Digital Geometric Methods in Image Analysis and Compression

Ari Gross[1] and Longin Latecki[2]

[1] University Graduate Center and Queens College,
City University of New York, Flushing NY 11367, USA
[2] University of Hamburg, Vogt-Kölln-Str. 30,
22527 Hamburg, Germany

Abstract. One of the important problems related to image analysis and compression is finding repeated structure. Although the focus of this paper is developing digital geometric models and methods for finding regular structure in digital document images, the applicability of the digital geometric approach is also demonstrated on images taken under orthographic and perspective projection. First, a fast linear-time algorithm is given to compute the static threshold that minimizes the non-well-composedness or weak connectivity of the document image. Next, a new digital similarity measure is introduced that outperforms the standard similarity measures, including the Hausdorff distance, with respect to determining if two discrete objects in the image are digitizations of the same prototype. This measure is then used in a model-based compression algorithm, and a variation of the algorithm is developed for finding structure in images taken under affine and perspective transformations.

1 Introduction

In previous work [3] [4] [8], the authors have focused on mathematically modeling the digitization process and on developing digitization rules and related algorithms that guarantee that a digitization is topology preserving. In this paper, we focus more on the application side, demonstrating how discrete spatial models and related digital similarity measures can be used in model-based compression of digital documents. Part of this paper focuses on the need for correct digital similarity measures. In particular, we will focus on the Hausdorff distance and its variations. There has been considerable important research on the Hausdorff distance [5] [10]. In this paper, however, we will focus on the relationship between the Hausdorff distance and the digitization process. It is shown that the Hausdorff distance is interesting exactly because it is closer to a digital similarity measure than the standard measures such as Hamming distance, weighted Hamming distance [11], residual entropy [6], and template distance [7].

The paper is structured as follows: First, a fast topology-preserving thresholding algorithm is presented. Next, we derive a new Hausdorff-based digital similarity measure and demonstrate its effectiveness with respect to document image compression. This measure is compared to the standard bidirectional Hausdorff distance used in document image analysis. Finally, a variation on the algorithm for finding repeated image structure is shown to be applicable to images taken under orthographic and perspective projection.

2 Computing a Topology Preserving Threshold

In previous work [2], the authors considered under what conditions a digital image is topology preserving. It was proven that for any **r parallel regular** set, i.e., any set that supports an inner and outer osculating ball of radius r at every point on its boundary (see [8]), a digitization resolution of r for the diameter of a grid square always guarantees that topology is preserved under monotonic digitization and that the resulting discrete set is well-composed, i.e., it has no checkerboard patterns. Well-composed sets have many desirable properties that make them particularly amenable to image processing algorithms, see [9]. It was shown in [2] that selecting a gray-level threshold that minimized the number of checkerboard neighborhoods also seemed to approximately minimize the sum of false topological connections and disconnections. This threshold also gave us very high recognition rates when applying subsequent OCR (using Omni-Page) to the thresholded binary document.

We now present an efficient algorithm for computing the threshold that minimizes the non-well-composedness of a gray-level image:

Starting at the top of the gray-level image, every 2 x 2 neighborhood is visited. For each such neighborhood, we consider the gray-level values for the 2 pairs of diagonal points a_1, a_2 and b_1, b_2. Let $a_{min} = \min(a_1, a_2)$, $a_{max} = \max(a_1, a_2)$, and similarly for b_{min} and b_{max}. Next, we consider the two closed intervals $[a_{min}, a_{max}]$ and $[b_{min}, b_{max}]$. If these two intervals intersect, then no threshold will cause this local 2 x 2 neighborhood to become non-well-composed and we need not consider it further. If, however, the two intervals are disjoint then assume, without loss of generality, that $a_{max} < b_{min}$. Assume further that when the threshold is set to some value t, $0 \leq t \leq 255$, that all pixels with gray-level values $\leq$ t are set to 0 (black), while all pixels with gray-level $>$ t are set to 255 (white). As we traverse the 2 x 2 neighborhoods of the image, two arrays plus_chk[] and minus_chk[] are maintained. These, respectively, keep track of the number of checked neighborhoods that are added or subtracted at each gray-level value. For each neighborhood with disjoint intervals, we increment the two arrays: ++plus_chk[a_{max}], ++minus_chk[b_{min}]. After the entire image has been traversed, these two arrays are used to compute the values for a third array chk[]. Initially, all the values in chk[] are set to zero. Then we compute the values of chk[] iteratively: chk[-1] = 0; chk[i]=chk[$i-1$]+plus_chk[i]-minus_chk[i].

The threshold value selected using this algorithm is appealing for several reasons. One important reason is that the algorithm minimizing the weak connectivity of the document consistently yields a value very close to the threshold value that minimizes the number of models required to match the connected components using the Hausdorff distance metric or a variation introduced in the next section.

3 Digitization Invariant Similarity Measures

In symbolic compression and model-based image coding [1] [11], there is a need for both fast and accurate techniques for comparing the similarity of two discrete planar shapes. A review of several of the similarity measures frequently used is given in [1] and includes the Hamming distance, the weighted Hamming distance, sum of weighted AND-NOTs, residual entropy, and degradation probability. These measures do not model the digitization process and are not invariant with respect to digitization. As a result, the problem of finding the correct number of models or prototypes is often considered a clustering problem, and it is accepted that a certain degree of mislabeling will occur. This mislabeling of connected components necessitates a correction phase, which takes the form of a residual map. This residual map ensures that the compression is lossless but is very expensive. As sensors improve, it is very desirable to put this model matching process into a more rigorous mathematical framework and discard the need for a corrective residual map. This is also desirable from the perspective of constructing an image compiler that converts a digital document back into MS Word or LaTeX format since a residual map is really not applicable in this context.

First, we define the Hausdorff distance and its relationship to monotonic topology-preserving digitization. The Hausdorff distance between two sets A and B is defined as

$$H(A,B) = \max(h(A,B), h(B,A)), \text{where}$$
$$h(A,B) = \sup_{a\in A} \inf_{b\in B} ||b-a||,$$

and $||\cdot||$ is some norm.

In [8], we proved the following theorem:

Theorem 1 *Let A be a par(r)-regular set. Then A and $Dig(A,r)$ are homotopy equivalent for every digitization $Dig(A,r)$, and $H(A, Dig(A,r)) \leq r$, where H is the Hausdorff distance.*

Using Theorem 1, we have $H(A, Dig(A,r)) \leq 1$, where the diameter of a grid square is presumed equal to 1. Moreover, since H is a distance metric satisfying the transitivity property, then for any two digitizations $Dig^1(A,r)$ and $Dig^2(A,r)$ we have the constraint $H(Dig^1(A,r), Dig^2(A,r)) \leq 2$. We have used this constraint effectively to find supersets of a given connected component (i.e., digital instance of a model), where the superset consists of all the connected components on the document that could conceivably be digitizations of the same underlying prototype. An example of this is shown in Fig 1.c, where all the letter "b"s are detected using Hausdorff distance 2, in addition to some extraneous letters. The "b"s that are undetected are either connected to adjacent letters or are substantially corrupted by noise. In all the examples that we considered, this constraint was always successful at finding a superset that included all unperturbed homeomorphic digital instances of the underlying model corresponding to a given connected component. Thus, the Hausdorff distance is useful as a necessary condition for two instances to belong to the same prototype.

On the other hand, using the Hausdorff distance as a sufficient condition for class membership does not work effectively nor does it correctly model the digitization process. As an example, consider the digital document shown in Fig 1.d. In this case, a Hausdorff distance of 1 was used as a criterion for grouping connected components together, where the model selected was an instance of the letter "b". As can be seen, this criterion is neither necessary nor sufficient. Some of the "b" instances are missing while other extraneous components have been included. Instead, we introduce a new variation of the Hausdorff distance that corresponds more closely to the digitization process and, in the many documents we considered, admitted no false positive matches. As a result, there may be some degree of model fragmentation but there is no need for maintaining a residual map.

Define $Q_I, Q_{II}, Q_{III}, Q_{IV}$ to be the closed first, second, third and forth quadrants of R^2. For example,

$$Q_I = \{(x,y) \in R^2 \mid x \geq 0 \text{ and } y \geq 0\}.$$

Then we can define the first quadrant directional Hausdorff distance between two sets in R^2 as follows:

$$h_I(A,B) = \max_{a\in A} \min_{\substack{b\in B\\ b-a\in Q_I}} ||b-a||, \tag{1}$$

and similarly for other quadrants.

Next we define

$$H_{i,j}(A,B) = \max(h_i(A,B), h_j(B,A)),$$

where i,j are quadrant numbers and $|i-j| = 2$.

Finally, the quadrant bidirectional Hausdorff is defined as

$$H_Q(A,B) = \min\{\ H_{I,III}(A,B), H_{III,I}(A,B), \\ H_{II,IV}(A,B), H_{IV,II}(A,B)\ \}\,. \tag{2}$$

This measure is a distance similarity measure that models the fact that, at the very least, a model and its digitization can vary by a quadrant Hausdorff distance equal to 1. To see this, consider the fact that even if a connected component A was arbitrarily close to the original model, another digitization of this "model" could have its centroid vary by a translation of up to 1 grid square diameter. Thus, any digitization resolution where $H_Q(A = Dig(M_1, r), B = Dig(M_2, r)) \leq 1$, for distinct models M_1 and M_2, cannot be model preserving. Assuming we originally had a model-preserving digitization, $H_Q(A, B) \leq 1$ can be used effectively as a sufficient condition for two instances A and B to belong to the same model class.

An example is shown in Fig 2.a, where the initial "model" was an instance of the letter "b". Grouping based on this sufficient condition, no mismatches occur although some instances are missed. This matching algorithm can be improved upon by iterating the matching algorithm on the recomputed model. As shown in the example, initially a single instance of the letter "b" was taken to be the model. After the first iteration, additional "b" connected components were matched, as shown in Fig 2.a. Once these instances of the model are found, the model is recomputed using the dilated binary connected components as masks onto the corresponding gray-level components. This process is repeated until convergence of the model. In the example shown, the final set of matched "b" connected components is shown in Fig 2.b, and the final reconstructed gray-level model is shown in Fig 2.c. The thresholded version of the reconstructed image, as shown in Fig 2.d, is closer to a monotonic digitization than the thresholded original image shown in Fig 1.b, as evidenced by the fact that linear segments remain digitally linear (see [4]). The converged set of gray-level models is shown in Fig 2.e.

4 Discrete Sets of Discrete Spatial Objects

Consider once again the document image shown in Fig 1.a. In the algorithm given in the previous section, the model evolved by starting with a single connected component. Every other connected component that matched this initial model with quadrant Hausdorff ≤ 1 was added to the list. After all the connected components that could be matched were added to the list, the model was recomputed by averaging all the matched connected components and rethresholding. This process was iterated until no new matches were found. Using the quadrant Hausdorff as a sufficient condition for two instances to belong to the same object class, we can alternatively keep the evolving model as the set of instances that have been matched so far. Instead of "learning" more about the model by averaging the instances together, we take their union as the representation of the current underlying model that is evolving. Every new instance of the model that is added to the set must eventually be expanded so that every connected component that it matches in a quadrant Hausdorff sense is also added to the set. This process continues until convergence. For the document shown in Fig 1.a, this algorithm generated the reconstructed document image shown in Fig 3.a.

This method of representing the model as a discrete set of discrete spatial objects is also very useful in finding patterns in images and can be used as part of an image search engine to classify images based on symmetric structure. Consider the window tiles shown in Fig 3.b. These tiles are not all scaled versions of each other since the image was taken under perspective projection. Rather than solve for the transformation parameters, the algorithm simply uses a version of the rank quadrant Hausdorff distance to compare connected components to each other. When the algorithm finally converges, nearly all of the window instances are recognized. For the flag image shown in Fig 3.c, the tesselating elements we want to match (i.e., the stars) are mapped onto a non-planar surface under perspective projection. The algorithm described recognized almost all the stars, as shown in Fig 3.c. For the tire tread, the tread elements are mapped to an approximately cylindrical surface and have considerable variation. Nevertheless, the tread elements are matched quite well, as can be seen in Fig 3.d. This method allows us to find shapes that deform smoothly over time. Consequently, it can serve as a useful tool in finding regular structure in images without regard to the underlying surface or the projective transformation.

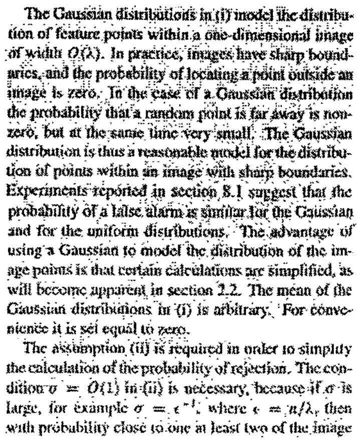
5 Conclusion

In this paper, a fast linear-time algorithm was presented to compute the static threshold that minimizes the non-well-composedness or weak connectivity of the document image. Next, a new digital similarity measure was introduced that outperforms the standard similarity measures, including the Hausdorff distance, with respect to determining if two discrete objects in the image are digitizations of the same prototype. This similarity measure was then used as the basis for a model-based compression algorithm. Finally, we demonstrated that a variation on the method can be extended to finding structure in images taken under affine and perspective transformations.

Acknowledgements: *The authors would like to acknowledge support for this research under NSF grant IRI-9707090 and the QC/CUNY Presidential Research Award. They would also like to acknowledge the very constructive support and assistance of Ruben Lusinyants, Ilya Dondoshansky, Elena Oranskaya, and Navdeep Tinna in this work.*

References

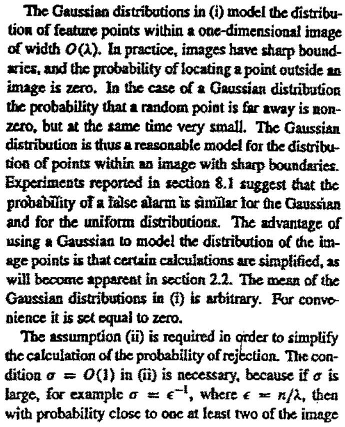
1. D. Doermann, *Document Image Understanding: Integrating Recovery and Interpretation*, PhD thesis, Univ of MD, College Park, 1993.
2. A. Gross and L. Latecki. Homeomorphic Digitization, Correction, and Compression of Digital Documents. *IEEE Workshop on Document Image Analysis*, Puerto Rico, June 1997.
3. A. Gross and L. Latecki. Digitizations Preserving Topological and Differential Geometric Properties. *Computer Vision and Image Understanding*, 62:370-381, Nov. 1995.
4. A. Gross and L. Latecki. A Realistic Digitization Model of Straight Lines. *Computer Vision and Image Understanding*, Vol. 67, No. 2, pp. 131-142, 1997.
5. D.P. Huttenlocher and W.J. Rucklidge, A multi-resolution technique for comparing images using the HHausdorff distance, In *Proceedings Computer Vision and Pattern Recognition*, pp. 705-706, NYC, NY, 1993.
6. S. Inglis and I. Witten. Compression-based template matching. In *Proceedings of the IEEE Data Compression Conference*, 1994.
7. T. Kanungo, R.M. Haralick, and I.T. Phillips. Global and local document degradation models. In *Proceedings of the International Conference on Document Analysis and Recognition*, pp. 730-734, 1993.
8. L. Latecki, C. Conrad, and A. Gross, Preserving Topology by a Digitization Process, to appear in *Journal of Mathematical Imaging and Vision*, 1997.
9. L. Latecki, U. Eckhardt, and A. Rosenfeld, Well-Composed Sets. *Computer Vision and Image Understanding*, 61:70–83, 1995.
10. W. Rucklidge. *Efficient Visual Recognition Using the Hausdorff Distance.* Number 1173 in Lecture Notes in computer Science. Springer-Verlag, 1996.
11. I. Witten, A. Moffat, and T. Bell. *Managing Gigabytes: Compressing and Indexing Documents and Images.* Van Nostrand Reinhold, 1994.



The Gaussian distributions in (i) model the distribution of feature points within a one-dimensional image of width $O(\lambda)$. In practice, images have sharp boundaries, and the probability of locating a point outside an image is zero. In the case of a Gaussian distribution the probability that a random point is far away is nonzero, but at the same time very small. The Gaussian distribution is thus a reasonable model for the distribution of points within an image with sharp boundaries. Experiments reported in section 8.1 suggest that the probability of a false alarm is similar for the Gaussian and for the uniform distributions. The advantage of using a Gaussian to model the distribution of the image points is that certain calculations are simplified, as will become apparent in section 2.2. The mean of the Gaussian distributions in (i) is arbitrary. For convenience it is set equal to zero.

The assumption (ii) is required in order to simplify the calculation of the probability of rejection. The condition $\sigma = O(1)$ in (ii) is necessary, because if σ is large, for example $\sigma = \epsilon^{-1}$, where $\epsilon = n/\lambda$, then with probability close to one at least two of the image

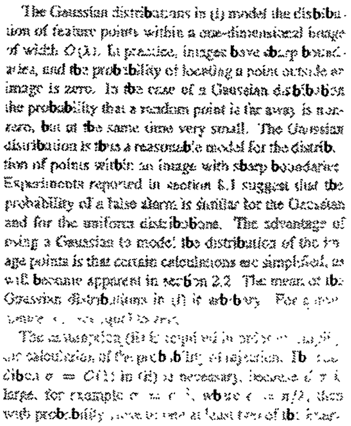
Fig 1a



The Gaussian distributions in (i) model the distribution of feature points within a one-dimensional image of width $O(\lambda)$. In practice, images have sharp boundaries, and the probability of locating a point outside an image is zero. In the case of a Gaussian distribution the probability that a random point is far away is nonzero, but at the same time very small. The Gaussian distribution is thus a reasonable model for the distribution of points within an image with sharp boundaries. Experiments reported in section 8.1 suggest that the probability of a false alarm is similar for the Gaussian and for the uniform distributions. The advantage of using a Gaussian to model the distribution of the image points is that certain calculations are simplified, as will become apparent in section 2.2. The mean of the Gaussian distributions in (i) is arbitrary. For convenience it is set equal to zero.

The assumption (ii) is required in order to simplify the calculation of the probability of rejection. The condition $\sigma = O(1)$ in (ii) is necessary, because if σ is large, for example $\sigma = \epsilon^{-1}$, where $\epsilon = n/\lambda$, then with probability close to one at least two of the image

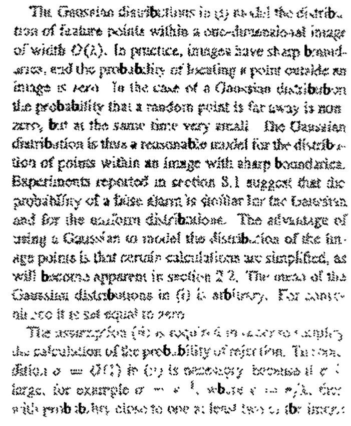
Fig 1b



The Gaussian distributions in (i) model the distribution of feature points within a one-dimensional image of width $O(\lambda)$. In practice, images have sharp boundaries, and the probability of locating a point outside an image is zero. In the case of a Gaussian distribution the probability that a random point is far away is nonzero, but at the same time very small. The Gaussian distribution is thus a reasonable model for the distribution of points within an image with sharp boundaries. Experiments reported in section 8.1 suggest that the probability of a false alarm is similar for the Gaussian and for the uniform distributions. The advantage of using a Gaussian to model the distribution of the image points is that certain calculations are simplified, as will become apparent in section 2.2. The mean of the Gaussian distributions in (i) is arbitrary. For convenience it is set equal to zero.

The assumption (ii) is required in order to simplify the calculation of the probability of rejection. The condition $\sigma = O(1)$ in (ii) is necessary, because if σ is large, for example $\sigma = \epsilon^{-1}$, where $\epsilon = n/\lambda$, then with probability close to one at least two of the image

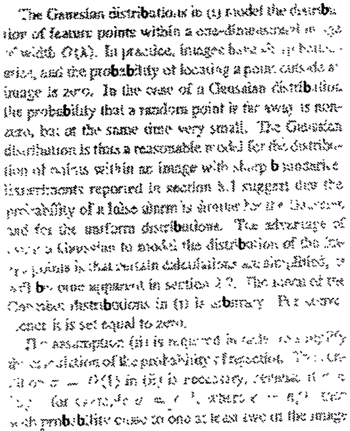
Fig 1c



The Gaussian distributions in (i) model the distribution of feature points within a one-dimensional image of width $O(\lambda)$. In practice, images have sharp boundaries, and the probability of locating a point outside an image is zero. In the case of a Gaussian distribution the probability that a random point is far away is nonzero, but at the same time very small. The Gaussian distribution is thus a reasonable model for the distribution of points within an image with sharp boundaries. Experiments reported in section 8.1 suggest that the probability of a false alarm is similar for the Gaussian and for the uniform distributions. The advantage of using a Gaussian to model the distribution of the image points is that certain calculations are simplified, as will become apparent in section 2.2. The mean of the Gaussian distributions in (i) is arbitrary. For convenience it is set equal to zero.

The assumption (ii) is required in order to simplify the calculation of the probability of rejection. The condition $\sigma = O(1)$ in (ii) is necessary, because if σ is large, for example $\sigma = \epsilon^{-1}$, where $\epsilon = n/\lambda$, then with probability close to one at least two of the image

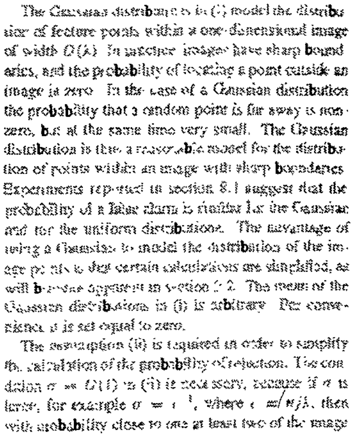
Fig 1d



The Gaussian distributions in (i) model the distribution of feature points within a one-dimensional image of width $O(\lambda)$. In practice, images have sharp boundaries, and the probability of locating a point outside an image is zero. In the case of a Gaussian distribution the probability that a random point is far away is nonzero, but at the same time very small. The Gaussian distribution is thus a reasonable model for the distribution of points within an image with sharp boundaries. Experiments reported in section 8.1 suggest that the probability of a false alarm is similar lor the Gaussian and for the uniform distributions. The advantage of using a Gaussian to model the distribution of the image points is that certain calculations are simplified, as will become apparent in section 2.2. The mean of the Gaussian distributions in (i) is arbitrary. For convenience it is set equal to zero.

The assumption (ii) is required in order to simplify the calculation of the probability of rejection. The condition $\sigma = O(1)$ in (ii) is necessary, because if σ is large, for example $\sigma = \epsilon^{-1}$, where $\epsilon = n/\lambda$, then with probability close to one at least two of the image

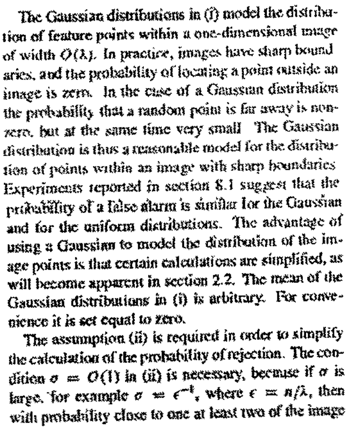
Fig 2a



The Gaussian distributions in (i) model the distribution of feature points within a one-dimensional image of width $O(\lambda)$. In practice, images have sharp boundaries, and the probability of locating a point outside an image is zero. In the case of a Gaussian distribution the probability that a random point is far away is nonzero, but at the same time very small. The Gaussian distribution is thus a reasonable model for the distribution of points within an image with sharp boundaries. Experiments reported in section 8.1 suggest that the probability of a false alarm is similar lor the Gaussian and for the uniform distributions. The advantage of using a Gaussian to model the distribution of the image points is that certain calculations are simplified, as will become apparent in section 2.2. The mean of the Gaussian distributions in (i) is arbitrary. For convenience it is set equal to zero.

The assumption (ii) is required in order to simplify the calculation of the probability of rejection. The condition $\sigma = O(1)$ in (ii) is necessary, because if σ is large, for example $\sigma = \epsilon^{-1}$, where $\epsilon = n/\lambda$, then with probability close to one at least two of the image

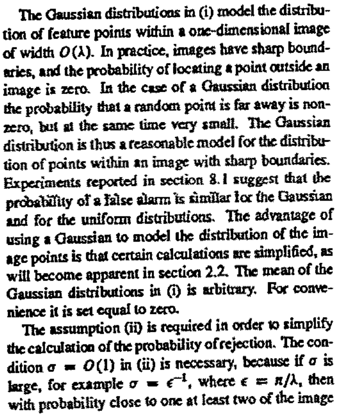
Fig2b



The Gaussian distributions in (i) model the distribution of feature points within a one-dimensional image of width $O(\lambda)$. In practice, images have sharp boundaries, and the probability of locating a point outside an image is zero. In the case of a Gaussian distribution the probability that a random point is far away is nonzero, but at the same time very small The Gaussian distribution is thus a reasonable model for the distribution of points within an image with sharp boundaries Experiments reported in section 8.1 suggest that the probability of a false alarm is similar lor the Gaussian and for the uniform distributions. The advantage of using a Gaussian to model the distribution of the image points is that certain calculations are simplified, as will become apparent in section 2.2. The mean of the Gaussian distributions in (i) is arbitrary. For convenience it is set equal to zero.

The assumption (ii) is required in order to simplify the calculation of the probability of rejection. The condition $\sigma = O(1)$ in (ii) is necessary, because if σ is large, for example $\sigma = \epsilon^{-1}$, where $\epsilon = n/\lambda$, then with probability close to one at least two of the image

Fig 2c



The Gaussian distributions in (i) model the distribution of feature points within a one-dimensional image of width $O(\lambda)$. In practice, images have sharp boundaries, and the probability of locating a point outside an image is zero. In the case of a Gaussian distribution the probability that a random point is far away is nonzero, but at the same time very small. The Gaussian distribution is thus a reasonable model for the distribution of points within an image with sharp boundaries. Experiments reported in section 8.1 suggest that the probability of a false alarm is similar lor the Gaussian and for the uniform distributions. The advantage of using a Gaussian to model the distribution of the image points is that certain calculations are simplified, as will become apparent in section 2.2. The mean of the Gaussian distributions in (i) is arbitrary. For convenience it is set equal to zero.

The assumption (ii) is required in order to simplify the calculation of the probability of rejection. The condition $\sigma = O(1)$ in (ii) is necessary, because if σ is large, for example $\sigma = \epsilon^{-1}$, where $\epsilon = n/\lambda$, then with probability close to one at least two of the image

Fig 2d

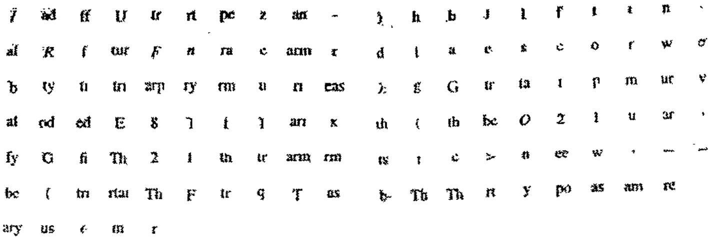
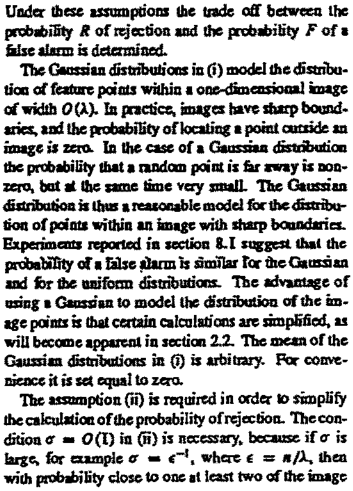

ĩ ad ff U tr rt pc z an - } h b J l f t t n `
al R f tur F n ra c arm r d l a e s c o r w σ
b ty ti tn arp ry rm u n eas); g G tr ta ı p m ur v
at od ed E 8] {] an x th (th be O 2 l u ar '
fy G fi Th 2 l th tr arm rm ts t c > n ee w ' — —
be (tn rtai Th F tr q T as b- Th Th n y po as am re
ary us ε m r

Fig 2e



Under these assumptions the trade off between the probability R of rejection and the probability F of a false alarm is determined.

The Gaussian distributions in (i) model the distribution of feature points within a one-dimensional image of width $O(\lambda)$. In practice, images have sharp boundaries, and the probability of locating a point outside an image is zero. In the case of a Gaussian distribution the probability that a random point is far away is nonzero, but at the same time very small. The Gaussian distribution is thus a reasonable model for the distribution of points within an image with sharp boundaries. Experiments reported in section 8.1 suggest that the probability of a false alarm is similar for the Gaussian and for the uniform distributions. The advantage of using a Gaussian to model the distribution of the image points is that certain calculations are simplified, as will become apparent in section 2.2. The mean of the Gaussian distributions in (i) is arbitrary. For convenience it is set equal to zero.

The assumption (ii) is required in order to simplify the calculation of the probability of rejection. The condition $\sigma = O(1)$ in (ii) is necessary, because if σ is large, for example $\sigma = \epsilon^{-1}$, where $\epsilon = n/\lambda$, then with probability close to one at least two of the image

Fig 3a

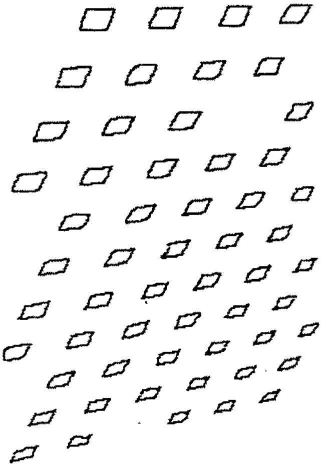


Fig 3b

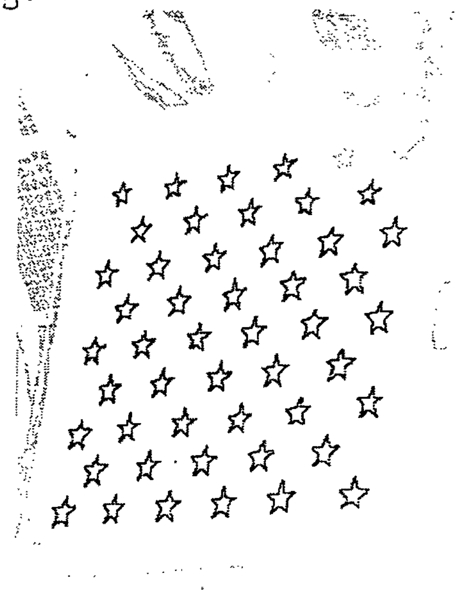


Fig 3c

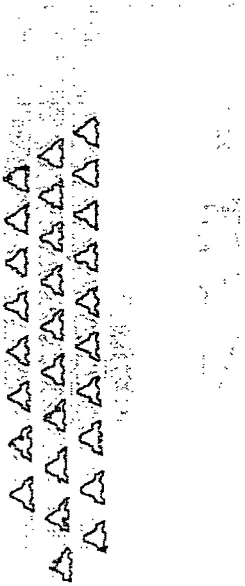


Fig 3d