W-operator Window Design by Maximization of Training Data Information

D. C. Martins-Jr, R.M. Cesar-Jr., J. Barrera

USP–Universidade de São Paulo
IME–Instituto de Matemática e Estatística
Rua do Matão, 1010 - Cidade Universitária
CEP: 05508-090, São Paulo, SP, Brasil
{davidjr,cesar,jb}@vision.ime.usp.br

Abstract

This paper presents a technique that gives a minimal window W for the estimation of a W-operator from training data. The idea is to choose a subset of variables W that maximizes the information observed in a set of training data. The task is formalized as a combinatorial optimization problem, where the search space is the powerset set of the candidate variables and the measure to be minimized is the mean entropy of the estimated conditional probabilities. As a full exploration of the search space requires an enormous computational effort, some heuristics of the feature selection literature are applied. The proposed technique is mathematically sound and experimental results show that it is adequate in practice.

1. Introduction

A W-operator is a binary image transformation that is locally defined inside a window W and translation invariant [1]. This means that it depends just on shapes of the input image seen through the window W and that the transformation rule applied is the same for all image pixels. A remarkable property of a W-operator is that it is characterized by a Boolean function which depends on $|W|$ variables, where $|W|$ is the cardinality of W. Examples of W-operators include erosion, dilation, opening, closing, edge detection, hit-miss, median filtering and skeletonization.

W-operators are used in practically any application of binary image processing from image restoration to pattern recognition. An important practical problem is designing W-operators to perform these tasks in specific contexts. There are several heuristic approaches to do that. A formal approach consists in estimating the target operator from collections of input-output image pairs, called training data, that describe the result of the desired transformation in some typical images of the considered domain. Technically, this problem is equivalent to the design of supervised classifiers, in statistical pattern recognition theory, or the learning of Boolean functions, in computational learning theory [2]. This kind of technique has been successfully applied, for example, in the digital documents industry.

Estimating a W-operator from training data is an optimization problem. The training data gives a sample of a joint distribution of the observed shapes and their classification (i.e., Boolean value associated to the observed shape in the output image). A loss function measures the cost of a shape miss classification. An operator error is the expectation of the loss function under the joint distribution. Given a set of W-operators, the target operator is the one that has minimum error. As, in practice, the joint distribution is known just by its samples, it should be estimated. This imply that operators error should also be estimated and, consequently, the target operator itself should be estimated. Estimating a W-operator is an easy task when the sampling of the joint distribution considered is large. However, this is rarely the case. Usually, the problem involves large windows with non concentrated probability mass joint distributions, which requires prohibitive amount of training data.

An approach for dealing with the lack of training data is constraining the considered space of operators. In fact, when the number of candidate operators decreases, it is necessary less training data to get good estimations of the best candidate operator [3]. Usually, the operator space is constrained based on some prior knowledge about desired characteristics of the target operator.

In this paper, we propose a formal technique for estimating a constrained operator space from training data. The idea is to estimate an optimal sub-window W^* that maximizes the available information about the unknown joint distribution, from a given window W and the available training data from the images through W. Choosing a sub-window is equivalent to clustering examples of the train-

ing data, since different shapes may become the same when seen by the sub-window. This, on one hand, decreases the estimation error of the joint distribution by making equivalent rarely observed shapes and, on other hand, increases its estimation error introducing noise in shape classification. The best window W^* should balance properly both effects. The constrained search space will be the set of W^*-operators.

The search space of this problem is the powerset of W, denoted $P(W)$. The criterion to be minimized is the degree of mixture of the observed classes, i.e. when there is more agreement about what class should be attributed to an observed shape, we should be looking to a better sub-window. A measure that reflects this property of a joint distribution is the mean conditional entropy. The important property of entropy explored here is that when the probability mass of a distribution becomes more concentrated somewhere in its domain, the entropy decreases. In other words, when there is a strong probability for a given class in the classification of a shape, the entropy of the conditional distribution should be low. Thus, the optimization algorithm consists in estimating the mean conditional entropy for the joint distribution estimated for each sub-window and choosing the one that minimizes this measure.

Each observed shape has a probability and a corresponding conditional distribution from which the entropy is computed . The mean conditional entropy is the mean of the computed entropies, weighted by the shape probabilities.

As $P(W)$ has an exponential size in terms of the cardinality of W, we adopted some heuristics to explore this space in reasonable computational time. The adopted heuristics were the SFS and SFFS feature selection algorithms [4].

Following this Introduction, Section 2 recalls the mathematical fundamentals of W-operators design. Section 3 introduces the definitions and properties of the mean conditional entropy and presents the proposed technique for generating the minimal window and, consequently, choosing a minimal family of W-operators. Section 4 presents some experimental results of the application of the proposed technique. Finally, Section 5 presents some concluding remarks and perspectives of future researches on this subject.

2. W-operator definition and design

In this section, we recall the notion of W-operator and the main principles for designing W-operators from training data.

2.1. W-operator definition and properties

Let E denote the integer plane and $+$ denote the translation on E. The opposite of $+$ is denoted $-$. A *binary image* or, simply, *image* is a function f from E to $\{0,1\}$. An image f can be represented, equivalently, by a subset X of E by the following relation $\forall x \in E, x \in X \Leftrightarrow f(x) = 1$.

The translation of an image $X \subseteq E$ by a vector $h \in E$ is the image $X_h = \{x \in E : x - h \in X\}$.

Let $P(E)$ be the powerset of E. An *image transformation* or *operator* is a mapping Ψ from $P(E)$ into itself.

An operator Ψ is called *translation invariant* iff (i.e., if and only if), for every $h \in E$,

$$\Psi(X_h) = \Psi(X)_h.$$

Let W be a finite subset of E. An operator Ψ is called *locally defined in the window* W iff, for every $x \in E$,

$$x \in \Psi(X) \Leftrightarrow x \in \Psi(X \cap W_x).$$

An operator is called a *W-operator* if it is both translation invariant and locally defined in a finite window W. Any W-operator Ψ can be characterized by a Boolean function ψ from $P(W)$ to $\{0,1\}$ through the relation, for every $x \in E$,

$$x \in \Psi(X) \Leftrightarrow \psi(X_{-x} \cap W) = 1.$$

Therefore, choosing a W-operator Ψ is equivalent to choose its corresponding Boolean function ψ.

2.2. W-operator design

Designing an operator means choosing an element of a family of operators to perform a given task. One formalization of this idea is as an optimization problem, where the search space is the family of candidate operators and the optimization criteria is a measure of the operator quality. In the commonly adopted formulation, the criteria is based on a statistical model for the images associated to a measure of images similarity, the loss function.

Let $\mathbf{S}$ and $\mathbf{I}$ be two discrete random sets defined on E, that is, realizations of $\mathbf{S}$ or $\mathbf{I}$ are images obtained according with some probability distribution on $P(E)$. Let us model image transformations in a given context by the joint random process $(\mathbf{S}, \mathbf{I})$, where the process $\mathbf{S}$ represents the input images and $\mathbf{I}$ the output images. The process $\mathbf{I}$ depends on the process $\mathbf{S}$ according to a conditional distribution.

Given a space of operators $\mathcal{F}$ and a loss function ℓ from $P(E) \times P(E)$ to $\Re^+$, the error $Er[\Psi]$ of an operator $\Psi \in \mathcal{F}$ is the expectation of $\ell(\Psi(\mathbf{S}), \mathbf{I})$, that is, $Er[\Psi] = E[\ell(\Psi(\mathbf{S}), \mathbf{I})]$. The *target* operator Ψ_{opt} is the one of minimum error, that is, $Er[\Psi_{opt}] \leq Er[\Psi]$, for every $\Psi \in \mathcal{F}$.

A joint random process $(\mathbf{S}, \mathbf{I})$ is jointly stationary in relation to a finite window W, if the probability of seeing a given shape in the input image through W together with a given Boolean value in the output image is the same for every translation of W, that is, for every $x \in E$,

$$P((S \cap W_x, I(x)) = P((S \cap W, I(o)),$$

where S is a realization of $\mathbf{S}$, I is the Boolean function equivalent to a realization of $\mathbf{I}$, and o is the origin of E.

For making the model usable in practice, from now on suppose that $(\mathbf{S}, \mathbf{I})$ is jointly stationary in relation to the finite window W. Under this hypothesis, the error of predicting an image from the observation of another image can be substituted by the error of predicting a pixel from the observation of a shape through W and, consequently, the optimal operator Ψ_{opt} is always a W-operator. Thus, the optimization problem can be, equivalently, formulated in the space of Boolean functions defined on $P(W)$, with joint random processes on $(P(W), \{0,1\})$ and loss functions ℓ from $\{0,1\} \times \{0,1\}$ to $\Re^+$.

In practice, the distributions on $(P(W), \{0,1\})$ are unknown and should be estimated, which implies in estimating $Er[\psi]$ and ψ_{opt} itself. When the window is small or the distribution has a probability mass concentrated somewhere, the estimation is easy. However, this almost never happens. Usually, we have large windows with non concentrated mass distributions, what require prohibitive amount of training data.

An approach for dealing with the lack of data is constraining the search space. The estimated error of an operator in a constrained space can be decomposed as the addition of the error increment of the optimal operator (i.e., increase in the error of the optimal operator by the reduction of the search space) and the estimation error in the constrained space. A constraint is beneficial when the constraint estimation error decreases (i.e., in relation to the estimation error in the full space) more than the error increment of the optimal operator. The known constraints are heuristics proposed by experts.

3. Window design by conditional entropy minimization

Information theory has its roots in Claude Shannon's works [5] and has been successfully applied in a multitude of situations. In particular, mutual information is a useful measure to characterize the stochastic dependence among discrete random variables [6, 7, 8]. It may be applied to feature selection problems in order to help identifying good subspaces to perform pattern recognition [9, 10]. For instance, Lewis [11] explored the mutual information concept for text categorization while Bonnlander and Weigend used similar ideas for dimensionality reduction in neural networks [12]. Additional works that may also be of interest include [13, 14]. An important concept related to the mutual information is the mean conditional entropy, which is explored in our approach.

3.1. Feature selection: problem formulation

Given a set of training samples T where each sample is a pair $(\mathbf{x}, y)$, a function ψ from $\{0,1\}^n$ to $\{0,1\}$, called a *binary classifier*, may be designed.

Feature selection is a procedure to select a subset $\mathcal{Z}$ of $C = \{1, 2, ..., n\}$ such that $\mathbf{X}_\mathcal{Z}$ be a good subspace of $\mathbf{X}$ to design a binary classifier ψ from $\{0,1\}^{|\mathcal{Z}|}$ to $\{0,1\}$.

The choice of $\mathcal{Z}$ creates a constraint search space for designing the binary classifier ψ. $\mathcal{Z}$ is a good subspace, if the classifier designed in $\mathcal{Z}$ from a training sample T has smaller error than the one designed in the full space from the same training sample T.

Clearly, there are too many possible subsets $\mathcal{Z}$ of C and two different aspects involves searching for most suitable ones: a criterion function and a search algorithm (often based on heuristics in order to cope with the combinatorial explosion) [15].

Next section explains how we explore the mean conditional entropy as a criterion function to distinguish between good and bad feature subsets.

3.2. Mean conditional entropy as criterion function

Let X be a random variable and P be its probability distribution. The *entropy* of X is defined as:

$$H(X) = -\sum_{x \in X} P(x) log_2 P(x) \qquad (1)$$

Similar definitions hold for random vectors $\mathbf{X}$. The motivation for using the entropy as a criterion function for feature selection is due to its capabilities of measuring the amount of information about labels (Y) that may be extracted from the features ($\mathbf{X}$). The more informative is $\mathbf{X}$ w.r.t. Y, the smaller is $H(Y|\mathbf{X})$. The basic idea behind this method is to minimize the conditional entropy of Y w.r.t the instances $\mathbf{x}_{\mathcal{Z}_\mathbf{i}}$ of $\mathbf{X}_\mathcal{Z}$.

This is illustrated in Figure 1 which shows the probability $P(Y|\mathbf{X} = \mathbf{x}_{\mathcal{Z}_\mathbf{i}})$ for some instance $\mathbf{x}_{\mathcal{Z}_\mathbf{i}}$. For the hypothetical situation 1(a) the conditional entropy $H(Y|\mathbf{x}_{\mathcal{Z}_\mathbf{i}})$ is small since $P(Y|X_\mathcal{Z} = \mathbf{x}_{\mathcal{Z}_\mathbf{i}})$ is concentrated around a peak. In other words, Y may be predicted from $\mathbf{X}_\mathcal{Z}$ with good confidence. On the other hand, in the case depicted by Figure 1(b) the entropy $H(Y|\mathbf{x}_{\mathcal{Z}_\mathbf{i}})$ should be large because it is obviously difficult to predict the value to be assumed by Y when $X_\mathcal{Z} = \mathbf{x}_{\mathcal{Z}_\mathbf{i}}$ is observed.

In order to define a criterion function, we average the conditional entropy for all possible instances $\mathbf{x}_{\mathcal{Z}_\mathbf{i}}$ weighted by the number of occurrences of each instance in the training set. We have estimated the *mean conditional entropy of Y given $\mathbf{X}_\mathcal{Z}$*, as shown in Equation 2:

$$\hat{E}[H(Y|\mathbf{X}_\mathcal{Z})] = \sum_{i=1}^{2^{|\mathcal{Z}|}} \frac{\hat{H}(Y|\mathbf{X}_{\mathcal{Z}_\mathbf{i}}) \cdot (o_i + \alpha)}{\alpha 2^{|\mathcal{Z}|} + t} \qquad (2)$$

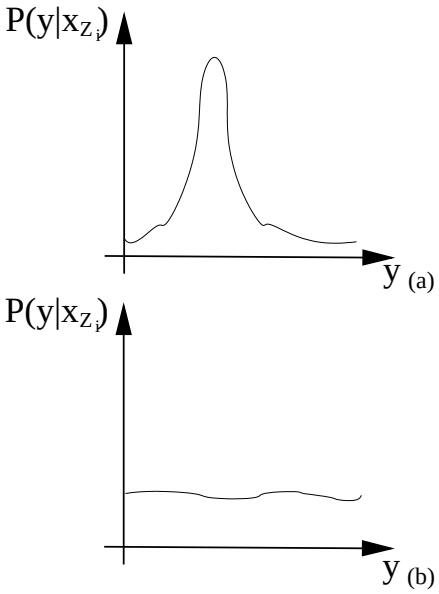


Figure 1. (a) low entropy; (b) high entropy

where $\hat{H}(Y|\mathbf{X}_{\mathcal{Z}_i})$ is the entropy of the estimated conditional probability $\hat{P}(Y|\mathbf{X}_{\mathcal{Z}_i})$, o_i is the number occurrences of $\mathbf{X}_{\mathcal{Z}_i}$ in the training set, t is the total number of training samples, $2^{|\mathcal{Z}|}$ is the number of possible instances of $\mathbf{X}_{\mathcal{Z}}$ and α is a weight factor used to model $P(\mathbf{X}_{\mathcal{Z}})$ and so circumvent problems when some instances of $\mathbf{X}_{\mathcal{Z}_i}$ are not observed in the training data. When $\mathbf{X}_{\mathcal{Z}_i}$ is not observed, $\hat{P}(Y|\mathbf{X}_{\mathcal{Z}_i})$ is considered uniform and $\hat{P}(\mathbf{X}_{\mathcal{Z}_i}) = \alpha/(\alpha 2^{|\mathcal{Z}|} + t)$. Thus, the mass of non observed shapes is $(2^{|\mathcal{Z}|} - N)\alpha/(\alpha 2^{|\mathcal{Z}|} + t)$, where N is the number of observed shapes. With these considerations, Formula 3 simplifies to

$$\hat{E}[H(Y|\mathbf{X}_{\mathcal{Z}})] = \frac{(2^{|\mathcal{Z}|} - N)\alpha}{\alpha 2^{|\mathcal{Z}|} + t} + \sum_{i=1}^{N} \frac{\hat{H}(Y|\mathbf{X}_{\mathcal{Z}_i}) \cdot (o_i + \alpha)}{\alpha 2^{|\mathcal{Z}|} + t} \tag{3}$$

, since $H([0.5, 0.5]) = 1$.

Usually N is small relatively to $2^{|\mathcal{Z}|}$ and so $\hat{E}[H(Y|\mathbf{X}_{\mathcal{Z}})]$ is easy to compute. We have adopted $\alpha = 1$ in our experiments.

3.3. Minimal components determination

Minimizing the mean conditional entropy $E[H(Y|\mathbf{X})]$ is equivalent to maximizing the mean mutual information. Therefore, feature selection may be defined as an optimization problem where we search for $\mathcal{Z}^* \subseteq C$ such that:

$$\mathcal{Z}^* : H(Y|\mathbf{X}_{\mathcal{Z}^*}) = min_{\mathcal{Z} \subseteq C}\{\hat{E}[H(Y|\mathbf{X}_{\mathcal{Z}})]\} \tag{4}$$

with $C = \{1, 2, ..., n\}$.

Usually, it is impossible to evaluate all subsets of C in order to calculate the optimum defined by Equation 4. Instead, a heuristic search algorithm is applied. There are many of such algorithms proposed in the literature and the author should refer to [16] for a comparative review. In this work we have explored the Sequential Forward Search approach (SFS), as explained below (we have also explored the SFFS algorithm, though these results are not shown here). The basic principle of this algorithm is very simple: given a subspace $\mathcal{Z}$ of components which is supposed to be the best set found until that step, look for a variable X_j such that $\mathcal{Z} \cup \{j\}$ is the best set among $\mathcal{Z} \cup \{i\}, 1 \leq i \leq n$ and $\mathcal{Z}$ itself. The algorithm stops and returns $\mathcal{Z}$ when there is no longer j such that $\mathcal{Z} \cup \{j\}$ is better than $\mathcal{Z}$. In our approach, ¨being better than¨ means presenting lower entropy as defined by Equation 2.

3.4. Minimal window determination

Section 2.2 explained the importance of selecting a suitable shape for the window W in the design of a W-operator. In order to explore the entropy concept described in the previous section to design W, the window positions that form its shape are taken as variables that compose a random vector $\mathbf{X}$. Hence, designing a W can be viewed as a feature selection problem that uses $\hat{E}[H(Y|\mathbf{X}_{\mathcal{Z}})]$ as a criterion function. Figure 2 illustrates this concept by showing two possible shapes for W. In that figure, the selected variables $\mathbf{X}_{\mathcal{Z}}$ are indicated as black cells, thus 5 variables have been selected in 2(a) while 13 have been selected in 2(b).

The estimation of Equation 2 is carried out from a training set consisting of input-output image pairs.

4. Experimental Results

Dimensionality reduction is intimately related to the so called U-curve problem where classification error is plotted against feature vector dimension (for an *a priori* fixed training set dimension). This plot leads to a U-shaped curve, implying that an increasing dimension initially improves the classifier performance. Nevertheless, this process reaches a minimum after which estimation errors degrades the classifier performance [15]. As it would be expected, the mean conditional entropy reflects this fact, thus corroborating its use for feature selection. Controlled experiments with a priori known distributions have been carried out and Figure 3 shows the plot of $\hat{E}[H(Y|\mathbf{X}_{\mathcal{Z}})] \times$ *dimensionality*, for a fixed amount of training data, illustrating the U-curve effect for the entropy.

The application of the mean conditional entropy to design W has been carefully analyzed in several binary image filtering experiments. Salt-and-pepper noise has been added to binary images and W-operators for filtering the

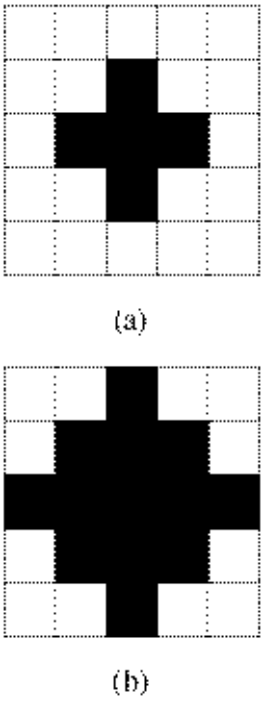


Figure 2. Windows with (a) 5 variables and (b) 13 variables

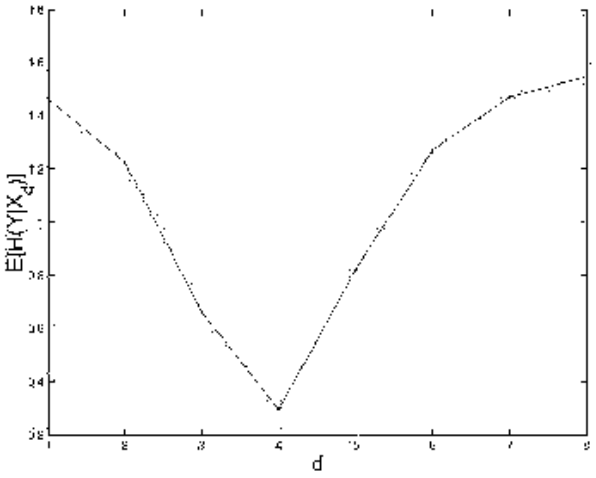


Figure 3. Plot of $\hat{E}[H(Y|\mathbf{X}_{\mathcal{Z}})]\times$ dimensionality

noisy images have been generated using the aforementioned methodology. Figure 4 presents an example of an original image and its noisy version.

In the first experiment we have analyzed the mean absolute error (MAE) for the image in Figure 4. We have chosen a simple image with added noise so that we could control all parameters in order to analyze MAE as a function of two variables: (1) the size of the training set used to select

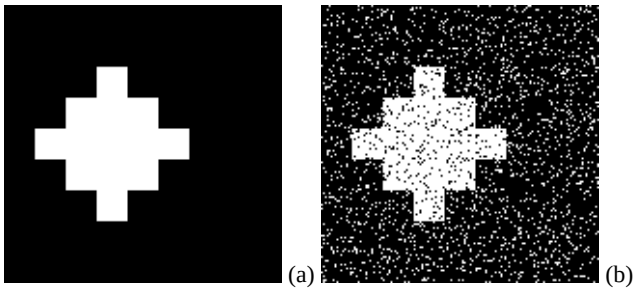


Figure 4. (a) ideal image; (b) observed image with salt and pepper noise (10%).

the best window for the W-operator and (2) the size of the training set used to design the W-operator. Four increasing sizes have been used: 1/4 of an image, 1/2 of an image, 1 image and 3 images (actually, pairs of images, since the process requires pairs of noisy-ideal images). Each experiment consisted of selecting a window and training the W-operator with training sets of increasing sizes and then applying to 10 noisy images generated by the same noise model. Table 1 resumes the MAE results (MAE_{min} means the minimum MAE for all 10 tries, while MAE_{avg} indicates the average from these 10 tries). It is important to note that using larger training sets to select the window and to train the W-operator improves the filter performance, as expected.

Table 1. Mean absolute error table for the filtering experiment: MAE_{min} means the minimum MAE for all 10 tries, while MAE_{avg} indicates the average from these 10 tries; TSSS: training set size used to select the window for the W-operator; TSSW: training set size used to design the W-operator.

	TSSS				
	1/4	1/2	1	3	TSSW
MAE_{min}	0.0056	0.0080	0.0057	0.0099	1/4
MAE_{avg}	0.0070	0.0087	0.0071	0.0118	
MAE_{min}	0.0046	0.0048	0.0036	0.0050	1/2
MAE_{avg}	0.0055	0.0058	0.0052	0.0067	
MAE_{min}	0.0036	0.0031	0.0029	0.0040	1
MAE_{avg}	0.0050	0.0038	0.0036	0.0045	
MAE_{min}	0.0038	0.0024	0.0027	0.0020	3
MAE_{avg}	0.0047	0.0035	0.0035	0.0030	

It is important to analyze the selected window for these experiments, which is illustrated in Figure 5. Each image in this figure is associated to the series of 10 window selection runs for a respective training set size. An accumulator array has been created where each cell corresponds to a variable in the window. For each run, the selected variables were incremented in the accumulator array (this is analogous to

the voting scheme of the traditional Hough transform [17]). The images in Figure 5 show the corresponding accumulator arrays with the gray-level coding the number of votes each cell received (darker gray-levels correspond to more voted cells). As it can be noted, the well voted cells form a window with a cross pattern (i.e. ¨+¨), which is the best window to design a W-operator for the considered type of image and noise. Note also that increasing the amount of training data improves the definition of the correct window. Besides, the subsets of more voted pixels in an increasing sequence of sample sizes form a sequence of included subsets. The other filtering experiments (explained below) lead to selected windows with its shape adapted to each particular case. This effect is used to define a more suitable window for designing the W-operator as follows: in the i-th experiment, the set of d_i variables, corresponding to the minimum of the U-shaped mean conditional entropy curve (Figure 3), is selected. The final number of variables is the (floor) average of all d_i, denoted as $\bar{d}$, i.e. the $\bar{d}$ most voted variables are selected.

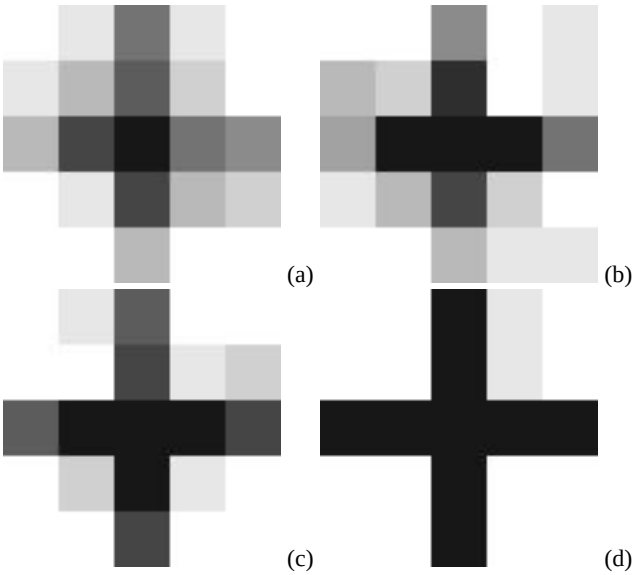


Figure 5. Accumulator arrays after 10 iterations. (a) 1/4 selection image; (b) 1/2 selection image; (c) 1 selection image; (d) 3 selection images

Salt-and-pepper noise is commonly cleaned in image processing by median filtering and we have compared this traditional technique to the proposed approach. Table 2 shows the MAE results for two increasing size median filters. Comparing these results with those presented in Table 1 indicates the superior performance our approach.

Some image filtering results are shown in Figures 6, 7 and 8. They show the original image, its noisy version and results produced by median filter and W-operator. Is

Table 2. MAE results for median filtering

	median 3×3	median 5×5
MAE_{min}	0.0044	0.0080
MAE_{avg}	0.0059	0.0097

is worth noting that median filtering severely affects thin patterns in the image, much more than the W-operator (see Figure 8 where $MAE = 0.0053$ for Figure 8(a) and $MAE = 0.0006$ for Figure 8(b)).

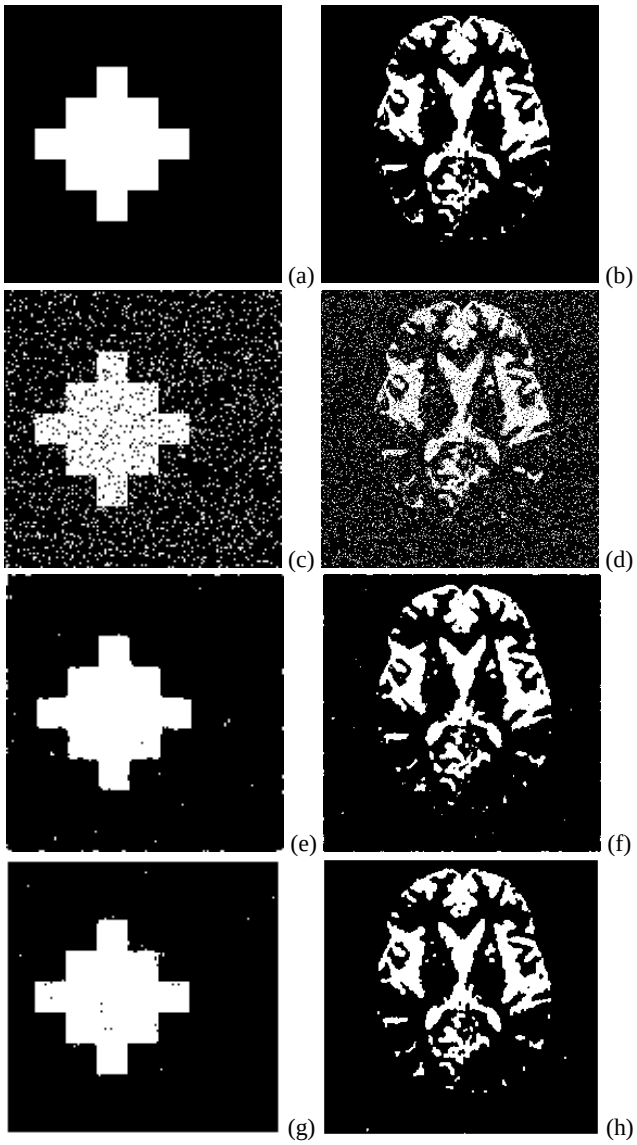
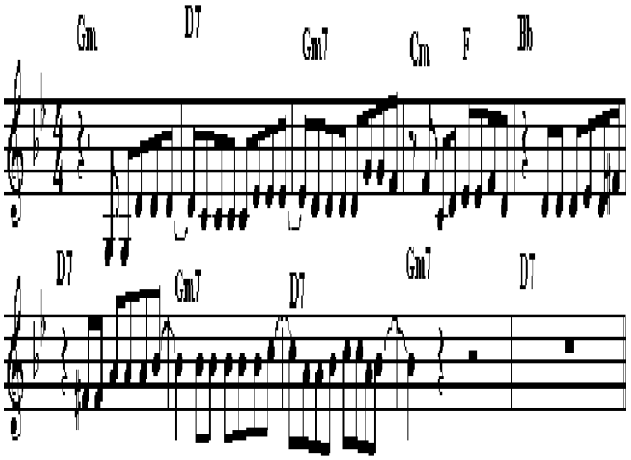
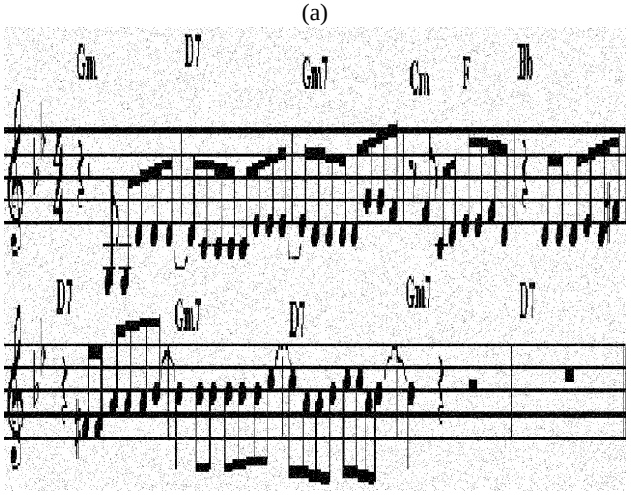


Figure 6. (a-b) Ideal images; (c-d) images with 10% salt and pepper noise; (e-f) final result by applying 3×3 median filter; (g-h) final result obtained by W-operators.

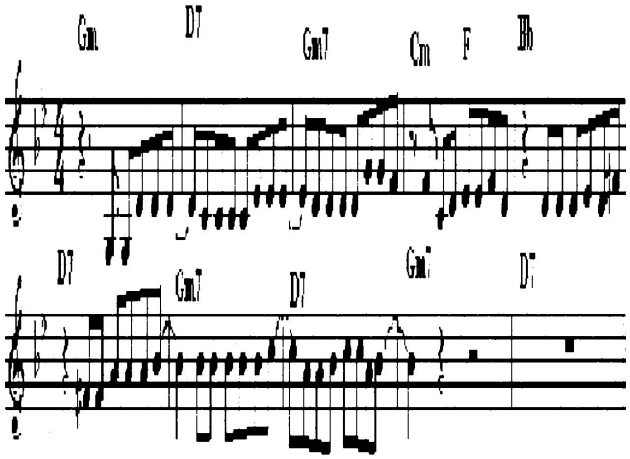


(a)

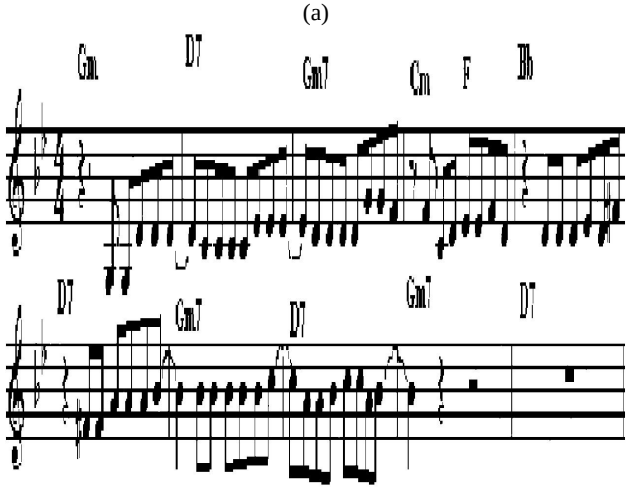


(b)

Figure 7. (a) ideal image; (b) image with 3% salt and pepper noise



(a)



(b)

Figure 8. (a) final result by applying 3×3 median filter; (b) final result by applying our method using 10 iteractions with 1 selection image each to obtain the accumulator array and 1 training image

5. Concluding Remarks

This paper presents a technique that gives a minimal window W for the estimation of a W-operator from training data. For applying this technique, it is necessary that the conditional probabilities of the pattern recognition problem studied have mass concentrated in one class, what is a reasonable hypothesis when the problem has a good solution. Experimental results corroborating our approach have been presented.

The problem of designing an window for a W-operator was approached recently by Dougherty et al [18] through the Coefficient of Determination (COD) concept. The COD is a kind of relative error of the estimated operator, that is used to measure the quality of the windows. Thus, it depends on an operator estimation algorithm. This approach may have two drawbacks: 1 - it does not separate the joint distribution estimation of the operator estimation; 2 - it may be very expensive computationally, since the operator design algorithm should be applied to every window in the search space. The approach proposed here sees the window design problem by a different perspective: to choose the window that gives the best estimation for the joint distribution under the hypothesis that the pattern recognition problem has a good solution. Thus, it does not have the drawbacks of the COD approach.

For the estimation of the mean conditional entropy, it is required the estimation of the conditional probabilities, $P(Y|\mathbf{X})$, and of the prior distribution, $P(\mathbf{X})$. The conditional probabilities $P(Y|\mathbf{X})$ are estimated based on simple counting of the observed classifications (i.e., the pixel observation in the ideal image) of a given shape. The entropy

for $\mathbf{X}$ is computed from the estimated distribution $\hat{P}(Y|X)$. Whenever there is no observations of $\mathbf{X}$, the distribution $P(Y|\mathbf{X})$ is considered uniform. Better estimations of $P(\mathbf{X})$ require a model for representing the distribution of $\mathbf{X}$, but modeling is very specific for each family of images. In this paper, we supposed that the probability is distributed uniformly outside the range of observed shapes and the percentage of the observed shapes probability mass is a parameter of the model. In the examples presented, this parameter was fixed in 1 (i.e. α in Equation 2). An interesting improvement of this research would be estimating α from the training data.

In the next steps of this research, we will also compare some heuristics to explore the search space and study the possibility of creating a kind of branch and bound algorithm for the full search.

It is important to note that, though this paper only presented binary image filtering experiments, the proposed technique is general and may be applied in several image processing problems such as texture segmentation, gray-level and color image processing, document analysis, etc. Some initial experiments for gray-level texture segmentation have lead to encouraging results and further advances on this will be reported in due time.

Acknowledgements

The authors are grateful to FAPESP (99/12765-2, 01/09401-0 and 02/04611-0), CNPq (300722/98-2, 52.1097/01-0 and 468413/00-6) and CAPES for financial support.

6. References

[1] J. Barrera, R. Terada, R. Hirata-Jr., and N. S. T. Hirata. Automatic programming of morphological machines by pac learning. *Fundamenta Informaticae*, pages 229–258, 2000.

[2] E. R. Dougherty and J. Barrera. Pattern recognition theory in nonlinear signal processing. *J. Math. Imaging Vis.*, 16(3):181–197, 2002.

[3] E. R. Dougherty, J. Barrera, G. Mozelle, S. Kim, and M. Brun. Multiresolution analysis for optimal binary filters. *J. Math. Imaging Vis.*, 14(1):53–72, 2001.

[4] P. Pudil, J. Novovicová, and J. Kittler. Floating search methods in feature selection. *Pattern Recognition Letters*, 15:1119–1125, 1994.

[5] C. E. Shannon. A mathematical theory of communication. *Bell System Technical Journal*, 27:379–423, 623–656, July, October 1948.

[6] T. M. Cover and J. A. Thomas. Elements of information theory. In *Wiley Series in Telecommunications*. John Wiley & Sons, New York, NY, USA, 1991.

[7] S. Kullback. *Information Theory and Statistics*. Dover, 1968.

[8] E. S. Soofi. Principal information theoretic approaches. *Journal of the American Statistical Association*, 95:1349–1353, 2000.

[9] R. O. Duda, P. E. Hart, and D. Stork. *Pattern Classification*. John Wiley & Sons, NY, 2000.

[10] M. A. Hall and L. A. Smith. Feature selection for machine learning: Comparing a correlation-based filter approach to the wrapper. In *Proc. FLAIRS Conference*, pages 235–239. AAAI Press, 1999.

[11] D. D. Lewis. Feature selection and feature extraction for text categorization. In *Proceedings of Speech and Natural Language Workshop*, pages 212–217, San Mateo, California, 1992. Morgan Kaufmann.

[12] B. V. Bonnlander and A. S. Weigend. Selecting input variables using mutual information and nonparametric density estimation. In *Proc. of the 1994 Int. Symp. on Artificial Neural Networks*, pages 42–50, Tainan, Taiwan, 1994.

[13] P. Viola and W. M. Wells, III. Alignment by maximization of mutual information. *Int. J. Comput. Vision*, 24(2):137–154, 1997.

[14] M. Zaffalon and M. Hutter. Robust feature selection by mutual information distributions. In *18th International Conference on Uncertainty in Artificial Intelligence (UAI)*, pages 577–584, 2002.

[15] T. E. Campos, I. Bloch, and R. M. Cesar-Jr. Feature selection based on fuzzy distances between clusters: First results on simulated data. In S. Singh, N. Murshed, and W. Kropatsch, editors, *Proc. ICAPR'2001 - International Conference on Advances in Pattern Recognition*, volume 2013 of *Lecture Notes in Computer Science, Springer-Verlag Press*, pages 186–195, Rio de Janeiro, Brasil, 2001.

[16] A. Jain and D. Zongker. Feature selection - evaluation, application, and small sample performance. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 19(2):153–158, 1997.

[17] L. F. Costa and R.M. Cesar-Jr. *Shape Analysis and Classification: Theory and Practice*. CRC Press, 2001.

[18] E. R. Dougherty, S. Kim, and Y. Chen. Coefficient of determination in nonlinear signal processing. volume 80, pages 2219–2235, 2000.