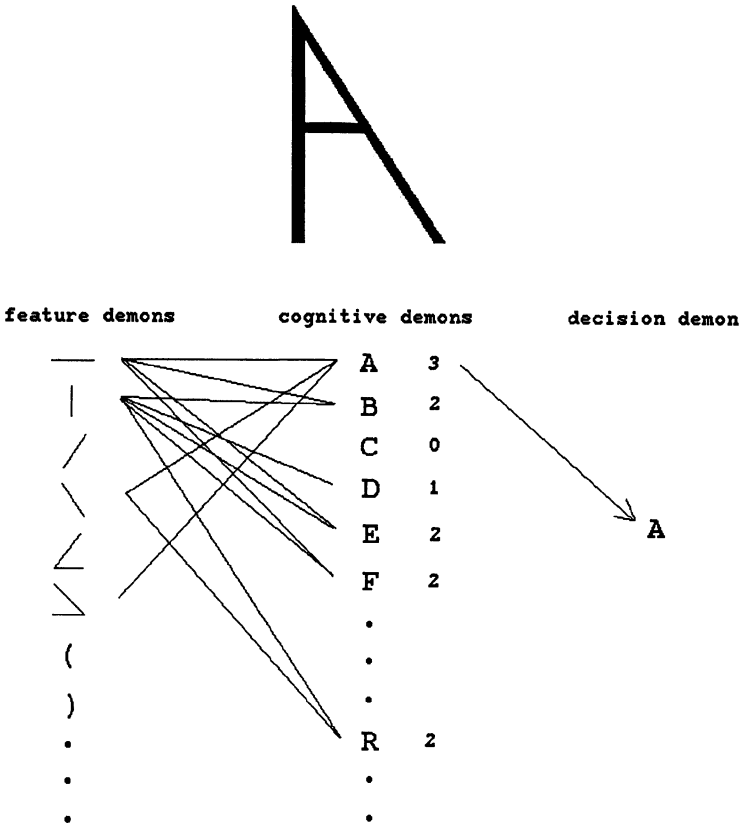
This refines the prototype model by further assuming that the prototypes are stored in memory as a network of discrete features. We then find a match between a list of features for the stimulus and a list of features stored in memory. A simple example of such a model is a pandemonium model. This has image demons, feature demons, cognitive demons and a decision demon. Thus, given an image, the first stage decides which feature demons are present. Each of these then "votes" for one of a set of cognitive demons, finally resulting in a decision. Continuing the previous example, suppose that the image demon is

Obviously, the voting here has been restricted - we might have included e.g. $\wedge$, $\vdash$ to have obtained a different outcome. However, feature models are challenged by experimental findings which show that the *context* of the stimulus influences the ability to identify or recognize the stimulus. This leads psychologists to believe that human recognition may involve both serial and parallel processing of information.

2.4 Representing the image

We now look at a different approach due to David Marr (1982), which looks at different ways of representing the image. A summary of this is given below.

The complexity varies from a grey level image, which can be easily handled automatically, to something which would be very difficult to program. There are three



Pandemonium Model.

	PURPOSE	PRIMITIVES
grey level image	represents intensity	intensity value at each point in the image
primal sketch	makes explicit information about the 2-dimensional image, primarily the intensity changes and their distribution and organization	zero-crossing, blobs, terminations, edge segments, virtual lines, groups, curvilinear organization, boundaries.
$2\frac{1}{2} - D$ sketch	makes explicit the information and rough depth of the visible surfaces, and contours of discontinuities in a viewer-centred co-ordinate frame	local surface orientation, distance from viewer, discontinuities in depth, discontinuities on surface orientation

stages involved in the processes that derive the primal sketch.

(i) The detection of zero-crossings:

A sudden change in intensity gives rise to a peak or trough in the first derivative, or, to a zero crossing in the second derivative. To detect these, apply the filter

$$\nabla^2 G = \frac{-1}{\pi\sigma^4}(1 - \frac{r^2}{2\sigma^2})e^{\frac{-r^2}{2\sigma^2}}$$

which will then transform a grey level image to a binary image. Although it seems implausible that humans use this method for edge detection, there is physiological evidence which allows Marr to claim:

It is not too unreasonable to propose that the $\nabla^2 G$ function is what is carried by the X cells of the retina and lateral geniculate body, positive values being carried by the on-centre X cells, and negative values by the off-centre X cells.

(ii) The formation of the raw primal sketch:

Obtain zero crossings for a number of different-sized channels (different σ). The question of how humans combine these channels is unanswered, but the raw primal sketch is nevertheless a very rich description of an image.

(iii) The creation of the full primal sketch:

This makes explicit important relationships. Starting with the raw primal sketch and operating on it with processes of selection, grouping and discrimination to form tokens, virtual lines and boundaries at different scales.

From a psychological distance measure perspective, we would like to be able to automatically classify and describe images as a full primal sketch. We can achieve this only in part.

3. Towards the desirable via the possible

The first step is to convert the grey level image to a binary image using filters or by thresholding. Then classify each pixel according to type by applying the filter

2	4	8
256	1	16
128	64	32

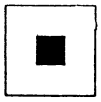
So, for example, the pixel which has 8-neighbor pattern  has type 391. This results in 512 types, and for each type we can compute an error rate by comparing the restored image with the true image. (Note that the error rate will not be symmetric. We can obtain a transition matrix N_{ij} for the number of times pixel type i in the true image corresponds to pixel type j in the restored image $0 \leq i, j \leq 511$.

In addition we can take the difference between the true and restored images and consider the pattern types which result. For small images, counts may be low. We can combine types which are inversions, rotations, reflections, and group together those types which are of no interest. Note that the ordering of the filter facilitates this process since a rotation through $\pi/4$ corresponds to a multiple of 4. Expected values for much of the above can be calculated for specific hypotheses, and tests on these can be performed. This approach is motivated by three factors:

(i) Many restoration algorithms consider 4– or 8–neighbor patterns, so this may yield further insight into which is preferable.

(ii) There is a need for a flexible, simple approach which can be further adapted and modified.

(iii) It seems to be within the spirit of psychometric models. We could consider the 512 types as demons or view them as local "primitives", e.g.

zero-crossings  , blobs  , terminations  .

This approach can be useful in the understanding and analysis of the performance of restoration algorithms, but for many purposes it is desirable to obtain an overall distance from one image to another. The pixel-by-pixel comparison is criticized because it ignores the spatial context. What is required is a new labeling which provides information not only about a pixel, but also pixels in that neighborhood. This new label should be constructed so that it could be used in a distance metric. Note that this is not the case for our 512 types since there is no straightforward way to measure the distance between any two types. We would like closeness in the labeled space to mimic closeness in the pixel neighborhoods.

We could associate a line with many of the types, and then measure the distance between the lines. This would give a nice interpretation to simple shifts and rotations, but unfortunately many of the interesting types require more than one line to describe them, and there is the added complication of dealing with inversions. A more flexible approach would be to associate a "black" circle and a "white" circle with each 3×3 neighborhood. For example, the black circle could be the smallest circle which would cover the centers of all the black pixels and none of the white pixels. This would give a 6-vector label for each neighborhood (center and radius for both circles), and enables us to describe 158 of the 512 types. We have adopted the convention that if only one circle is needed (as for a uniform array), then the other circle is given by the same center with a negative radius - as though it had been "cut away". We can then obtain an overall measure by reverting to a pixel-by-pixel comparison of the new labels. Here, we have used the minimum distance between corresponding white or black circles, where distance between circles is taken as $((x_1 - x_2)^2 + (y_1 - y_2)^2 + (r_1 - r_2)^2)^{1/2}$. This is then averaged over all pixels.

We have conducted a few experiments to determine the characteristics of this measure. For large scale details the measure behaved in a similar fashion to the ordinary pixel-by-pixel measure, but for small scale details, we obtained slightly preferable results. Some of the large-scale test images and corresponding transformations were:

image	transformation	distances (ordinary, new)
half plane ($y > 0$)	half plane ($y > 1$)	.016, .027
	half plane ($y > 2$)	.031, .054
	half plane ($y > 3$)	.047, .099
	half plane ($y > .05x$)	.012, .020
	half plane ($y > .1x$)	.025, .047
	half plane ($y > .2x$)	.050, .112

image	transformation	distances (ordinary, new)
circle radius 20	translate (1,0)	.095, .054
	translate (2,0)	.039, .089
	translate (1,1)	.028, .067
	translate (2,1)	.044, .010
image	transformation	distances (ordinary, new)
bar (3 wide)	translate 1	.031, .054
	translate 2	.062, .105
	translate 3	.094, .193
	remove	.047, .099

Similar results were obtained for a diagonal line, and squashing this circle. The important differences in the measures are seen in the small-scale details. The ordinary error rate gives the same distances for the removal of a 2 pixel-wide bar, or a 2×2 square, as for a single translation of the same object. The new distance measure gives clear preference to the translations.

image	transformation	distances (ordinary, new)
square (2×2)	translate (1,0)	.016, .061
	translate (2,0)	.031, .121
	translate (1,1)	.023, .070
	translate (2,1)	.031, .129
	remove	.016, .075

4. Examples

These images are taken from Ripley (1986) (Figures 2, 3) and Ripley & Taylor (1987) (Figures 4 - 6). For each method of reconstruction we display the binary images, together with the "differenced image".

Error rates are shown for each pixel type of interest. The images are relatively small, so we have combined rotations, reflections and inversions.

All of the non-spatial restorations (Figures 3, 4) have similar error rates (non-significant χ^2) for each pixel type, as expected.

An interesting point is that the error rates for the "corner" pixel-types are very different for the restoration methods which use 4-neighbors rather than 8-neighbors; compare Figures 5 and 6. This is because the corners, when viewed through a 4-neighborhood perspective, could be straight edges, which are very plausible.

Note that small translations of a part of the image (as seen in Figure 2) are easily seen in the differenced image, and that the error rate for the edge-type pixel is very much greater than the overall error rate.

The distances (D) of the new circle measure for these images have a similar ordering to the overall error rate. A desirable difference of these values is that the small translation in Figure 2 is now preferred to the restoration of Figure 5, and the non-spatial restorations are relatively more costly.

Figure 2. Distribution of types in true image, with error rate for each type.

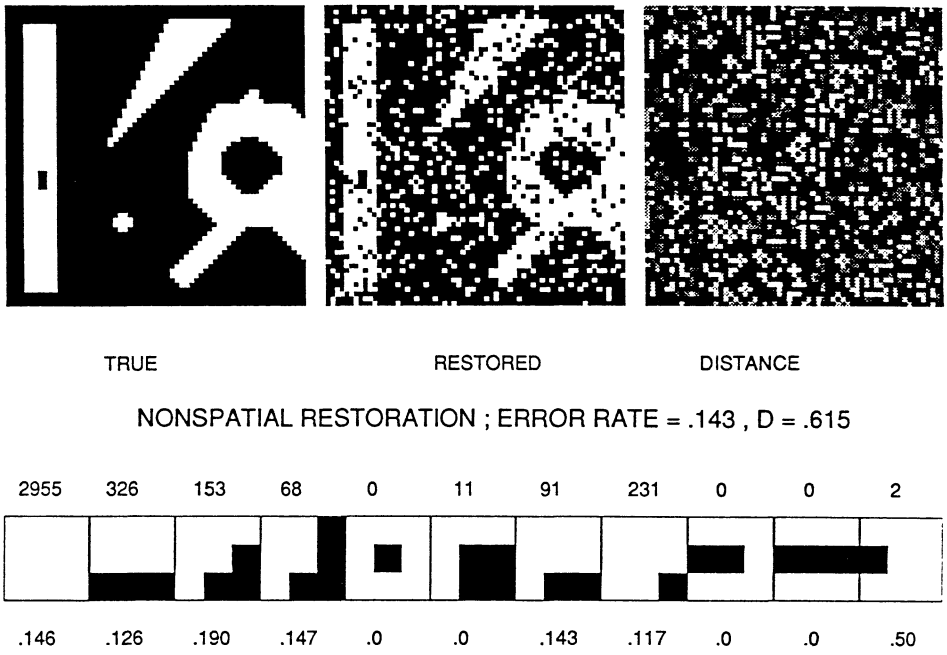
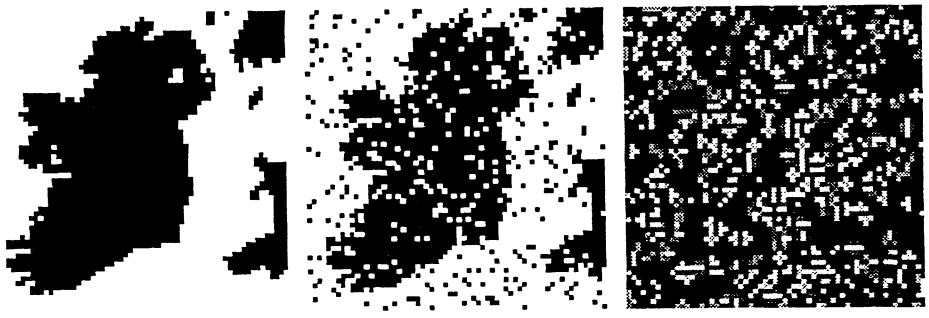


Figure 3. Distribution of types in true image, with error rate for each type.

Clearly, this new error measure still has shortcomings. Many of the possible pixel types have no circle representations as they are too disconnected (Ireland has 1% of such types); here we have chosen to ignore these, but other ad hoc approaches are possible.



NONSPATIAL RESTORATION ; ERROR RATE = .093 , D = .449

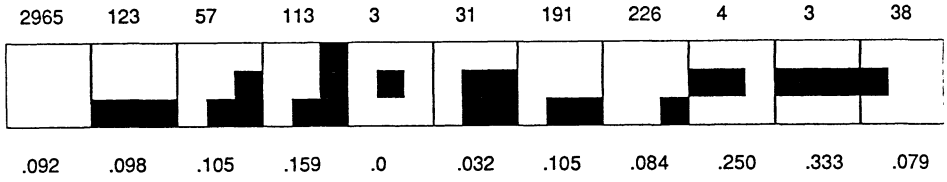


Figure 4. Distribution of types in true image, with error rate for each type.

BESAG SIMULTANEOUS (4) ; ERROR RATE = .022 , D = .078

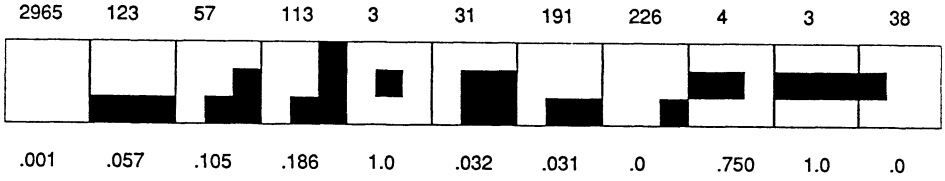


Figure 5. Distribution of types in true image, with error rate for each type.

In addition, other circle representations, perhaps using the largest circle, are possible, although not so easy to compute. Obviously, our set of examples is small, but further experiments should give more insight into the error measures described here.

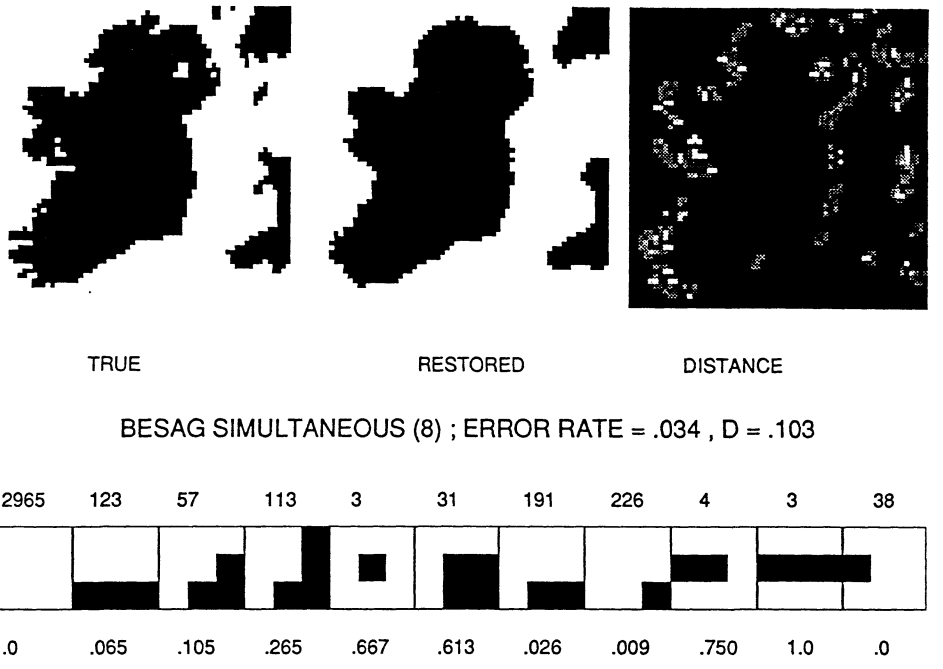


Figure 6. Distribution of types in true image, with error rate for each type.

Acknowledgements

I am grateful to Brian Ripley for allowing me access to these image files. They were drawn using the interactive image analysis program 'Z' (Baddeley, 1988) whilst visiting CSIRO Division of Mathematics and Statistics, Sydney.

References

Baddeley, A. J. (1987). A class of image metrics. *Proceedings of the ANZAAS Congress*, Townsville, Queensland, Australia.

Marr, D. (1982). *Vision: a computational investigation into the human representation and processing of visual information.* San Francisco: W. H. Freeman.

Ripley, B. D. (1986). Images and pattern recognition (with discussion). *Canadian Journal of Statistics* **14**, 83-111.

Ripley, B. D., & Taylor, C. C. (1987). Pattern Recognition. *Science Progress* **71**, 413-428.