Método de Análise de Características para Teste de Intersecção entre Classes

Leimar S. S. Mafort, Aline R. Gesualdi, Márcio P. de Albuquerque, Eduardo A. B. da Silva, Eugênio S. Caner, Marcelo P. de Albuquerque

***Resumo*— Neste artigo apresentamos um método de análise da qualidade das características utilizadas para descrever padrões observados em sinais genéricos, com ênfase especial em padrões observados em imagens. A utilização do método apresentado torna mais simples a eliminação de características que não contribuem significativamente para o sistema de classificação, permitindo também a identificação de onde encontram-se as maiores chances de classificações equivocadas e entre quais estão as maiores vulnerabilidades. São apresentados testes realizados em imagens de letras e algarismos extraídos de placas de automóveis. Estes demonstram que o método tem bastante utilidade para o projetista de sistemas de classificação, indicando quais características devem ser consideradas. Além disso, ele provê uma base de comparação para que parâmetros de pré-processamento possam ser alterados a fim de se chegar a valores ótimos de cada um deles, ou seja, valores que retornem as maiores taxas de acerto.**

***Palavras-Chave*— Processamento de imagens, sistemas de classificação, análise de características.**

***Abstract*— In this article we present a feature quality analysis method used to describe patterns observed in generic signals, with special emphasis in images. The presented method indicates relevant features needed by a classification system. In addition, it sorts these features by classification efficiency revealing the most important ones for the system. Tests carried out with alphanumeric car plate images are presented. These demonstrate that the method is useful for a classification system designer, indicating what is relevant to be considered as a standard feature. Moreover, it provides an attribute comparison key that can be used in a pre-processing step in order to make an optimum selection of the input to the classifier.**

***Keywords*— Image procesing, classification system, characteristic analysis.**

I. Introdução

Técnicas de reconhecimento de padrões se mostram importantes quando nos interessa separar e classificar automaticamente estruturas encontradas em sinais. Estas podem ser encontradas em diversas formas de sinais, como, por exemplo, sinais ultra-sônicos, de som ou de imagem [1]. Um exemplo típico no contexto desse último exemplo repousa na visão computacional.

Apesar de o método proposto ser genérico, neste artigo ele será aplicado no reconhecimento de caracteres encontrados em placas de automóveis.

Leimar S. S. Mafort, Márcio P. de Albuquerque, Eugênio S. Caner e Marcelo P. de Albuquerque¸ Coordenação de Atividades Técnicas, Centro Brasileiro de Pesquisas Físicas, Rio de Janeiro, Brasil, E-mails: mafort@cbpf.br, mpa@cbpf.br, marcelo@cbpf.br. Aline R. Gesualdi e Eduardo A. B. da Silva, Laboratório de Processamento de Sinais, Universidade Federal do Rio de Janeiro (COPPE/UFRJ), Rio de Janeiro, Brasil, E-mails: aline@lps.ufrj.br, eduardo@lps.ufrj.br.

A. Motivação - o porquê da seleção de um grupo de características

Por hipótese, há um banco de sinais – imagens com *l* pixels de largura por *h* pixels de altura. Também por hipótese, essas imagens podem ser classificadas em *n* grupos diferentes, ou seja, existem *n* padrões distintos.

A primeira observação a se fazer está na grande dimensionalidade dos sinais. Sendo $l = h = 128$, as imagens (128×128) podem ser representadas como pontos no espaço *n*-dimensional com $n = 16384$; os eixos desse espaço representam o nível de cinza de cada pixel, que neste exemplo poderia ser representado por 8 bits. Pode-se concluir, dessa forma, que o número máximo de diferentes imagens é $2^{8\times128\times128} \approx 10^{39500}$. Como existem apenas algumas centenas de imagens no banco de imagens, podemos concluir que imagens ainda não vistas serão classificadas pelo sistema.

Um número tão grande de variáveis de entrada pode levar a problemas sérios em um sistema de reconhecimento de padrões [2]. Se, por exemplo, um classificador neuronal estivesse sendo utilizado e se cada pixel da imagem fosse tomado como entrada, haveria 16385 pesos a serem ajustados para cada neurônio da primeira camada interna. Isto impõe que o conjunto de treinamento seja muito grande para garantir que os pesos estejam bem determinados, além de exigir muitos recursos computacionais para encontrar uma função mínima adequada [3].

Uma forma de contornar esse problema é combinar algumas variáveis de entrada para formar um número menor de variáveis chamadas características. A essa etapa no processo de classificação é dado o nome de *extração de características*. A tarefa de projetar um extrator de características não apresenta nem na ciência da computação nem na psicologia uma solução ótima, sendo neste sentido uma *ciência intuitiva* e *ad hoc* [4].

Conhecida a aplicação e o tipo de sinal, pode-se determinar um conjunto de características para descrever os padrões. Em se tratando de imagens, é comum usar a *área*, o *perímetro*, o *centro de massa* ou uma combinação de operações entre linhas e colunas da imagem. Não há limite para o número de características que podem ser definidas. O método proposto neste trabalho analisa algumas características e elege, dentre as inicialmente definidas, quais são as mais adequadas para a separação dos padrões no problema específico.

II. O método aplicado

O método de análise de características proposto será aplicado a um exemplo prático: reconhecimento de algarismo

Caractere	Número de amostras
“0”	37
“1”	36
“2”	44
“3”	44
“4”	31
“5”	28
“6”	31
“7”	37
“8”	42
“9”	35

TABELA I
NÚMERO DE AMOSTRAS DE CADA CARACTERE NO BANCO DE IMAGENS.

extraídos de placas de automóveis. As imagens do banco de sinais representam cada uma um caractere. As imagens foram segmentadas pelo processo de segmentação descrito em [5] resultando em imagens binárias (1 bit). O banco de sinais contém aproximadamente 40 amostras de cada um dos 10 padrões, como apresenta a Tabela I.

A. Definindo um grupo de características

Conforme comentado anteriormente, o número de características que pode ser definido para um problema é virtualmente ilimitado. Conhecida a forma de sinal e o problema específico, foram consideradas as seguintes características [6]:

1) **Área**: número de pixeis do objeto, considerando uma imagem binária.
2) **Centróide X**: coordenada X do centro de massa da imagem.
3) **Centróide Y**: coordenada Y do centro de massa da imagem.
4) **Comprimento do eixo maior**: escalar que representa o comprimento em pixeis do eixo maior da elipse que circunscreve o objeto.
5) **Comprimento do eixo menor**: escalar que representa o comprimento em pixeis do eixo menor da elipse que circunscreve o objeto.
6) **Excentricidade**: escalar que representa a excentricidade da elipse que envolve o objeto. A excentricidade é definida como a razão entre a distância entre os focos e o eixo maior da elipse.
7) **Orientação**: escalar que representa o ângulo em graus entre o eixo x e o eixo maior da elipse que circunscreve o objeto.
8) **Área Preenchida**: o número de pixeis contidos na *Filled Image* (uma imagem binária do mesmo tamanho do *Bounding Box* - o menor retângulo que contém o objeto).
9) **Área Convexa**: o número de pixeis contidos na *Convex Image* - menor polígono convexo que contém o objeto.
10) **Número de Euler**: escalar igual ao número de objetos na imagem menos o número de buracos nestes objetos.
11) **Diâmetro Equivalente**: o diâmetro de um círculo de mesma área que o objeto.
12) **“Solidez”**: razão entre a Área e a *Convex Area* (área da *Convex Image* - uma imagem que representa o menor polígono convexo que contém o objeto).
13) **“Extensão”**: escalar calculado a partir da divisão da *Área* pela área do *Bounding Box* (o menor retângulo que contém o objeto).

B. Análise estatística das características

Para cada uma das amostras do banco de imagens calculamos um vetor de características. Esse vetor tem tamanho *x* para o caso de considerarmos as *x* características. Haverá *n* grupos diferentes de vetores (1 para cada um nos *n* padrões). No caso específico, tem-se que $x = 13$ características e $n = 10$ algarismos.

Cada um dos padrões (algarismos encontrados nas placas dos automóveis) será representado matematicamente por uma linha das duas matrizes Υ e Σ. A primeira representa as médias das características e a segunda seus respectivos desvios padrões. O valor de $\mu_{i,j}$ representa a média da j-ésima característica para o i-ésimo padrão. O valor de $\sigma_{i,j}$ representa o desvio-padrão da j-ésima característica para o i-ésimo padrão.

$$\Upsilon = \begin{bmatrix} \mu_{1,1} & \mu_{1,2} & \cdots & \mu_{1,x} \\ \mu_{2,1} & \mu_{2,2} & \cdots & \mu_{2,x} \\ \vdots & \vdots & \ddots & \vdots \\ \mu_{n,1} & \mu_{n,2} & \cdots & \mu_{n,x} \end{bmatrix} \quad (1)$$

$$\Sigma = \begin{bmatrix} \sigma_{1,1} & \sigma_{1,2} & \cdots & \sigma_{1,x} \\ \sigma_{2,1} & \sigma_{2,2} & \cdots & \sigma_{2,x} \\ \vdots & \vdots & \ddots & \vdots \\ \sigma_{n,1} & \sigma_{n,2} & \cdots & \sigma_{n,x} \end{bmatrix} \quad (2)$$

De modo a estabelecer a relação de relevância entre as características escolhidas para a classificação de padrões, o método proposto calcula a probabilidade de intersecção entre as classes para uma ou mais característica de acordo com a equação:

$$m_{p_i,p_s,c_e} = P([L_i, L_s])_{p_s,c_e} * P([L_i, L_s])_{p_i,c_e} \quad (3)$$

onde p_i e p_s são os padrões escolhidos, c_e a característica escolhida, L_i o limite inferior do intervalo de intersecção e L_s o limite superior. A Figura 1 apresenta um exemplo entre as classes ‘0’ e ‘1’ do banco de imagens considerando a Área como característica. Notar que as distribuições foram consideradas Gaussianas.

A probabilibade de intersecção m_{p_i,p_s,c_e} proporciona a relação entre as classes para uma determinada característica, podendo variar no intervalo $[0, 1]$. Para cada uma das x características podemos criar uma matriz quadrada M de ordem igual ao número de classes n, de acordo com a equação 4. A matriz M tem como elementos os indicadores do grau de separação entre as classes considerando cada uma delas isoladamente.

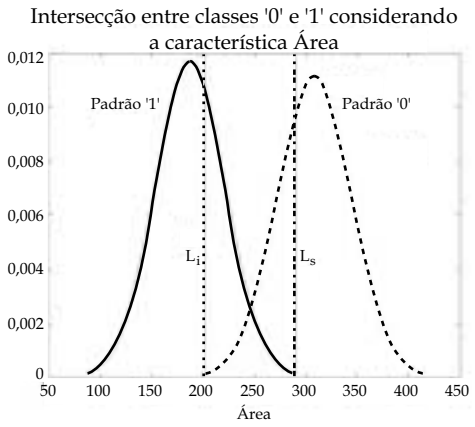

Fig. 1. Intersecção entre as classes geradas para os padrões '0' e '1', levando em consideração a característica área. L_i é o limite inferior do intervalo de intersecção e L_s, o superior.

$$M([1:n],[1:n],x) = \begin{bmatrix} 1 & m_{1,2,x} & \cdots & m_{1,n,x} \\ m_{2,1,x} & 1 & \cdots & m_{2,n,x} \\ \vdots & \vdots & \ddots & \vdots \\ m_{n,1,x} & m_{n,2,x} & \cdots & 1 \end{bmatrix} \tag{4}$$

Processos de clasificação que levam em conta apenas uma característica são extremamente raros. Dentro de limites razoáveis e não incorrendo nos erros que o fenômeno do "*curse of dimensinality*"[7] pode nos trazer, tomamos como hipótese a afirmação de que a adição de uma nova característica faria com que as classes ficassem mais separadas. Assim propomos o cálculo da matriz F, obtida a partir da equação 5.

$$F = \begin{bmatrix} 1 & \prod_{i=x_i}^{x_f} m_{1,2,i} & \cdots & \prod_{i=x_i}^{x_f} m_{1,n,i} \\ \prod_{i=x_i}^{x_f} m_{2,1,i} & 1 & \cdots & \prod_{i=x_i}^{x_f} m_{2,n,i} \\ \vdots & \vdots & \ddots & \vdots \\ \prod_{i=x_i}^{x_f} m_{n,1,i} & \prod_{i=x_i}^{x_f} m_{n,2,i} & \cdots & 1 \end{bmatrix} \tag{5}$$

onde x_i e x_f são as características inicial e final, respectivamente. Podemos combinar as características de modo a obtermos $\frac{13!}{(13-x)!x!}$ matrizes F diferentes. Para determinarmos quais características são mais relevantes para o sistema de classificação, basta escolher qual matriz F apresenta as classes de forma mais espaçada, ou seja, cujos elementos estejam mais próximos de zero. Com o objetivo de determinar tal proximidade a zero, utilizamos o coeficiente Λ definido pela equação 6. Aquela que retornar um menor valor para esse coeficiente pode ser considerada como a que contém as características que melhor separam as classes.

$$\Lambda = \sum_{i=0}^{n} \sum_{j=0}^{n} F(i,j) \qquad \forall \ i \neq j \tag{6}$$

Observando a matriz F e o coeficiente Λ podemos extrair a informação de quais classes estão mais próximas e nos concentrar especificamente em criarmos características que consigam separar essas classes. Nesse sentido o método auxilia o projetista de sistemas de classificação, dirigindo os maiores esforços para os setores que precisem ser melhorados.

III. Problema de características correlacionadas

É importante notar que este processo de separação de classes pode mascarar uma melhora que na verdade não existe de fato. Notar que a matriz F, a menos dos elementos da diagonal principal, que são iguais a 1, é o produto das matrizes M_x para cada característica x. Assim, se na formação de F, multiplicamos duas matrizes M_{x_i} e M_{x_j} tais que as características x_i e x_j possuam alta correlação entre si, tenderemos a ter valores pequenos para os elementos em F, e, em conseqüência, valores pequenos de Λ. Entretanto, isto certamente não significará uma melhora no processo de separação entre classes. Isto pode ser corrigido, levando em consideração a correlação entre as matrizes M no cálculo de Λ, isto é, duas características com alta correlação não devem entrar ambas no cálculo de Λ.

Desta forma, levamos em conta esse problema para o cálculo de Λ, testando a correlação entre as características i e j (veja Tabela II).

IV. Critério "força bruta" de avaliação

É necessário estabelecermos um critério que mostre que o coeficiente Λ está relacionado com a taxa de acerto do sistema de classificação. Utilizamos para tal um sistema de classificação com x características. Para cada um desses sistemas, calculamos o índice de acerto médio e comparamos esses resultados com os valores de Λ. O sistema de classificação consiste em um classificador por distâncias Mahalanobis [8] definido por

$$D_i^2(x) = (x - \mu_i) S_t^{-1} (x - \mu_i) \tag{7}$$

onde D_i é a distância quadrática para cada elemento i; S_i representa a matrix de desvios padrões; μ_i é o vetor de médias das características e x é a característica de teste.

Cada classe é representada pelas médias e desvios padrões das x características que entram no sistema. Calculando-se a distância do ponto de teste até todas as classes, saberemos que ele pertencerá à classe que menos diste dele. Se utilizássemos $k = 4$ características teríamos 715 sistemas de classificação: todas as combinações 4 a 4 das 13 características utilizadas; se $k = 5$, teríamos 1287. Seria inviável mostrarmos todos esses resultados - os valores das taxas de acerto e Λ para todas as combinações que valores de k entre 1 e 12 gerariam. Por esse motivo, mostramos na Tabela III o coeficiente de correlação entre a taxa de acertos e o Λ obtido para sistemas com $k = 1$ até 12. Uma vez que o valor de Λ é inversamente proporcional à taxa média de acertos, podemos concluir que quão mais próximo o coeficiente de correlação estiver de -1, maior é a relação entre os valores de acerto e Λ, visto que à medida que a taxa de acerto aumenta, mais próximo de zero fica o valor de Λ.

	1	2	3	4	5	6	7	8	9	10	11	12	13
1	1,00	0,65	0,22	0,09	0,94	-0,81	-0,11	0,79	0,92	-0,49	1,00	-0,65	0,02
2	0,65	1,00	0,31	0,06	0,78	-0,60	-0,31	0,27	0,77	0,02	0,68	-0,82	-0,62
3	0,22	0,31	1,00	-0,30	0,23	-0,36	0,01	0,15	0,15	-0,06	0,23	-0,01	-0,20
4	0,09	0,06	-0,30	1,00	0,02	0,41	0,05	-0,20	0,36	0,50	0,08	-0,53	0,08
5	0,94	0,78	0,23	0,02	1,00	-0,89	-0,08	0,74	0,91	-0,36	0,95	-0,73	-0,18
6	-0,81	-0,60	-0,36	0,41	-0,89	1,00	-0,02	-0,79	-0,65	0,56	-0,81	0,37	0,11
7	-0,11	-0,31	0,01	0,05	-0,08	-0,02	1,00	-0,13	-0,12	0,25	-0,15	0,20	0,55
8	0,79	0,27	0,15	-0,20	0,74	-0,79	-0,13	1,00	0,56	-0,82	0,77	-0,18	0,27
9	0,92	0,77	0,15	0,36	0,91	-0,65	-0,12	0,56	1,00	-0,15	0,93	-0,89	-0,17
10	-0,49	0,02	-0,06	0,50	-0,36	0,56	0,25	-0,82	-0,15	1,00	-0,47	-0,24	-0,39
11	1,00	0,68	0,23	0,08	0,95	-0,81	-0,15	0,77	0,93	-0,47	1,00	-0,68	-0,03
12	-0,65	-0,82	-0,01	-0,53	-0,73	0,37	0,20	-0,18	-0,89	-0,24	-0,68	1,00	0,47
13	0,02	-0,62	-0,20	0,08	-0,18	0,11	0,55	0,27	-0,17	-0,39	-0,03	0,47	1,00

TABELA II

COEFICIENTES DE CORRELAÇÃO ENTRE OS VETORES DAS MÉDIAS DE CADA CARACTERÍSTICA PARA CADA PADRÃO. AS CARACTERÍSTICAS ESTÃO AQUI REPRESENTADAS POR NÚMEROS DE 1 A 13.

K características consideradas simultaneamente	Número de sistemas	Coeficiente de correlação entre Λ e porcentagem de acertos
1	13	-0,7856
2	78	-0,8267
3	286	-0,8402
4	715	-0,8563
5	1287	-0,8614
6	1716	-0,8628
7	1716	-0,8693
8	1287	-0,8830
9	715	-0,9006
10	286	-0,9230
11	78	-0,9465
12	13	-0,9647

TABELA III

CORRELAÇÃO ENTRE A TAXA MÉDIA DE ACERTOS E O COEFICIENTE Λ.

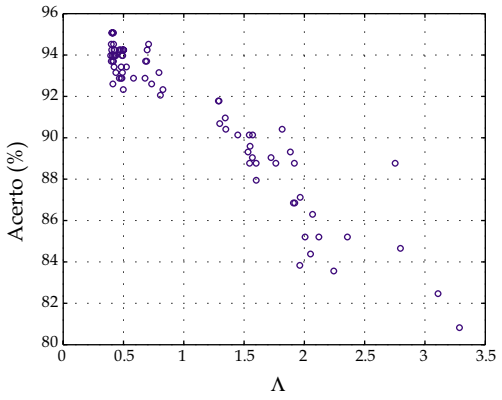

Fig. 2. Taxa de acerto em função de Λ correspondentes a todos os sistemas de classificação gerados com $k = 11$ para algarismos.

V. RESULTADOS

O gráfico Fig. 2 apresenta o resultado do número de acertos de todos os sistemas de classificação em função de Λ para $k = 11$. Observamos que os menores valores de Λ resultam nas mais elevadas eficiências.

Da mesma forma, utilizamos este método no reconhecimento de letras extraídas de placas de automóveis, onde aplicamos as mesmas características e o mesmo pré-processamento empregados na análise dos algarismos. As amostras foram extraídas de 532 placas de automóveis, com aproximadamente 36 imagens de cada uma das 26 letras do alfabeto. Montamos as matrizes M e F, e calculamos o coeficiente Λ para todas as possíveis combinações de características. Podemos observar no gráfico da Fig. 3 que os maiores valores de eficiência correspondem aos menores valores de Λ.

Para o caso de classificação de letras considerando apenas uma característica, a máxima taxa de acerto conseguida foi de 31,1%. Deste modo, podemos concluir que o valor de Λ está relacionado com a taxa de acerto quando ela representa um valor razoavelmente alto. Podemos comprovar essa afirmativa com o seguinte exemplo: o coeficiente de correlação entre Λ e a taxa de acerto para os possíveis sistemas que utilizavam 11 características simultaneamente é de -0.7 e a maior taxa de acerto para esse grupo de sistemas é 85,9%. Notar a baixa correlação entre o Λ e a taxa média de acerto encontrada para sistemas de classificação de letras utilizando poucas características simultaneamente, pois nesses casos, os sistemas de classificação são muito ruins e não são frequentemente utilizados.

VI. CONCLUSÕES

Perante os resultados obtidos podemos concluir que efetivamente o coeficiente Λ, resultado final do método de análise de características, está relacionado com a taxa de acerto. Isso pôde ser visto tanto no teste de classificação de algarismos como no de classificação de letras extraídos de placas de automóveis.

O coeficiente Λ é um valor importante que nos diz se o sistema, na média, terá uma taxa alta ou baixa de acerto. Um

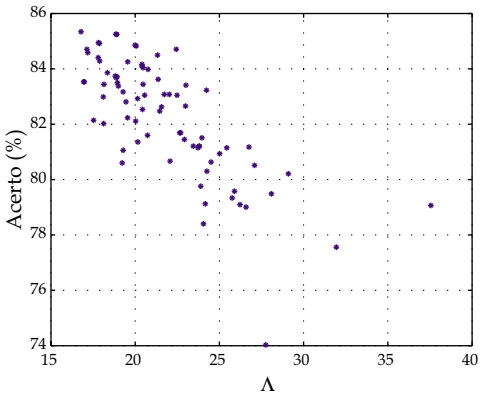

Fig. 3. Taxa de acerto em função de Λ correspondentes a todos os sistemas de classificação gerados com $k = 11$ para letras.

outro elemento importante que não pode passar despercebido é a matriz F, que indica onde se encontram as maiores vulnerabilidades e as maiores virtudes do sistema. Essa matriz indica também quais classes estão separadas e mostra entre quais deve ser despendida maior atenção com o intuito de aumentar a separação entre elas. Com a matriz F, conseguimos isolar problemas em classificação, fazendo com que o tempo não seja desperdiçado em setores que já apresentem uma boa separação.

AGRADECIMENTOS

Este trabalho foi apoiado pelo Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq) e pela Fundação Carlos Chagas Filho de Amparo à Pesquisa do Estado do Rio de Janeiro (FAPERJ).

REFERÊNCIAS

[1] GONZALEZ, Rafael C., “Digital Image Processing”, Second edition, Addison-Wesley Publishing Company, ISBN 0-201-11026-1, **1987**.

[2] JAIN, Anil/ZONGKER, Douglas, “Feature Selection: Evaluation, Application, and Small Sample Performance”, IEEE Transactions on PAMI, 19, 153–158, **1997**.

[3] BISHOP, Christopher M., “Neural Networks for Pattern Recognition”, Oxford University Press, **1997**.

[4] O’ROURKE, Joseph e TOUSSAINT, Godfried, “Handbook of Discrete and Computational Geometry - Chapter 43 - Pattern Recognition”, Editores J. E. Goodman e J. O’Rourke, CRC Press, Nova Iorque, pp. 797-813, **1997**.

[5] ALBUQUERQUE, Marcelo P. de, ALBUQUERQUE, Márcio P. de, GESUALDI, Aline da R., ESQUEF, Israel A., “Image thresholding using Tsallis entropy”. Pattern Recognition Letters. , **2004**.

[6] Matlab Funciton Reference, http://www.mathworks.com/access/helpdesk_r12p1/help/techdoc/ref/ref.shtml, **2004**.

[7] BELLMAN, R., “Adaptive Control Processes: A Guided Tour”, Princeton, **1961**.

[8] DUDA, Richard O.; HART, Peter E. e STORK, David G., “Pattern Classification”, John Wily & Sons, Inc., November, Nova Iorque, Segunda Edição,ISBN 0-471-05669-3, **2000**.