BAYESIAN SWITCHING ALGORITHM FOR THE OPTIMAL INCREASING BINARY FILTER

Nina S. T. Hirata[1], *Edward R. Dougherty*[2] *and Junior Barrera*[1]
[1] Dept. of Computer Science, University of São Paulo,
Rua do Matão, 1010, 05508-900 São Paulo, BRAZIL
e-mail: nina@ime.usp.br, jb@ime.usp.br
[2] Dept. of Electrical Engineering, Texas A & M University
College Station, 77840 - TX, United States
e-mail: edward@ee.tamu.edu

ABSTRACT

The optimal windowed, translation-invariant binary image operator depends on conditional probabilities $p(\mathbf{Y}|\mathbf{x})$, where $\mathbf{Y}$ is a pixel value in a window about the pixel. The switching algorithm [1] derives an optimal increasing filter from the optimal operator by switching observation vectors in or out of the kernel in such a way as to obtain an increasing filter with minimal increase in error over the optimal operator. These operators are usually designed by estimating the conditional probabilities from observed-ideal pairs of images. However, samples are typically too small to obtain good estimates of these probabilities. This paper discusses the design of increasing optimal filters by the switching algorithm using prior distributions for the conditional probabilities.

1 INTRODUCTION

A number of methods have been proposed to estimate an optimal binary window filter (equivalently, an optimal stack filter) from data [2]. This paper addresses the switching method, in which the optimal nonincreasing (unconstrained) filter is estimated from the sample data and the optimal increasing filter is derived by switching observation vectors in or out of the kernel in such a way as to obtain an increasing filter with a minimal increase in filter error over the optimal unconstrained filter [3, 1]. As the size of the window increases, the number of potential switches grows exponentially, and therefore it is necessary to design an efficient algorithm, not one that is mainly brute force [3].

This paper applies the switching algorithm in the context of a Bayesian cost function, characterizes the estimated error of the optimal increasing filter in terms of the prior distribution, and compares the expected positive bias of the MAE of the derived optimal filters using both nonBayesian and Bayesian switching (recognizing that in both cases an increasing filter is obtained).

2 SWITCHING ALGORITHM

Denoting an observation vector in the window by $\mathbf{x}$ and the variable to be estimated by $\mathbf{Y}$, the kernel of the optimal filter ψ_{opt} is $\mathcal{K}[\psi_{opt}] = \{\mathbf{x} : p_{\mathbf{x}} = P(\mathbf{Y} = 1|\mathbf{x}) > 0.5\}$. The inversion set of ψ_{opt} consists of all $\mathbf{x} \in \mathcal{K}[\psi_{opt}]$ for which there exists $\mathbf{z}$ such that $\mathbf{x} \leq \mathbf{z}$ and $\mathbf{z} \notin \mathcal{K}[\psi_{opt}]$, together with all $\mathbf{x} \notin \mathcal{K}[\psi_{opt}]$ for which there exists $\mathbf{z}$ such that $\mathbf{z} \leq \mathbf{x}$ and $\mathbf{z} \in \mathcal{K}[\psi_{opt}]$. ψ_{opt} is increasing if and only if its inversion set is null. The switching algorithm begins with the inversion set of ψ_{opt} and efficiently derives a sequence of diminishing inversion sets until it arrives a null inversion set. The sequence proceeds in such a fashion as to produce an increasing filter possessing minimal error among all increasing filters (see [1] for more details).

The algorithm is applied using a cost of switching and the best filter is the one obtained by a switching sequence having minimal cost. If the conditional probabilities $p_{\mathbf{x}}$ and the observation probabilities $P(\mathbf{x})$ are known, then the cost for MAE optimization is $c_{\mathbf{x}} = |2\,p_{\mathbf{x}} - 1|\,P(\mathbf{x})$. In practice, we have only estimates $\hat{p}_{\mathbf{x}}$ and $\hat{P}(\mathbf{x})$ of the probabilities. Since it is not uncommon to have good estimates of the observation probabilities but not of the conditional probabilities (which actually determine the optimal filter), for the sake of simplicity, here we will assume that $P(\mathbf{x})$ is known. If the $p_{\mathbf{x}}$ estimates are very good, then the empirical cost will be close to the true cost $c_{\mathbf{x}}$, but in practice samples are typically too small to obtain good estimates except for a very small set of observations.

3 BAYESIAN SWITCHING COST COMPUTATION

The Bayesian switching algorithm will use the conditional probabilities under the assumption that $p_{\mathbf{x}}$ possesses a prior distribution $f(p_{\mathbf{x}})$ [4]. If we have no knowledge of these probabilities, then we assume $p_{\mathbf{x}}$ to be uniformly distributed over $[0, 1]$.

To analyze the switching cost relative to the observed data, consider the vector $\mathbf{x}$ for which $\hat{p}_{\mathbf{x}} \leq 0.5$. If we make the switch from 0 to 1 and (for the true probability) $p_{\mathbf{x}} > 0.5$, then there is an error decrease of $|2\,p_{\mathbf{x}} - 1|\,P(\mathbf{x})$; if $p_{\mathbf{x}} \leq 0.5$, then there is an error increase of $|2\,p_{\mathbf{x}} - 1|\,P(\mathbf{x})$. Given $\hat{p}_{\mathbf{x}}$, and the conditional

density $f(p_{\mathbf{x}} \mid \hat{p}_{\mathbf{x}})$, the expected error increase is

$$a_{\mathbf{x}} = \Big[\int_0^{0.5} |2p_{\mathbf{x}} - 1|\, f(p_{\mathbf{x}}|\hat{p}_{\mathbf{x}})\, dp_{\mathbf{x}} - \int_{0.5}^{1} |2p_{\mathbf{x}} - 1|\, f(p_{\mathbf{x}}|\hat{p}_{\mathbf{x}})\, dp_{\mathbf{x}} \Big] P(\mathbf{x}) \quad (1)$$

After multiplying through by $P(\mathbf{x})$, the first integral gives the expected increase in error from the switch 0 to 1 from $p_{\mathbf{x}}$ being less than 0.5 given the estimated probability and the second gives the expected decrease in error from $p_{\mathbf{x}}$ exceeding 0.5 given the estimated probability.

Looking at $a_{\mathbf{x}}$, we see that if $\hat{p}_{\mathbf{x}} \leq 0.5$ and $n_{\mathbf{x}}$ is large, then the first integral dominates; however, if $\hat{p}_{\mathbf{x}} > 0.5$ and $n_{\mathbf{x}}$ is large, then the second integral dominates. On the other hand, if $n_{\mathbf{x}}$ is small, then $\hat{p}_{\mathbf{x}}$ provides little conditioning and the integrals tend to depend on the prior distributions of $p_{\mathbf{x}}$. The cost of changing the output value for $\mathbf{x}$ to 1 depends on $n_{\mathbf{x}}$, which is intuitive because there is significant cost in changing when a vector is observed many times, but not when it is rarely observed. The Bayesian switching algorithm is based on the cost $a_{\mathbf{x}}$, not $c_{\mathbf{x}}$.

We simplify the notation denoting $p_{\mathbf{x}}$ by p and $\hat{p}_{\mathbf{x}}$ by $\hat{p}$, to show that Eq. 1 can be simplified :

$$\begin{aligned}
a(\mathbf{x}) &= \Big[\int_0^{0.5} (1-2p)\, f(p|\hat{p}) dp - \int_{0.5}^{1} (2p-1)\, f(p|\hat{p}) dp \Big] P(\mathbf{x}) \\
&= \Big[\int_0^{0.5} f(p|\hat{p})\, dp - 2\int_0^{0.5} p\, f(p|\hat{p})\, dp - 2\int_{0.5}^{1} p\, f(p|\hat{p})\, dp + \int_{0.5}^{1} f(p|\hat{p})\, dp \Big] P(\mathbf{x}) \\
&= \Big[\int_0^{0.5} f(p|\hat{p})\, dp + \int_{0.5}^{1} f(p|\hat{p})\, dp - 2\Big(\int_0^{0.5} p\, f(p|\hat{p})\, dp + \int_{0.5}^{1} p\, f(p|\hat{p})\, dp \Big) \Big] P(\mathbf{x}) \\
&= \Big[1 - 2\int_0^{1} p\, f(p|\hat{p})\, dp \Big] P(\mathbf{x}) \\
&= \Big(1 - 2\, E\,[p|\hat{p}\,] \Big) P(\mathbf{x}) \qquad (2)
\end{aligned}$$

For the case where switches are from 1 to 0, by similar arguments we obtain

$$a(\mathbf{x}) = \Big(2\, E\,[p|\hat{p}\,] - 1 \Big) P(\mathbf{x}) \qquad (3)$$

Hence, the Bayesian switching cost can be expressed by

$$a(\mathbf{x}) = \Big| 2\, E\,[p|\hat{p}\,] - 1 \Big|\, P(\mathbf{x}) \qquad (4)$$

4 EXPERIMENTAL RESULTS

We assume the prior distribution of $p_{\mathbf{x}}$ is a beta distribution. It has two parameters, α and β, and $f(p_{\mathbf{x}}) = p_{\mathbf{x}}^{\alpha}(1-p_{\mathbf{x}})^{\beta} B(\alpha,\beta)^{-1}$, where $B(\alpha,\beta)$ is the beta function. It has mean $\alpha/(\alpha+\beta)$ and variance $\alpha\beta/(\alpha+\beta)^2(\alpha+\beta+1)$. Given observations of image realizations (and therefore $\hat{p}_{\mathbf{x}}$), the conditional density $f(p_{\mathbf{x}}|\hat{p}_{\mathbf{x}})$ is also a beta distribution with parameters $\alpha' = \alpha + n_{\mathbf{x}}$ and $\beta' = \beta + n_{\mathbf{x}} - u_{\mathbf{x}}$, where $n_{\mathbf{x}}$ is the number of times configuration $\mathbf{x}$ has been observed and $u_{\mathbf{x}}$ is the number of times it has been observed with $Y = 1$. Under these conditions, Bayesian switching cost is given by

$$a_{\mathbf{x}} = \Big| 2\,\frac{\alpha'}{\alpha'+\beta'} - 1 \Big|\, P(\mathbf{x})\,. \qquad (5)$$

We use three beta prior distributions for the conditional probabilities to estimate optimal increasing filters for edge noise filtering. We assume the edge noise intensity is parametrized by δ. The noise intensity is inversely proportional to the value of δ. δ ranges from 21 to 49 and its distribution is given by the curve shown in Fig. 1. Figure 2 shows, respectively, part of an ideal image and noise realizations for $\delta = 45, 35, 25$ and 21.

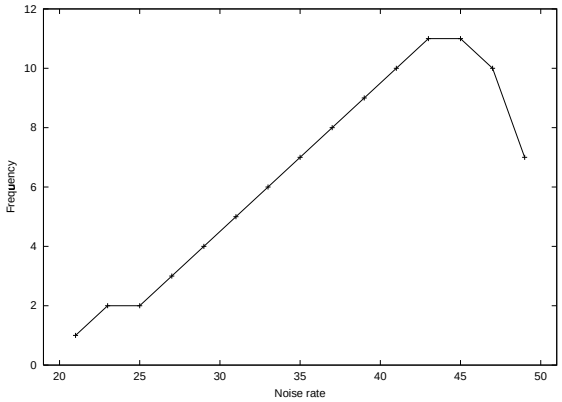

Figure 1: Distribution of the noise intensity parameter.

Let $fq(\delta)$ denote the relative frequency of the noise intensity parameter δ. For our model, $fq(21) = 2fq(23) = 2fq(25) = 3fq(27) = 4fq(29)$ and so on. To estimate a prior distribution, we considered $m * fq(\delta)$ images for each δ and estimated $p_{\mathbf{x}}$ for each group of m images. Therefore, for each $\mathbf{x}$ at most $\sum_{\delta} fq(\delta) = 96$ estimates of $p_{\mathbf{x}}$ have been computed. These estimates were considered as realizations of the distribution $f(p_{\mathbf{x}})$ and then its parameters α and β have been estimated using the maximum likelihood estimation procedure to find the distribution that most likely generated these samples. If most of the estimates $p_{\mathbf{x}}$ were not reliable or if $\mathbf{x}$ were not observed in most of the image groups, then we assumed $\alpha = \beta = 1$. The basic difference among the priors is the amount of data used to estimate $p_{\mathbf{x}}$. Prior 1, 2 and 3 have been obtained using $m = 2, 10$ and 50, respectively (that means prior 1 was obtained from 192 images, prior 2 from 960 images, and prior 3 from 4800 images).

Estimates obtained from large samples are more precise than those obtained from small samples. Therefore,

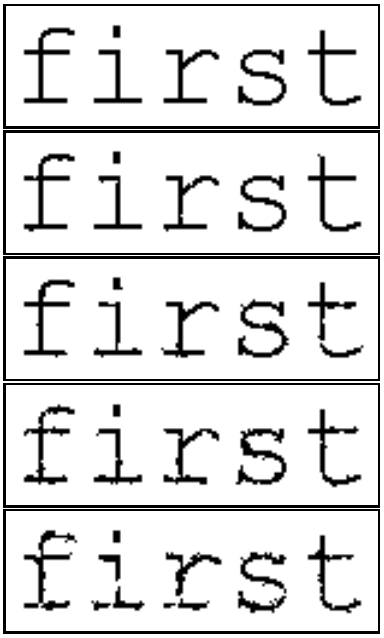

Figure 2: Image samples.

we expect prior 3 being more precise than prior 2, and prior 2 more precise than prior 1.

In fact, expected results have been observed through experimental results using a 5×3 window. For a fixed value of δ, we designed optimal increasing filters and observed the error curve as the number of training data increased. Next we analyze the results obtained for $\delta = 45, 35, 25$ and 21.

For $\delta = 45$ (Fig. 3), prior 3 does as good as possible with no training. Similar effect is observed for prior 2. This reflects the low error to begin with. Prior 1 starts out below no prior but no prior slowly catches up, as we would expect.

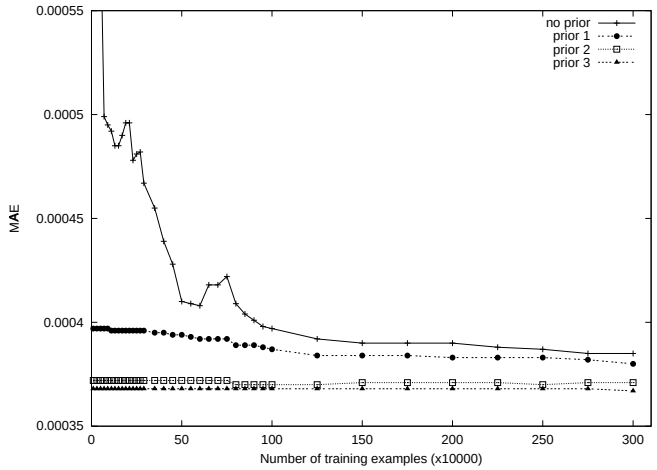

Figure 3: Error curve for $\delta = 45$.

For $\delta = 35$ (Fig. 4), prior 3 also does as good as possible with no training. However, for prior 2, the error curve presents a small decrease as the training data increases. Prior 1 starts out below no prior but no prior catches up, faster than for $\delta = 45$.

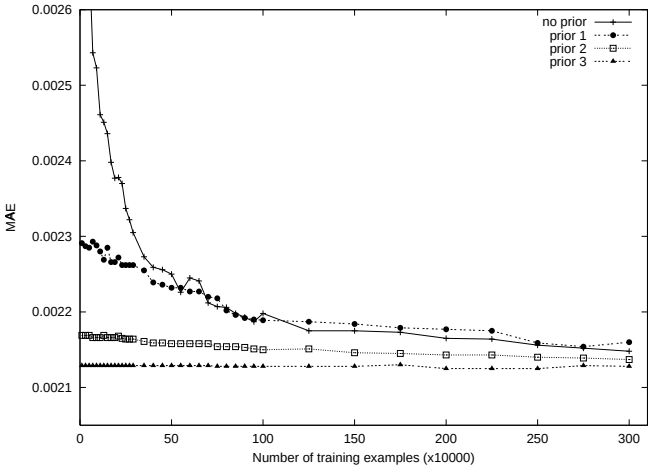

Figure 4: Error curve for $\delta = 35$.

For $\delta = 25$ (Fig. 5), prior 3 does as would be expected, and the decrease of the error curve of prior 2 is more accentuated. The interesting thing is how fast no prior catches up to prior 1. This seems to reflect the lack of training for prior 1. The fact that no prior goes a bit below prior 1 for a while seems to mean that the inaccuracy of prior 1 is putting a drag on no prior. This is the kind of thing we see for bad priors.

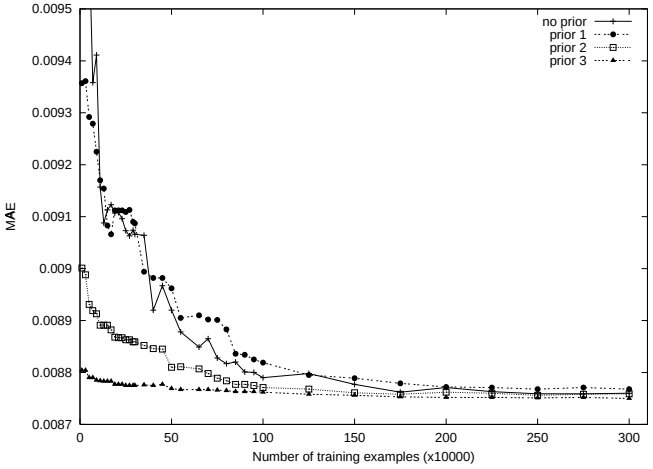

Figure 5: Error curve for $\delta = 25$.

Finally, for $\delta = 21$ (Fig. 6), prior 1 only helps where there is essentially no training. Very quickly it puts a drag on no prior. No prior also catches up to prior 2 and 3.

In all cases prior 3 always does better than prior 2, which in its turn does better than prior 1, as we would expect. Moreover, as the training data increases, training with no prior catches up to training with prior. How

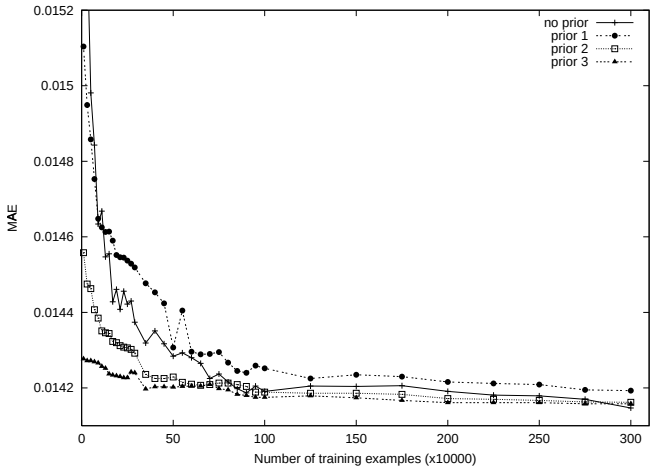

Figure 6: Error curve for $\delta = 21$.

fast it catches depend on the goodness of the prior. If prior is not good, then training with no prior results in smaller error than training with prior for relatively small amount of training data. In the example above, use of prior is advantageous for the noise intensities that are relatively frequent in the distribution considered (Fig. 1).

5 CONCLUSION

For small amount of training data, use of prior data has proven useful. As the training data increases, non-Bayesian training provides better results than those using non adequate priors. However, if prior is correct, it provides better results than nonBayesian training, even for a considerably large amount of training sample.

6 ACKNOWLEDGEMENTS

N. S. T. Hirata acknowledges support from FAPESP (process 98/14328-6).

References

[1] N. S. T. Hirata, E. R. Dougherty, and J. Barrera. A Switching Algorithm for Design of Optimal Increasing Binary Filters Over Large Windows. *Pattern Recognition*, 33(6):1059–1081, June 2000. Special Issue on Mathematical Morphology & Nonlinear Image Processing.

[2] J. Barrera, E. R. Dougherty, and N. S. Tomita. Automatic Programming of Binary Morphological Machines by Design of Statistically Optimal Operators in the Context of Computational Learning Theory. *Electronic Imaging*, 6(1):54–67, January 1997.

[3] N. S. T. Hirata, E. R. Dougherty, and J. Barrera. Efficient Switching Algorithm for Designing Increasing Binary Filters. In E. R. Dougherty and J. T. Astola, editors, *Nonlinear Image Processing X*, volume 3646 of *Proc. SPIE*, pages 185–196, San Jose, CA, January 1999.

[4] E. R. Dougherty and J. Barrera. Bayesian Design of Optimal Morphological Operators Based on Prior Distributions for Conditional Probabilities. *Acta Stereologica*, 16(3):167–174, 1997.