A Statistical Approach to Multi-Scale Edge Detection.

S.M. Konishi
Smith-Kettlewell Eye Research Institute
San Francisco, CA 94115

J.M. Coughlan
Smith-Kettlewell Eye Research Institute
San Francisco, CA 94115

Alan Yuille
Department of Statistics
University of California at Los Angeles
Los Angeles, CA 90095
yuille@stat.ucla.edu

In Image and Vision Computing. Vol. 21. No. 1. pp 37-48. January. 2003.

A Statistical Approach to Multi-Scale Edge Detection

Scott Konishi, Alan Yuille, and James Coughlan
Smith-Kettlewell Eye Research Institute
2318 Fillmore Street, San Francisco, CA 94115

Abstract— **We propose a statistical approach to combining edge cues at multiple scales using data driven probability distributions. These distributions are learnt on the Sowerby and South Florida datasets which include the ground truth positions of edges. We evaluate our results using Chernoff information and conditional entropy. Our results demonstrate the effectiveness of multi-scale processing and validate previous heuristics such as coarse-to-fine edge tracking.**

Keywords— **Bayesian inference, edge detection, multi-scale processing, empirical evaluation.**

I. Introduction

IT is generally agreed that edge detection should be performed at multiple scales, see [13] for a historical perspective. There is less agreement on precisely how to combine the results of edge detectors at different scales.

In this paper, we propose a statistical approach for combining edge detectors at different scales. This develops from our previous work on evaluating the effectiveness of different edge cues [10], [11], see also [2]. We use joint probability distributions to combine edge cues at different scales. In addition, we study the effectiveness of cues for the *localization* task of determining how close pixels are to an edge. We also explore how much information is lost (if any) when we decimate images to perform multi-scale processing. Our approach is developed and evaluated on the Sowerby and South Florida datasets (see figure (14) for examples of images and ground truth edges for these datasets).

We also relate our work to two alternative methods for combine edge cues based on: (i) coarse-to-fine tracking using scale-space, and (ii) logical operations. The next two paragraphs give backgrounds on these two approaches.

One approach to combining edge cues at different scales is to detect edges at coarse scales, where they are presumably poorly localized, and then track the edges at finer scales to determine the localization. This strategy has been used both for edge detection and for algorithms for solving the correspondence problem for binocular stereopsis, see [13]. This strategy partially motivated the study of scale-space [18], [9]. This study resulted in theorems which supported coarse-to-fine tracking by proving that edges which existed at coarse scales would continue to exist at small scales [19],[1]. These results used Gaussian filters to perform multi-scale processing. They also concentrated on the Laplacian of a Gaussian filter (for historical reasons) which has since been shown to have weak empirical performance, for example see [10]. More sophisticated methods of multi-scale processing have also been proposed, most notably non-linear diffusion [15]. These methods also enable tracking of edges at different scales.

Another approach is to combine the results of edge detectors at multiple scales using logical operations such as AND or OR (often used in industrial applications, S. Geman – personal communication). These approaches are intuitive and simple to implement.

An advantage of our statistical approach is that it enables us to get samples of images containing edges. We have generated such samples in previous work using less realistic edge models [20],[21] which have been used to evaluate human ability to detect edges in images [8]. The more realistic edge probabilities described in this paper will enable us to generate more realistic images containing edges.

In section (II), we give some background on our statistical approach to edge detection [10]. We describe the relationship to generative models in section (III). Section (IV) gives empirical results of our approach and compares them with methods based on logical combinations of cues. In section (V), we extend our statistical approach to estimating the distances of pixels to the nearest edge (which subsumes edge detection as a special case). This enables us to relate our work to coarse-to-fine approaches. Section (VI) demonstrates that little information is lost by decimating the image when performing multi-scale processing (leading to gains in computational efficiency).

II. Background

The background material was first reported in [10], see [12] for a more detailed version.

Statistical edge detection involves learning the conditional probability distributions $P(\phi|\textit{on-edge})$ and $P(\phi|\textit{off-edge})$ for the filter response ϕ conditioned on whether the filter is evaluated *on* or *off* an edge. We can then use the log-likelihood ratio test,

$$\log \frac{P(\phi(I(x))|\ \textit{on-edge})}{P(\phi(I(x))|\ \textit{off-edge})} > T, \tag{1}$$

E-mail: konishi/yuille/coughlan@ski.org.

to determine if a pixel x in image $I(x)$ is an edge, where T is a suitable threshold.

We used two image datasets with ground truth specified. The Sowerby dataset contains one hundred pre-segmented colour images. The South Florida dataset contains fifty grey-scale images. These datasets differ both by the nature of the images and by the methods used to construct the segmentations (the ground truth). The Sowerby dataset consists of outdoor images of the English countryside. The South Florida dataset consists largely of indoor images in Florida supplemented with a few photographs of magazine images. More detailed differences between the datasets are described in [12].

We evaluate the effectiveness of different edge filters, and their combinations, using *performance criteria*. This requires representing the conditional probability distributions by *adaptive non-parametric representations* (e.g. histograms), see [10]. The performance criteria are also used to determine the adaptive non-parametric representations by evaluating the effectiveness of the probability distributions induced by the different choices of bin boundaries. For each edge filter, we choose those bin boundaries which give highest performance, using six bins per filter dimension. Different edge cues were combined by their joint distributions. These were represented by multi-dimensional histograms with bin boundaries determined for the individual edge filters (as above).

Two performance criteria are used. The first criterion, Chernoff Information $C(p,q)$ [5] is a measure of the ease in determining which of two distributions $p(.)$ and $q(.)$ generates a set of samples (all members of the set must be sampled from the same distribution). It is given by

$$C(p,q) = -\min_{0\leq\lambda\leq 1} \log\{\sum_{j=1}^{J} p^{\lambda}(y_j)q^{1-\lambda}(y_j)\}. \qquad (2)$$

It arises in theoretical studies [20] of the difficulty of detecting roads in aerial images [6]. The second criterion is conditional entropy used in section (V) to determine the effectiveness of our approach to localize pixels relative to the nearest ground truth edge.

In addition, we used a decision tree method [16] to select those bin boundary cuts which best help discrimination. This enables us to understand the "guts" of the probability distributions and to determine which aspects are most important for segmentation. In this paper, for example, we use decision trees to find the most effective scales for edge detection. The decision tree approach was also used [12] to prevent *overlearning* [17] since by restricting the number of cuts we can acquire sufficient data to learn distributions even for high-dimensional edge filters (cross-validation [16] was used to determine if we were overgeneralizing).

III. Relationship to Generative Models

One advantage of our statistical approach to edge detection is that it enables us to generate sample images containing edges [20], [21]. These sample images have been used for psychophysics experiments to investigate the ability of human subjects to detect edges [8]. Visual inspection of the sampled images helps determine whether there are sufficient statistical cues to enable edges to be detected.

In more detail, Geman and Jedynak [6] proposed a Bayesian model for detecting roads in aerial images. The road was represented by a contour X. It was assumed that this contour was generated by a prior probability distribution $P(X)$ which imposed smoothness on the shape of the road using a Markov model. The imaging model assumed that filter responses were generated by distributions $P(\phi = y|on)$ and $P(\phi = y|off)$ depending on whether the filter ϕ was evaluated *on* or *off* the road. This gave a probabilistic model $P(Y|X)$ to generate the observed filter responses Y over the entire image conditioned on the position of the contour X. The model assumed that filter responses were independent, conditioned on the position X of the road curve.

Yuille and Coughlan [20] analyzed the Geman and Jedynak model and showed that the detectability of curves in the image depended on the Chernoff information between the distributions $P(\phi = y|on)$ and $P(\phi = y|off)$ and the entropy of the prior distribution $P(X)$. Edge detection gets easier as the Chernoff information increases and the entropy decreases. Yuille and Coughlan illustrated these theoretical results by sampling from the distributions $P(Y|X)P(X)$. Kersten and Schrater used similar stimuli to determine the ability of human observers to detect edges in images and noted that observers are better at detecting straight curves [8].

The empirical probability distributions $P(\phi|on), P(\phi|off)$ reported in this paper can be used to generate more accurate sample stimuli which can be used either to evaluate human performance at curve detection or for generating realistic images for other purposes.

IV. Multiscale Edge Detection

This section studies the effectiveness of edge detection at multiple scales. We concentrate mostly on using the magnitude of the gradient operator $|\vec{\nabla}(.)|$ (which was among the best of the filters evaluated in [10]), but other filters give similar results [12]. The operators are applied at different scales σ by smoothing the image with a Gaussian filter of variance σ^2. We apply the filters to the three colour bands Y, I, Q for the Sowerby dataset and the single grey-scale band Y for the South Florida dataset. In addition, we study the grey-scale band Y of Sowerby and the *chrominance bands* defined by $I/Y, Q/Y$ where we have normalized out the grey-scale Y.

We summarize our findings in the following results.

RESULT I: *Multi-scale filtering is very effective.* This results is not surprising but is a pre-requisite for the rest of the paper. This result is illustrated in figure (1) which was first published in [10]. It shows that multi-scale filtering gives a major improvement in performance for edge detection on the Sowerby dataset. This holds for full colour, grey-scale, and chrominance. The improvement is less dramatic for the South Florida dataset where most of the image structure seems to occur at a single (small) scale (and the background of the image is less cluttered so edge detection is comparatively easy [12]). Multiscale is better able to discriminate between texture edges (which should be discounted) and the edges which correspond to boundaries. It is also able to detect edges of different widths (which occur in Sowerby but rarely in South Florida). The differential operators are the magnitude of the image gradient $|\vec{\nabla}|$, the Nitzberg operator $\vec{N}$ [14], and the Laplacian ∇^2 [13].

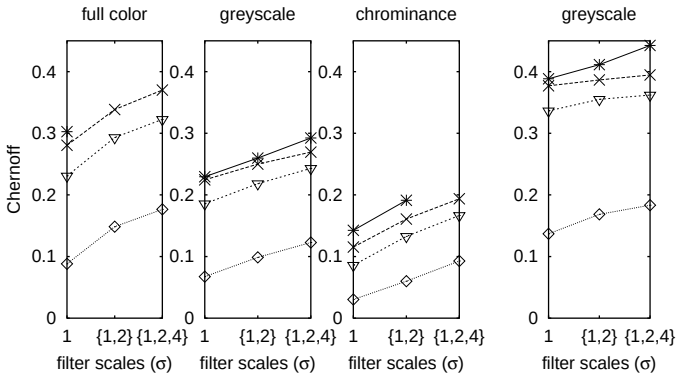

Fig. 1. The advantages of using multi-scale filters on Sowerby (LEFT THREE PANELS) and South Florida (RIGHT PANEL). The edge detector operators are labelled by stars for (N_1, N_2), crosses for N_1, triangles for $|\vec{\nabla}|$, and diamonds for ∇^2. The Chernoff information is shown for: 1 the filter at scale $\sigma = 1$, $\{1, 2\}$ the coupled filter for scales $\sigma = \{1, 2\}$, and $\{1, 2, 4\}$ the coupled filter for scales $\sigma = \{1, 2, 4\}$. The Chernoff always increases as we add larger-scale filters. Decision trees are required when applying filters $\nabla^2, |\vec{\nabla}|$ to (Y, I, Q) at scales $\sigma = 1, 2, 4$, and when applying (N_1, N_2) to chrominance at scales $\sigma = 1, 2$.

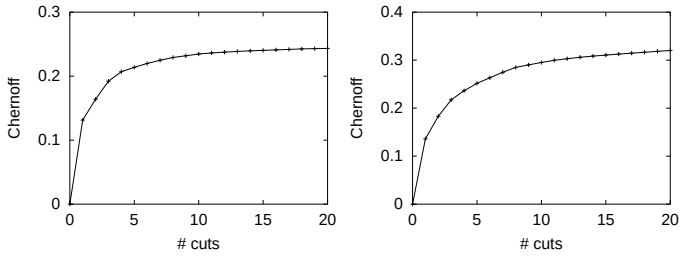

Fig. 2. Left Panel: Decision tree for Sowerby in grey-scale, $|\nabla|_{\sigma=0,1,2,4}(Y)$. Right Panel: Decision tree for Sowerby in full colour, $|\nabla|_{\sigma=0,1,2,4}(Y, I, Q)$.

We now seek to understand the effectiveness of multi-scale edge detection in more detail. Our intention is understand which scales convey most information and to compare the results to more heuristic techniques such as combining filters at different scales by taking thresholds at each scale and then performing logical operations.

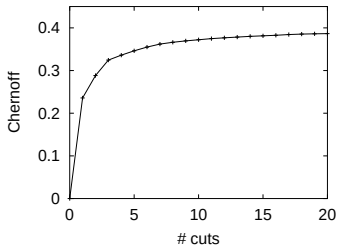

Fig. 3. The Chernoff for the magnitude of gradient operator asymptotes quickly with the number of decision cuts. South Florida dataset.

$\sigma = 2$	$\sigma = 1$	$\sigma = 4$	$\sigma = 1$	$\sigma = 0$	$\sigma = 4$	$\sigma = 4$
0.0082	0.0107	0.0029	0.0346	0.0145	0.0018	0.0067

TABLE I

THE POSITIONS OF THE BIN BOUNDARIES FOR THE SOWERBY DATASET. THE ORDERING IS FROM LEFT TO RIGHT (I.E. THE FIRST BIN BOUNDARY CHOSEN IS AT THE EXTREME LEFT). THE VALUE OF σ LABELS THE SCALE AND THE NUMBER BELOW IT INDICATES THE POSITION OF THE BIN BOUNDARY (IN TERMS OF THE LOG-LIKELIHOOD RATIO).

RESULT II: *The Chernoff information rapidly asymptotes for a small number of decision cuts.* This is shown for Sowerby in figure (2) and for South Florida in figure (3). A small number of cuts is sufficient to give much of the discrimination performance. It is most effective to have one cut per each scale but with a bias towards extra cuts at the smaller scales. The positions of these cuts are given in tables (I,II). This motivates a *poor man's multi-scale edge detector* where one can combine edge detector filters at multiple-scales using decision tree cuts. Moreover, these decision cuts can be determined empirically *without* needing to learn the full joint probability distributions. It would only require a simple learning stage to determine effective positions for the cuts.

RESULT III: *Intermediate filter scales are most effective (when used alone).* The effectiveness of different scales is shown in figure (4) and the left panel of figure (6). The intermediate scales are most effective. Too big is bad (due to poor localization) and too small is bad (due to false positives).

RESULT IV: *Chrominance is most effective at large scales.* This is illustrated by figure (4). This result agrees with studies of biological vision (eg. receptive fields for colour are larger than those for grey-scale).

RESULT V: *The absolute performance of logical AND or*

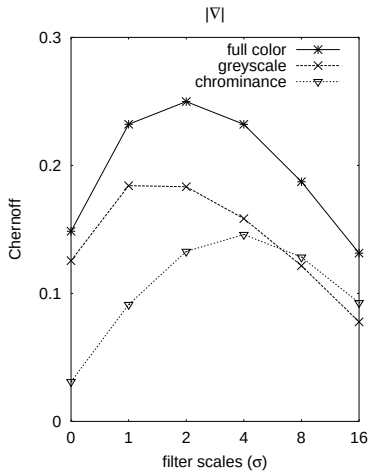

Fig. 4. Chernoffs for magnitude of gradient at a single scale for grey-scale, chrominance, and full colour on the Sowerby dataset. Observe that intermediate scales are most effective. Note also that the chrominance is most effective at larger scales (in agreement with studies of human vision). Sowerby dataset.

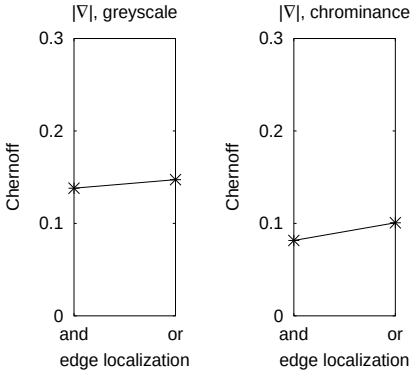

Fig. 5. The logical AND and OR of filters at scales $\{0, 1, 2, 4\}$ are not very effective. Left Panel: grey-scale. Right Panel: chrominance.

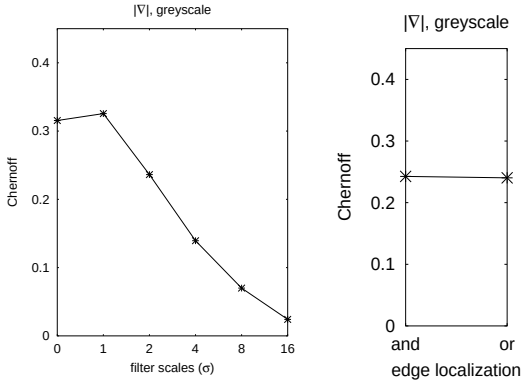

Fig. 6. Left Panel: the Chernoff for the magnitude of gradient operator at a single scale for the South Florida dataset. Observe that, in contrast to the results for Sowerby, the most effective scale is the smallest one. Right Panel: the logical AND and the logical OR are also not very effective on South Florida.

$\sigma = 0$	$\sigma = 1$	$\sigma = 1$	$\sigma = 1$	$\sigma = 2$	$\sigma = 0$	$\sigma = 0$
0.0331	0.0061	0.0509	0.0183	0.0027	0.0801	0.0140

TABLE II
THE POSITIONS OF THE BIN BOUNDARIES FOR SOUTH FLORIDA, SIMILAR CONVENTIONS TO PREVIOUS FIGURE.

OR filters is comparatively disappointing. This is shown in figure (5) and the right panel of figure (6). Better performance can be obtained using a small number of decision tree cuts, as described in RESULT II.

V. Localization of Edges

To study the effectiveness of multi-scale edge detection we now turn to the harder task of classifying all the pixels in the image depending on their distance from the nearest edge. We determine how effective the different scales are at these tasks.

Estimating the localization of a pixel (relative to the nearest edge) is a straightforward application of Bayesian decision theory. For each pixel we compute the probability that it is a specific number of pixels away from an edge. From this, we compute the Chernoff information, and obtain ROC curves, for binary decision tasks, such as whether a pixel is less than, or more than, two pixels from an edge. In addition, we use conditional entropy to evaluate the effectiveness of our filters for simultaneously classifying pixels into multiple classes based on their distance from the nearest edge.

A. Binary Classification

In this subsection, we study edge localization by classifying pixels depending on whether they are less than (or equal to) or greater than w pixels from an edge. Let $\omega(x)$ be the distance of a pixel x to the nearest edge. For each w, we classify pixels into the following two classes: (i) $\alpha_1 = \{x : \omega(x) \leq w\}$, (ii) $\alpha_2 = \{x : \omega(x) > w\}$.

We now learn the probability distributions $P(\phi = y|\alpha_1), P(\phi = y|\alpha_2)$ exactly as we learnt $P(.|on\text{-}edge)$ and $P(.|off\text{-}edge)$ in the previous sections (decision trees are used if necessary). The priors $P(\alpha)$ for $\alpha = (\alpha_1, \alpha_2)$ are also learnt. We evaluate the distributions by their Chernoff information as in previous section. (The next section evaluates these distributions for multiple classifications).

Our main findings are summarized by RESULTS VI-IX. In addition, we found that RESULTS II,V on decision tree cuts and on AND and OR filters are essentially unaltered, see [12]. (I.e (i) the Chernoff reaches an asymptote very quickly as a function of the number of decision cuts, and (ii) the AND and OR rules give disappointing absolute performance.)

RESULT VI: *Localization is possible and higher scale information helps* We obtain reasonable Chernoffs for local-

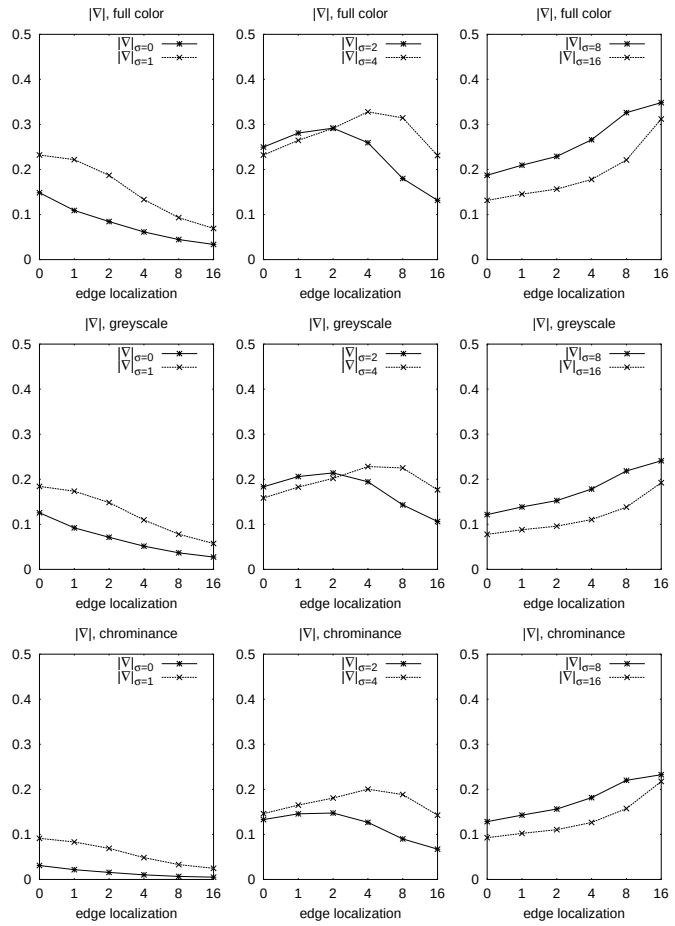

Fig. 7. These figures show that edge localization to a specified degree of accuracy is, not surprisingly, best performed by filters tuned to this degree of localization. The results are for full colour (Top Panels), grey-scale (Middle Panels), and chrominance (Bottom Panels). The figures show the Chernoff for magnitude of gradient at a single scale on the Sowerby dataset. Short scale filters ($\sigma = 0, 1$) are in the Left Panels, mid-scale filters ($\sigma = 2, 4$) in the Center Panels, and large scale filters ($\sigma = 8, 16$) in the Right Panels. The horizontal-axis is accuracy of edge localization, in pixels (e.g., an edge localization of 2 means that the Chernoff information is calculated relative to whether each pixel is within 2 pixels of an edge vs. greater than 2 pixels from an edge.)

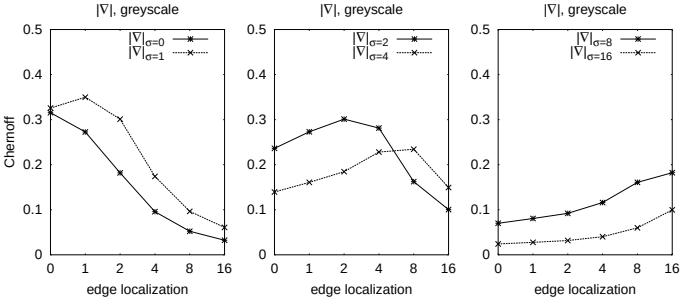

Fig. 8. Chernoff for magnitude of gradient at a single scale, S.Florida dataset. Same conventions as previous figure.

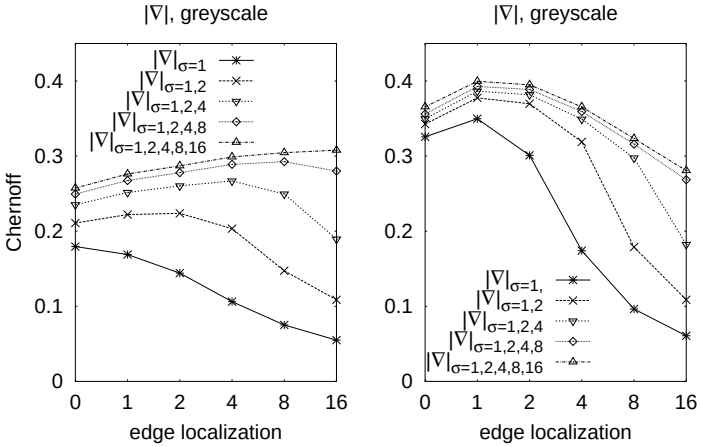

Fig. 9. Chernoff information vs. edgel localization for multi-scale filters, using the greyscale image band. Left: Sowerby dataset. Right: S.Florida dataset.

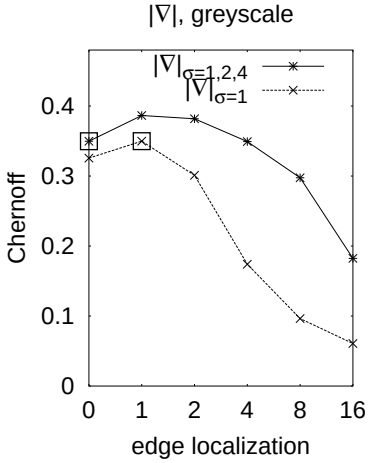

Fig. 10. Justification for coarse-to-fine tracking on South Florida. We plot the Chernoff for the multi-scale filter $|\nabla|_{\sigma=1,2,4}$ and the filter at scale $\sigma = 1$ (the best single-scale filter for South Florida) as functions of edge localization. Observe that the Chernoff for the filter at scale $\sigma = 1$ for localizing to width $w = 1$ is almost identical to the Chernoff for multi-scale to localize the edge to width $w = 0$ (the white boxes indicate the relevant data points).

ization, see figures (7, 8). This result is a pre-requisite for the remaining results in this section. It implies that the quality of the datasets and the ground truths are adequate for our analysis. In addition, the quality of the localization results improve as we add higher scales (particularly for Sowerby), see figure (9).

RESULT VII: *Localization to width w is best done by a filter at scale w.* This is illustrated by figures (7, 8). This result is not surprising, but it has never been empirically demonstrated.

RESULT VIII: *Chrominance localization improves with scale and approaches grey-scale localization at large scales.* This is shown in the lower two panels in figure (7). It is per-

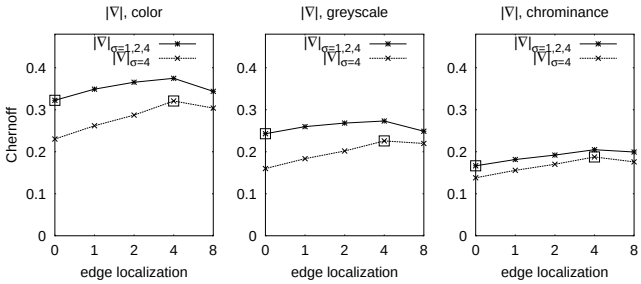

Fig. 11. Coarse-to-fine tracking on Sowerby with full colour (LEFT PANEL), grey-scale (CENTRE PANEL), and chrominance (RIGHT PANEL). The plots show the Chernoffs of the multi-scale filters $|\nabla|_{\sigma=1,2,4}$ and the filter at $\sigma = 4$ (the best single-scale filter for Sowerby) as functions of edge localization. Observe that for all panels, the Chernoff for the filter at $\sigma = 4$ for localizing to width $w = 4$ is almost identical to the Chernoff for the multi-scale filters localizing to width $w = 0$, the square boxes indicate the relevant data points.

haps not surprising that chrominance localization improves with scale (studies of biological vision suggest this). It is unexpected that chrominance localization approaches grey-scale localization at large scale (our definition of chrominance has normalized out the grey-scale intensity).

RESULT IX: *Justification for the strategy of coarse-to-fine edge tracking.*. For both datasets there is an optimal scale σ^*, which is $\sigma^* = 4$ for Sowerby and $\sigma^* = 1$ for South Florida. In both cases, the Chernoffs for using the filter at the optimal scale σ^* to localize the edge to width $w = \sigma^*$ is only slightly smaller than the Chernoff using multi-scale filters to localize the edge precisely at $w = 0$, see figures (10,11). This validates the strategy of detecting the edge at a coarse scale σ^* with only approximate localization, $w = \sigma^*$, and then "tracking" the edge by using smaller scale filters to localize it precisely. But observe that the choice of coarse-scale is dataset dependent.

B. Multiple classification

We now estimate the localization of a pixel (relative to the nearest edge) by computing the probability that it lies within a set of distances from an edge. As in the previous subsection, we let $\omega(x)$ be the distance of a pixel x to the nearest edge. We classify pixels into the following five classes: (i) $\alpha_1 = \{x : \omega(x) = 0\}$, (ii) $\alpha_2 = \{x : \omega(x) = 1\}$, (iii) $\alpha_3 = \{x : 1 < \omega(x) \leq 2\}$, (iv) $\alpha_4 = \{x : 2 < \omega(x) \leq 4\}$, (v) $\alpha_5 = \{x : \omega(x) > 4\}$.

For this multiple-classification task, using filter ϕ, we learn the conditional probability distributions $P(\phi = y|\alpha_i)$ and the priors $P(\alpha_i)$ for $i = 1, ..., 5$. We classify a pixel $\vec{x}$ as α^* where

$$\alpha^* = \arg \max_{i=1,...,5} P(\phi(x)|\alpha_i)P(\alpha_i). \quad (3)$$

To evaluate the performance of filter ϕ, we use the *conditional entropy* $H(\phi|y)$ [5] defined by:

$$H(\phi|y) = -\sum_{y} \sum_{i=1}^{5} P(\alpha_i|\phi = y)P(y) \log P(\alpha_i|\phi = y). \quad (4)$$

We use Bayes rule to compute the posterior distributions $P(\alpha_i|\phi = y)$ from the likelihood functions $P(\phi = y|\alpha_i)$ and the priors $P(\alpha_i)$.

The conditional entropy of a random variable α is a measure of how much uncertainty remains about its value *after* an observation $\phi = y$ has been made (averaged over the possible values of y). The smaller the conditional entropy, the greater the certainly of the value of α after making the observation. Hence we prefer filters with low conditional entropy (the ideal filter would have zero conditional entropy). The conditional entropy can be compared with the entropy $H = -\sum_{i=1}^{6} P(\alpha_i) \log P(\alpha_i)$ of the prior distribution of α before any observations have been made. The conditional entropy is always lower than the entropy because making the observation must, on average, decrease the uncertainty of α.

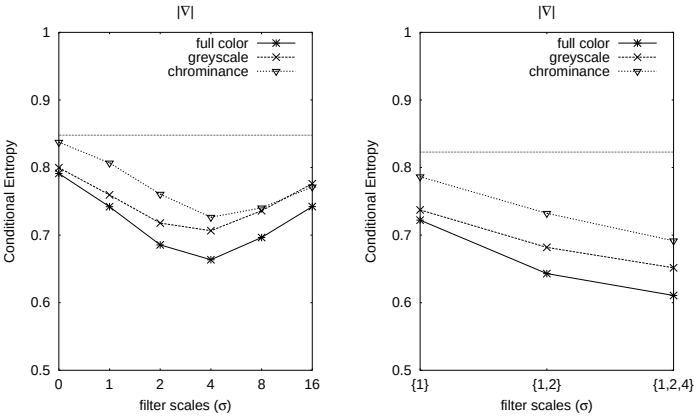

Fig. 12. The conditional entropy for the magnitude of gradient operator on the Sowerby dataset for single-scale filters (LEFT PANEL) and multi-scale filters (RIGHT PANEL). The horizontal axis is the filter scale, e.g., a filter scale of 2 is the filter $|\nabla|_{\sigma=2}$. The dotted line is the entropy (observe that the entropies differ between the plots because running certain large-scale filters requires a modification of the images to remove boundary artifacts, which causes changes in the prior distributions). The classification is localization of pixels to an accuracy of being on-edge to within 0, 1, $(1, 2]$, $(2, 4]$, > 4 pixels.

The results, see figures (12,13), show the advantages of multi-scale filtering for both the Sowerby and South Florida datasets. It can be shown [12] that these results mean that we have a 55 percent chance of correctly classifying a pixel into one of the five categories (for the Sowerby dataset). This is compared to random guessing which would yield 20 percent accuracy. But overall, the results confirm the well-established belief that edge detection is hard!

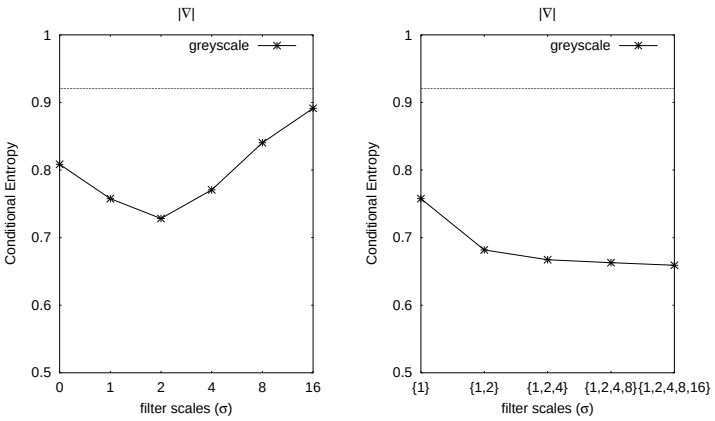

Fig. 13. The conditional entropy for the magnitude of gradient operator on the South Florida dataset. Same conventions as previous figure.

VI. Decimation

Finally, we address the question of how to perform multi-scale processing efficiently. Burt [3] proposed representing an image by a pyramid constructed by repeatedly convolving the image by a Gaussian filter and sub-sampling. This representation is very efficient. But it unclear whether we lose information by performing edge detection on such a pyramid. What are the trade-offs between the amount of edge information we destroy by decimating the image and the potential speed up in computation?

Our result in this section show that hardly any information (eg. ability to detect edges) is lost if we perform multi-scale analysis using a pyramid representation. This has many computational advantages. Suppose we want to detect a hand in an image by a deformable template (see, for example, [4]). Each decimation (by a factor of 2x2) will speed up the algorithm by a factor of 4. Moreover, we can save memory by only storing the edge filter responses on the decimated images.

To decimate an $M \times N$ image by a factor of k, we do a simple average of each $k \times k$ pixel region. (Our results show, somewhat surprisingly, that simple averaging is sufficient and we do not need to smooth the image as we decimate it). This gives the intensity values on the $(M/k) \times (N/k)$ decimated image. We use $k = 1, 2, 4, 8, 16$ on Sowerby images and $k = 1, 2, 4, 8$ on South Florida (because the South Florida images are smaller). To decimate the ground truth for Sowerby by a factor of k, we define a pixel to be an edge provided at least k out of the $k \times k$ pixels in the region are labelled edges. For South Florida, we define a pixel to be an edge provided at least $k/2$ out of the $k \times k$ pixels in the region are labelled edges. We use a different procedure for the two datasets because the edges in Sowerby tend to be twice as thick as those in South Florida. The resulting edge maps are checked visually, see figure (14), to ensure that the resulting edges are plausible (i.e. thin and correctly

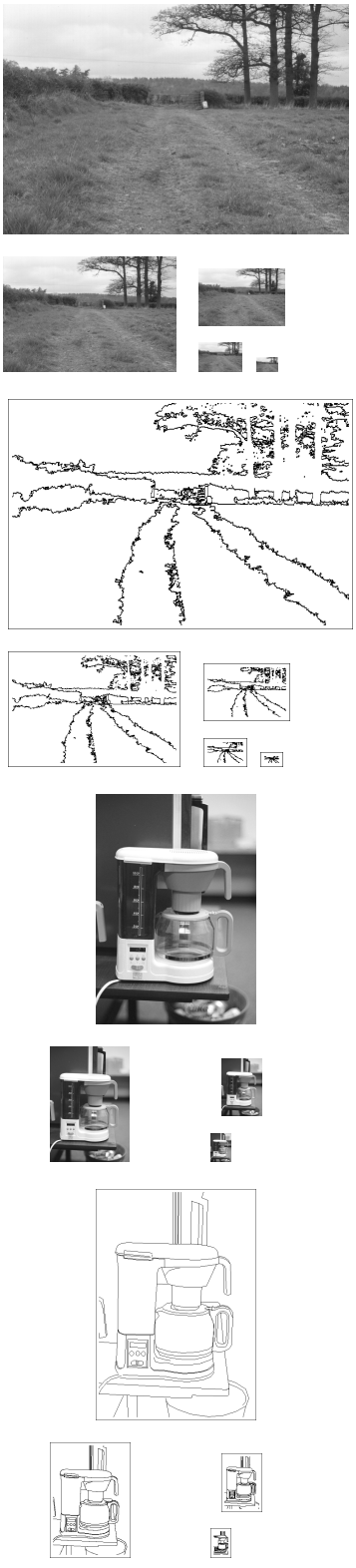

Fig. 14. Decimations of Sowerby and South Florida images. Far left, a typical Sowerby image at five decimations. Centre left, its edge maps for these decimations. Similarly for South Florida in centre right and far right.

located).

Our results show that performing decimation at multiple scales loses very little information (so more sophisticated decimation methods, involving image smoothing, can hardly do any better). To understand this, let the decimation factor be k and the filters have scale σ_k (for $\sigma_k = 1, 2, 4$). In a undecimated image this corresponds to an *effective scale* $\sigma = k \times \sigma_k$. If no information is lost by decimation, then we expect that filters with the same effective σ should have the same Chernoff information. This is shown to be true in figure (15) where we plot the curves of the Chernoff information as a function of edge localization. The curves with a common effective scale are practically identical regardless of the decimation.

Our results on how performance degrades as we decimate the images shows different responses for Sowerby and South Florida, see figure (16). Overall we see surprisingly good Chernoff information even when the images are severely decimated by factors of up to 8 or 16, at least for the Sowerby dataset. For the Sowerby dataset, the information content is roughly constant, or slightly increasing, with increasing effective scale. In fact, the edges are approximately self-similar at each decimation level. This can be seen by noting that the curves (left panel, figure (16)) are essentially translated versions of each other. This means that any edge on any decimated image (at least up to a 16×16 decimation) will look statistically like an edge in the original (non-decimated) image. On the other hand, for the South Florida dataset, the information content drops sharply with increasing effective scale. For large decimations, the edges become very difficult to detect. (This is due to an inherent decrease in the information content, not due to a poor decimation. As noted above, little or no information is lost due to the decimation.)

A surprising implication of this is that, for 4×4 decimations or larger, the Sowerby dataset is actually *easier* to segment (i.e., determine the edges) than the South Florida dataset. For the largest scale we study (8x8 decimation), edges in the South Florida dataset have an information content of only around 0.125 — which is the limit of discriminability, so edges will be very difficult to detect. This occurs despite the large amount of texture in the Sowerby dataset, and the relative lack in the South Florida dataset.

We see two reasons for this. Firstly, the Sowerby dataset contains a lot of "texture edges" which make the images harder to segment than the South Florida set. As the decimation increases the texture edges start becoming weaker (i.e., smoothed out) so the background for Sowerby becomes more distinct from edges faster than the background for South Florida (where edges might start to get "washed out"). Secondly, it appears that the ground truth segmentation for South Florida segmentation shows great attention to precise localization of sharp edges but typically ignores large-scale broad edges such as the folds in a carpet

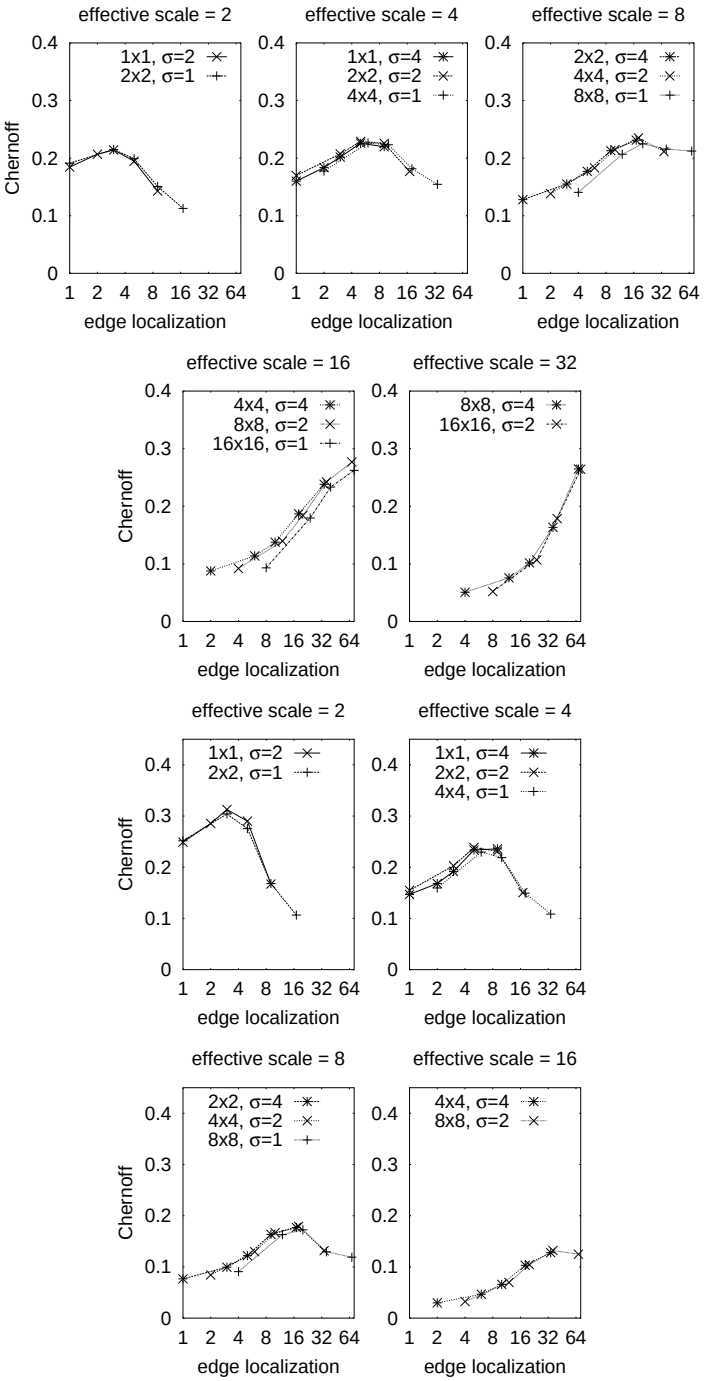

Fig. 15. The overlaying of these curves shows that little information is lost by decimation. The curves plot the Chernoff information as a function of the edge localization (measured in the undecimated image). The filters are the magnitude of the gradient on the grey-scale image. The σ above each graph gives the effective scale. For each effective scale σ we plot all the combinations of decimation factors k and scale σ_k such that $\sigma = k \times \sigma_k$. For example, for scale = 4 we plot the undecimated image at scale $\sigma = 4$, the image decimated by 2×2 at scale $\sigma_2 = 2$, and the image decimated by 4×4 at scale $\sigma_4 = 1$. Top two rows, Sowerby. Bottom two rows, South Florida.

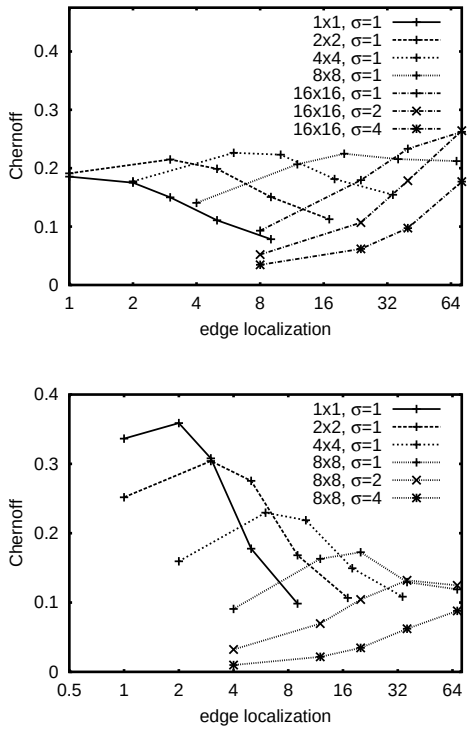

Fig. 16. This figure shows that for Sowerby there is more information at high scales while the opposite is true for South Florida. Left panel: for each effective scale $(1, 2, 4, 8, 16, 32, 64$ for Sowerby we plot the Chernoff information as a function of edge localization. Right panel: similarly for South Florida (with effective scales $1, 2, 4, 8, 16, 32$).

or out-of-focus edges. In contrast, the Sowerby segmentation does the opposite and appears to concentrate on labelling all edges, including those which are broad and hard to localize, at the cost of precise localization. The first observation is an inherent feature of the two datasets, and so would imply that differences are real. The second, however, would mean that some of the differences in information content is an artifact of how the ground truths of the datasets were determined. It remains to be seen which of these two effects is the more prominent effect.

VII. Conclusion

This paper introduced the idea of performing multi-scale edge detection by statistical inference (by extending work reported in [10]). We have summarized our main findings in RESULTS I-IX and by our section on decimation.

Our work shows that edge cues can be combined by statistical inference and that this approach outperforms other methods based on logical combination of cues. But we also show that simple decision rules are often sufficient to achieve close to optimal performance (subject to our evaluation criteria). Our work also gives some justification for the coarse-to-fine strategy used in scale-space. We have also evaluated the effectiveness of colour cues and of decimating the image when performing multi-scale processing.

These empirical probability distributions $P(\phi|\textit{on-edge})$ and $P(\phi|\textit{off-edge})$ can also be used to generate samples of realistic edges which can be used to determine human ability to detect edge contours [8] and determine how this ability relates to theoretical limits [20],[21].

Acknowledgments

We would like to acknowledge funding from the National Institute of Health (NEI) with grant number RO1-EY 12691-01, and from the Smith-Kettlewell core grant. We gratefully acknowledge the use of the Sowerby image dataset from Sowerby Research Centre, British Aerospace. We thank Andy Wright for bringing it to our attention. We also thank Prof. K. Bowyer for allowing us to use the South Florida dataset.

References

[1] J. Babaud, A.P. Witkin, M. Baudin and R.O. Duda. "Uniqueness of the Gaussian Kernel for Scale-Space Filtering". *PAMI(8), No. 1*, January 1986, pp. 26-33. 1986.

[2] K. Bowyer, C. Kranenburg, and S. Dougherty. "Edge Detector Evaluation Using Empirical ROC Curves". In *Proc. Computer Vision and Pattern Recognition*. CVPR'99, Fort Collins, Colorado. pp 354-359. 1999.

[3] P.J. Burt. "Fast filter transforms for image processing". *Computer Graphics and Image Processing*. Vol. 16, pp 20-51. 1981.

[4] J. Coughlan, D. Snow, C. English, and A.L. Yuille. "Efficient Optimization of a Deformable Template Using Dynamic Programming". In *Proceedings Computer Vision and Pattern Recognition. CVPR'98*. Santa Barbara. California. 1998.

[5] T.M. Cover and J.A. Thomas. **Elements of Information Theory**. Wiley Interscience Press. New York. 1991.

[6] D. Geman. and B. Jedynak. "An active testing model for tracking roads in satellite images". *IEEE Trans. Patt. Anal. and Machine Intel.* Vol. 18. No. 1, pp 1-14. January. 1996.

[7] D. M. Green and J. A. Swets. *Signal Detection Theory and Psychophysics*. 2nd Edition. Peninsula Publishing P.O. Box 867. Los Altos, California 94023. 1988.

[8] D. Kersten and P. Schrater. "The Tuning of Vision to Natural Contours: Straighter is Better". *The Annual Meeting of the Vision Sciences Society ARVO*. 2001.

[9] J. J. Koenderink. "The Structure of Images". *Biological Cybernetics*, vol. 50, pp. 363-370, 1984.

[10] S. M. Konishi, A.L. Yuille, J.M. Coughlan and S. C. Zhu. "Fundamental Bounds on Edge Detection: An Information Theoretic Evaluation of Different Edge Cues." In *Proceedings Computer Vision and Pattern Recognition CVPR'99*. Fort Collins, Colorado. 1999.

[11] S. M. Konishi, A.L. Yuille, J.M. Coughlan and Song Chun Zhu. "Statistical Edge Detection: Learning and Evaluating Edge Cues." *Pattern Analysis and Machine Intelligence*. In Press. 2002.

[12] S. Konishi. **PhD Thesis.** In preparation. 2002.

[13] D. Marr. **Vision**. W.H. Freeman and Company. San Francisco. 1982.

[14] M. Nitzberg, D. Mumford, and T. Shiota, **Filtering, Segmentation and Depth,** Springer-Verlag, 1993.

[15] P. Perona and J. Malik. Scale-Space and Edge Detection Using Anisotropic Diffusion. *IEEE Transactions on Pattern Analysis and Machine Intelligence, 12(7)*:629–639, 1990.

[16] B.D. Ripley. **Pattern Recognition and Neural Networks.** Cambridge University Press. 1996.

[17] V.N. Vapnik. **Statistical Learning Theory**. John Wiley and Sons, Inc. New York. 1998.

[18] A. P. Witkin Scale-Space Filtering, In: Proc. 8th Int. Joint Conf.

on Artificial Intelligence, (Karlsruhe, Germany), pp. 1019-1022, 1983.

[19] A.L. Yuille and T. Poggio. "Scaling Theorems for Zero-Crossings". *IEEE Transactions on Pattern Analysis and Machine Intelligence, PAMI-8*, pp 15-25. 1986.

[20] A. L. Yuille and J. M. Coughlan. "Fundamental Limits of Bayesian Inference: Order Parameters and Phase Transitions for Road Tracking" . *Pattern Analysis and Machine Intelligence* PAMI. Vol. 22. No. 2. February. 2000.

[21] A.L. Yuille, J.M. Coughlan, Y-N. Wu and S.C. Zhu. "Order Parameters for Minimax Entropy Distributions: When does high level knowledge help?" *International Journal of Computer Vision*. 41(1/2), pp 9-33. 2001.