Avaliação de Índice Invertido em Busca de Imagens por Conteúdo

Tauller Matos[1], Ilmério Silva[1], Celia Z. Barcelos[2] e Patrícia Proença[1]

[1] Universidade Federal de Uberlândia, Faculdade de Computação,
Uberlândia, Brasil, 38400-902
tauller@yahoo.com.br, ilmerio@facom.ufu.br, patriciaproenca@gmail.com
[2] Universidade Federal de Uberlândia, Faculdade de Matemática,
Uberlândia, Brasil, 38400-902
celiazb@ufu.br

Resumo Apresentamos uma proposta inovadora de utilização do Índice Invertido para recuperação de imagens baseado em conteúdo (CBIR). O objetivo é acelerar o processamento de consultas, sem perda de qualidade na resposta. Para avaliar a eficácia da proposta foram desenvolvidos dois sistemas de recuperação de imagens. Um baseado nas técnicas normalmente utilizadas em CBIR, a saber, distância euclidiana. O segundo, indexa as imagens através de uma estrutura da recuperação textual conhecida como índice invertido e calcula a similaridade através da medida do co-seno. Por meio de experimentos, fazemos uma análise comparativa do número de operações aritméticas efetuadas pelos dois sistemas no cálculo da similaridade mostrando o ganho significativo no tempo de processamento.

1 Introdução

O número de imagens digitais disponíveis na Web tem crescido exponencialmente nos últimos anos. O baixo custo dos equipamentos juntamente com dispositivos de alta capacidade de armazenamento e compartilhamento de informações têm contribuído para isto.

Neste cenário, encontrar informações relevantes para o usuário, se torna cada vez mais uma tarefa difícil. A grande quantidade de imagens disponíveis aumenta a necessidade de melhores algoritmos de recuperação de imagens, métodos de indexação e técnicas de classificação.

Existem duas abordagens principais de recuperação de imagens: recuperação baseada em anotações textuais e recuperação baseada no conteúdo visual. A primeira utiliza-se de anotações manuais para descrever o conteúdo das imagens. Neste caso, consultas são expressas por palavras chaves e efetuadas por técnicas de gerenciamento de banco de dados. Esta abordagem é limitada, pois é inviável prover anotações de imagens para grandes coleções, além disso, anotações manuais são subjetivas - uma mesma imagem pode ser interpretada de diferentes maneiras por diferentes pessoas.

XXXIV Conferencia Latinoamericana de Informática (CLEI 2008), pp. 1080-1089, 2008.

A segunda técnica bastante utilizada para recuperação de imagens é conhecida como Recuperação de Imagens Baseada no Conteúdo (CBIR). Seu principal objetivo é encontrar imagens relevantes conforme a necessidade do usuário, através de características visuais automaticamente extraídas das imagens. Atualmente, os métodos de representação de características das imagens mais utilizados usam: cor, textura e forma como atributos de indexação, os quais são extraídos da imagem de maneira independente (veja [1] e [2]). Uma das diferenças dessa abordagem com a recuperação baseada em anotações textuais é conter uma imagem exemplo como consulta e não palavras-chaves.

Uma vez que as características das imagens tenham sido extraídas e armazenadas em vetores de característica, fazem-se necessárias medidas que comparem vetores de característica das imagens do banco de dados com o vetor de característica da imagem exemplo. Essas medidas, são normalmente baseadas em distância entre vetores das imagens.

Uma medida comumente utilizada é a distância euclidiana. Vários trabalhos utilizam distância euclidiana para calcular a similaridade entre a imagem de consulta e as imagens do banco de dados. Dentre os quais, destacamos o Sistema MARS [6] que calcula a similaridade de características de textura entre imagens. O Sistema Netra [3], também a utiliza para calcular a similaridade baseada em cor e forma. Em CIRES [4] e Blobworld [5] a comparação de características de forma e textura também é obtida por meio desta medida de distância. Em todos estes sistemas, o processo de recuperação inclui o cálculo da distância entre o vetor da imagem de consulta e todos os vetores das imagens do banco de dados, tornando a recuperação computacionalmente cara. Este processo é inviável para grandes coleções de imagens, por exemplo a Web.

Diante disso, propomos uma nova abordagem, tanto para indexação das imagens quanto para o cálculo da similaridade em CBIR e comparamos o desempenho da proposta. Para isto, desenvolvemos dois sistemas, aqui chamados de CBIR-Cor e CBIR-Índice. O primeiro sistema utiliza as técnicas convencionais de CBIR e no segundo propomos uma abordagem para interpretação das características de baixo nível baseado no modelo vetorial utilizado em recuperação textual. Neste segundo sistema, a indexação das imagens é feita através da estrutura chamada de índice invertido e para calcular a similaridade utilizaremos a medida do co-seno. Segundo Baeza-Yates et. al. [7] o uso do índice invertido em recuperação textual tem resultado em ganho na velocidade de recuperação de documentos sem perda de qualidade na recuperação. Estes conceitos, bem como estes dois sistemas serão descritos mais adiante.

Este trabalho tem como objetivo comparar o número de operações aritméticas realizadas utilizando a distância euclidiana e a do co-seno, com o intuito de mostrar a possibilidade de ganho de tempo no processo de recuperação, além de uma avaliação na qualidade da recuperação.

Este artigo é distribuído da seguinte forma: na Seção 2 descrevemos o sistema CBIR-COR; é feito um estudo da distribuição dos vetores de característica gerados pelo CBIR-Cor na Seção 3; Na Seção 4 definimos o CBIR-Índice juntamente com os conceitos de índice invertído, modelo vetorial e da medida de similari-

dade do co-seno, na Seção 5 compara-se o número de operações aritméticas dos dois sistemas além dos resultados das medidas de precisão encontradas, e para finalizar na Seção 6 concluímos e abordamos os trabalhos futuros.

2 Sistema CBIR-Cor

Para a construção dos dois sistemas de recuperação de imagens por conteúdo foi utilizado o extrator de característica de baixo nível Momentos de Cor [8]. Este caracteriza as imagens em termos da distribuição das cores nos *pixels* da imagem. São normalmente dados por três medidas estatísticas: média, desvio padrão e obliquidade. Os momentos de cor podem ser derivados do espaço de cor HSV correspondendo aos canais matiz, saturação e intensidade da seguinte forma:

Sendo N o número de pixels de uma imagem e p_{ij} o valor do j-ésimo pixel no i-ésimo canal de cor, os três primeiros momentos: média (E_i), desvio padrão (σ_i) e obliquidade (s_i) são computados por:

$$E_i = \frac{1}{N}\sum_{j=1}^{n} p_{ij} \tag{1}$$

$$\sigma_i = \sqrt{\left(\frac{1}{N}\sum_{j=1}^{n}(p_{ij} - E_i)^2\right)} \tag{2}$$

$$s_i = \sqrt[3]{\left(\frac{1}{N}\sum_{j=1}^{n}(p_{ij} - E_i)^3\right)} \tag{3}$$

Em resumo, convertemos as imagens do espaço de cor RGB para o espaço de cor HSV. Então, calculamos os três momentos acima para cada canal H, S e V. Com isto, o vetor de característica da imagem terá nove posições distribuídas conforme Tabela 1.

O CBIR-Cor indexa as imagens com o valores encontrados nas equações 1, 2 e 3, para cada canal de cor HSV e calcula a similaridade baseada na distância euclidiana, dada pela equação 4.

$$d(x, y) = \sum_{k=1}^{9} |x_k - y_k|^2 \tag{4}$$

onde x_k é o valor correspondente a um elemento do vetor x que representa uma imagem do banco de dados de característica e y_k é a k-ésima posição do vetor de característica que representa a imagem de consulta.

Para grandes coleções, esta forma de calcular a similaridade torna a recuperação das imagens computacionalmente cara. Pois o processo de recuperação inclui o cálculo da distância entre o vetor da imagem de consulta e todos os

Tabela 1. Descrição do vetor de característica resultante da extração de característica da imagem usando os três momentos do espaço de cor HSV

Vetor	Descrição
1	média da matiz (MH)
2	desvio padrão da matiz (DH)
3	obliquidade da matiz (IH)
4	média da saturação (MS)
5	desvio padrão da saturação (DS)
6	obliquidade da saturação (IS)
7	média da intensidade (MV)
8	desvio padrão da intensidade (DV)
9	obliquidade da intensidade (IV)

vetores das imagens do banco de dados. Por este motivo, este processo é inviável para grandes coleções de imagens.

Com o objetivo, de construir um sistema viável para recuperação de imagens em grandes bases de dados como a *Web*, propomos uma nova abordagem tanto para indexação das imagens quanto para o procedimento do cálculo da similaridade em CBIR. Para elaboração deste novo sistema foi necessário analisar a distribuição dos valores dos vetores de características gerados pelo CBIR-Cor.

3 Vetores de características

A distribuição dos valores em cada uma das nove posições do vetor (ver Tabela 1), quando observada em um histograma obtido da coleção utilizada neste trabalho (Corel 1000, que será abordada mais adiante), tem comportamento semelhante à distribuição normal de probabilidades. Dentro das propriedades da distribuição normal o desvio padrão é usado para determinar a percentagem de ocorrências dentro de um intervalo [10]. Com intuito de agrupar as imagens semelhantes e baseado na propriedade do desvio padrão optamos por aplicar este domínio em nossas distribuições separando cada posição do vetor em 10 classes, usando a tabela da distribuição normal padronizada [9], conforme a Figura 1.

As faixas E e F contém as características mais frequentes na base de dados e portanto são poucas discriminantes. Assim como em [11] usamos os 30% dos termos mais comuns para definir as faixas dos valores de características mais frequentes. Assim, teremos 15% dos termos à direita e 15% a esquerda. As outras faixas foram definidas diminuindo a frequência de 2% em 2% para direita e para esquerda. Com exceção das faixas A e J que ficaram com a percentagem restante para os 100% da amostra.

Desta forma, cada intervalo terá a seguinte percentagem aproximada de imagens:

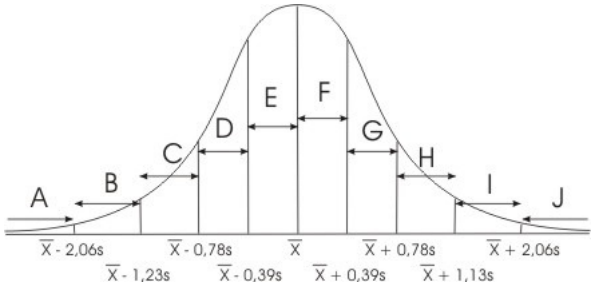


Figura 1. Distribuição adotada para classificação das faixas. Onde $\bar{X}$ é a média e s o desvio padrão.

- 15% dos casos estão incluídos no intervalo [$\bar{X}$ - 0.39s, $\bar{X}$] (correspondente a faixa E da Figura 1) e 15% está no intervalo [$\bar{X}$, $\bar{X}$ + 0.39s] (correspondente a faixa F da Figura 1);
- 13% dos casos estão incluídos no intervalo [$\bar{X}$ - 0.78s, $\bar{X}$ - 0.39s] (faixa D) e 13% está no intervalo [$\bar{X}$ + 0.39s, $\bar{X}$ + 0.78s] (faixa G);
- 11% dos casos estão incluídos no intervalo [$\bar{X}$ - 1.23s, $\bar{X}$ - 0.78s] (faixa C) e e 11% está no intervalo [$\bar{X}$ + 0.78s, $\bar{X}$ + 1.23s] (faixa H);
- 09% dos casos estão incluídos no intervalo [$\bar{X}$ - 2.06, $\bar{X}$ + 1.23s](faixa B) e 09% está no intervalo [$\bar{X}$ + 1.23s, $\bar{X}$ + 2.06s] (faixa I);
- 02% dos casos estão incluídos no intervalo [0, $\bar{X}$ - 2.06] (faixa A) e 02% está no intervalo [$\bar{X}$ + 2.06s, 1] (faixa J);

4 Sistema CBIR-Índice

A partir do mapeamento dos grupos de valores de características de baixo nível de imagens digitais, criamos os identificadores de indexação da coleção. Portanto, as imagens são indexadas através de uma estrutura popular em recuperação de informação textual chamado índice invertido. A idéia é limitar a comparação somente com o subconjunto de imagem que contenham pelo menos uma característica comum a imagem de consulta. Portanto não há necessidade de se percorrer toda a base de dados.

4.1 Índice Invertido

Possui duas partes principais: uma estrutura de busca, chamada de vocabulário, contendo todos os termos distintos existentes nos textos indexados e, para cada termo, uma lista invertida que armazena os identificadores dos registros contendo o termo, a saber, documentos onde a palavra ocorre [7], como ilustrado na Figura 2 utilizando imagens. Neste caso, os termos são obtidos da união das 10 faixas da tabela 1. Por exemplo, MH forma MHA, MHB, ..., MHJ.

Consultas nos arquivos invertidos são feitas tomando-se a lista invertida correspondente aos termos procurados.

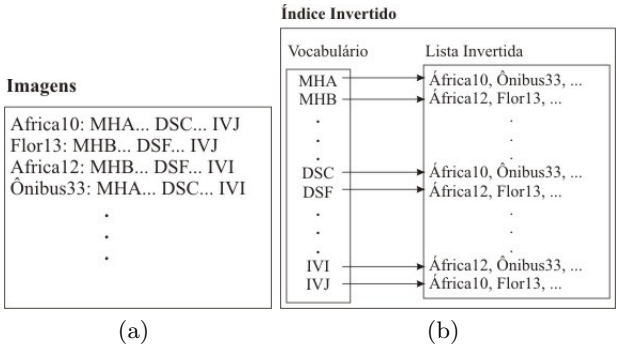


Figura 2. (a) Banco de dados de Característica (b) Estrutura de busca (vocabulário) e a lista invertida

4.2 Modelo Vetorial

Proposto inicialmente por [12], este modelo representa documentos e consultas como vetores de termos. Os termos que compõem o sistema de recuperação são modelados como elementos pertencentes a um espaço vetorial. Cada termo possui um valor (peso) associado a um documento que indica o grau de importância deste termo no documento. Assim, cada documento é constituído por pares de elementos na forma $[termo_i, peso_i]$, representado pelo vetor $\mathbf{d} = (W_{1d}, W_{2d}, ..., W_{nd})$. As consultas também são representadas por vetores de termos $\mathbf{q} = (W_{1q}, W_{2q}, ..., W_{nq})$.

Existem diversas formas de se calcular o peso de um termo no documento. Normalmente estes cálculos se fundamentam no número de ocorrências do termo no documento e na coleção (frequência). Neste trabalho consideramos o número de ocorrências do termo no documento como booleano. Se o termo estiver presente no documento ele terá o peso igual a 1; caso contrário ele terá peso igual a 0. A frequência na coleção é dada pelo IDF (*inverse document frequency*), que é calculado da seguinte forma:

$$idf_t = \log \frac{N}{df_t} \tag{5}$$

onde N é o número total de documentos da coleção e df_t é número de vezes que o termo ocorre na coleção. Assim, o IDF de um termo raro é elevado, enquanto que o IDF de termos comuns decresce em escala logarítmica.

A utilização de uma mesma representação para documentos e consultas permite o cálculo da similaridade entre uma consulta q e um documento d_j. Este cálculo é realizado através da correlação entre os vetores que os representam, quantificada pelo co-seno do ângulo formado por d_j e q. Esta métrica é conhecida como medida de similaridade do co-seno. Desta forma, em um espaço vetorial de dimensão n, a similaridade (sim) entre dois vetores $\mathbf{q}$ e $\mathbf{d}$ é calculada através do co-seno do ângulo formado por estes vetores, através da seguinte fórmula [11]:

$$sim(\boldsymbol{q}, \boldsymbol{d}) = \frac{\sum_{i=1}^{n} W_{qi} W_{di}}{\sqrt{\sum_{i=1}^{n} W_{qi}^2} \sqrt{\sum_{i=1}^{n} W_{di}^2}} \quad (6)$$

onde **q** é o vetor de termos da consulta; **d** é o vetor de termos do documento; W_{qi} é o peso do termo i da consulta q e W_{di} é o peso do termo i no documento d.

Os valores da similaridade entre uma expressão de busca e cada um dos documentos da coleção são utilizados na ordenação dos documentos resultantes. Portanto, no modelo vetorial os resultados são ordenados de acordo com a medida de similaridade do co-seno formando o *ranking*.

Para o contexto deste trabalho fazemos as seguintes analogias entre recuperação de imagens e recuperação textual: imagens correspondem aos documentos; faixas são os termos; e imagens exemplos ou imagens de consulta são as consultas na recuperação textual.

5 Resultados Experimentais

A coleção utilizada para extração dos vetores de características das imagens de ambos os sistemas foi a Corel-1000. Esta base é um subconjunto do banco de dados Corel, contendo 1000 imagens com resolução 384×256 por 256×384 *pixels*. O banco de dados é composto por 10 categorias (África, praia, edifícios, ônibus, dinossauros, elefantes, flores, comidas, cavalos e montanhas), com 100 imagens em cada categoria.

Para avaliarmos a qualidade de recuperação, entre os dois sistemas implementados, utilizaremos a medida de desempenho conhecida como precisão (P). Dada uma consulta, a precisão é definida como a fração entre o número de imagens relevantes recuperadas, sobre o número total de imagens recuperadas:

$$P = \frac{Imagens\_relevantes\_recuperadas}{imagens\_recuperadas} \quad (7)$$

Para efetuar a medida da precisão baseamos no argumento encontrado em [11], que para muitas aplicações, particularmente a busca na *Web*, o importante é que os documentos mais relevantes estejam na primeira página e no máximo na terceira. Portanto, calculamos a precisão em 4 posições do *ranking*. Na posição P0, significa calcular a precisão para a primeira imagem relevante encontrada no *ranking*, com exceção da imagem de consulta. P5, P10 e P20, são respectivamente a 5^a, 10^a e 20^a posição do *ranking*. Para a realização das consultas foram tomadas 5 imagens de cada classe relacionadas aleatoriamente.

A Tabela 2 mostra os valores encontrados das precisões P0, P5, P10 e P20 do sistema CBIR-COR, relativo a média de 5 consultas de cada classe, juntamente com o valor médio da precisão de 50 consultas da coleção. Na Tabela 3 encontramos estes mesmos resultados para o sistema CBIR-Índice.

De acordo com os resultados podemos observar que o CBIR-Índice apresentou um ganho na qualidade de recuperação no ponto P0. Na precisão P5, a percentagem média foi bem similar. Nos pontos de precisão P10 e P20, o CBIR-Índice,

Tabela 2. Cálculo da precisão no sistema CBIR-Cor

Classe	P0	P5	P10	P20
África	69,8%	68%	60%	60%
Praia	70,7%	46%	46%	41%
Construções	77,5%	60%	50%	40%
Dinossauros	100%	100%	98%	97%
Ônibus	82,4%	64%	60%	50%
Cavalos	100%	100%	88%	76%
Flores	90%	70%	50%	37,5%
Comidas	86,6%	88%	84%	79%
Elefantes	100%	88%	74%	60%
Montanhas	49,4%	52%	46%	49%
Média Geral	**82,6%**	**73,6%**	**65,6%**	**59%**

Tabela 3. Cálculo da precisão no sistema CBIR-Índice

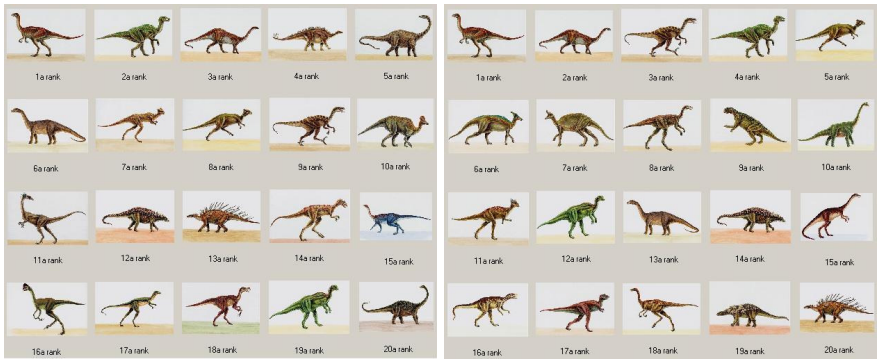
Classe	P0	P5	P10	P20
África	73,2%	84%	60%	55%
Praia	87,5%	52%	46%	41%
Construções	55%	56%	48%	38%
Dinossauros	100%	96%	94%	94%
Ônibus	73,2%	72%	60%	45%
Cavalos	90%	88%	80%	70%
Flores	100%	60%	57,5%	42,5%
Comidas	86,6%	76%	66%	58%
Elefantes	100%	80%	74%	54%
Montanhas	66,6%	60%	42%	39%
Média Geral	**83,2%**	**72,4%**	**62,8%**	**53,7%**

obteve uma perda pequena na qualidade de recuperação. Observa-se então que, o *ranking* gerado com o método proposto neste trabalho tem melhor qualidade nas regiões de alta precisão.

Também foi comparado nos dois sistemas o número de operações aritméticas efetuadas para o cálculo da similaridade, entre a imagem de consulta e a base de dados.

No CBIR-Cor, de acordo com a equação 4, e o vetor de característica da Tabela 1, para cada consulta fazemos as seguintes operações: i) 9 subtrações; ii) 9 multiplicações; iii) 8 adições; iv) 1 raiz. Como mencionado na seção 2 o CBIR-Cor, faz a comparação do vetor de consulta com todas as imagens da coleção, com o objetivo de se obter o *ranking* de similaridade final. Tendo a base de dados 1000 imagens ao final de uma consulta totaliza-se 28.000 operações aritméticas.

Entretanto no CBIR-Índice, por meio da lista invertida não é necessário varrer toda a coleção, mas somente aquelas que casarem com pelo menos um termo da consulta. Para cada característica (termo) presente faremos as seguintes operações: i) $z * 9$ somas, onde z é o tamanho médio das listas invertidas; ii) y



(a) (b)

Figura 3. (a) resultado de uma consulta no sistema CBIR-Cor; (b) resultado de uma consulta no sistema CBIR-Índice. Ambos na classe dinossauro

divisões, onde y é o número total de imagens encontradas nas listas. Em nossos experimentos z e y tiveram valores médios de 1038 adições e 677 divisões respectivamente. Totalizando 1715 operações aritméticas por consulta. Valor bem abaixo dos 28.000 encontrados no CBIR-Cor. O denominador da equação 6 não é necessário calcular no momento da consulta, pois ele é calculado no momento da indexação das imagens. Isto indica um ganho de tempo no processamento da consulta utilizando o índice inverto e o cálculo do co-seno no processo de recuperação, similarmente em grandes coleções de imagens.

A Figura 3 mostra o resultado de uma consulta nos dois sistemas. Onde a primeira imagem do ranking corresponde a imagem de consulta.

Uma abordagem que compara a distância euclidiana com a distância do co-seno foi abordada em [13]. Provou-se que ambas as distâncias trazem resultados bem similares. A ordem das imagens recuperadas é que pode sofrer alterações devido a variância da amostra. Mas em [13] não foi utilizada a metodologia de indexação das imagens proposta aqui, com o uso do índice invertido por meio do mapeamento em faixas.

6 Conclusão e trabalhos futuros

Este trabalho fez a comparação do desempenho entre dois sistemas. O primeiro utiliza a medida da distância euclidiana para o cálculo da similaridade e o outro calcula através da medida do co-seno. Além de indexar as imagens através do índice invertido. Mostrou-se através de experimentos, que o segundo sistema, tem um ganho significativo no tempo de processamento por realizar menos operações aritméticas, pois não há necessidade de se varrer toda a base de dados. Na avaliação do resultado de recuperação o sistema manteve a qualidade nas regiões de alta precisão, mas com uma pequena queda à medida que se avança no *ranking*.

Como trabalhos futuros, pretendemos melhorar a qualidade do resultado por meio de novos algoritmos para definir as faixas dos vetores de características e de outras características de baixo nível, como textura e forma.

Referências

1. Smeulders Awm, Worring M, Gupta A, and Jain R. Content based image retrieval at the end of the early years. IEEE Transactions on Image Processing Intell 22, pages 1349-1380, 2000.
2. Ryszard S. Choras, Tomasz Andrysiak, and Michal Choras. Integrated color, texture and shape information for content based image retrieval. Pattern Anal Applic, 2007.
3. Ma, W. Y. and Manjunath, B. S. Netra: A toolbox for navigating large image databases. Multimedia Systems, 7(3):184-198, 1999.
4. Iqbal, Q. and Aggarwal, J. K. CIRES: A System for Content-Based Retrieval in Digital Image Libraries. In Seventh International Conference on Control, Automation, Robotics and Vision, pages 205-210, 2002.
5. Carson, C., Thomas, M., Belongie, S., Hellerstein, J. M., and Malik, J. Blobworld: A system for region-based image indexing and retrieval. In Proceedings of the Third International Conference VISUAL, Lecture Notes in Computer Science, Spring, Amsterdam, Netherlands, 1999.
6. Rui, Y., T.S.Huang, and Mehrotra, S. . Content-based image retrieval with relevance feedback in MARS. In Proceedings of International Conference on Image Processing, volume 2, pages 815-818, 1997.
7. R. Baeza-Yates and B. Ribeiro-Neto. Modern Information Retrieval. Addison- Wesley Essex, UK, 1999.
8. Stricker, M. and Orengo, M. Similarity of Color Images. In Proceedings of IS&T and SPIE Storage and Retrieval of Image and Video Databases III,San Jose, USA, 1995.
9. Levine, D. M. and Berenson, M. L., Stephan, D. Estatística: Teoria e Aplicações: Usando Microsoft Excel. 1998 tr. It. de Teresa Cristina Padilha de Souza, LTC, Rio de Janeiro, 2000.
10. Spiegel, M. R. Scaum´s Outline of Theory and Problems of Estatistics. Schaum Publishing Co. E.U.A 1970. tr. It. de Pedro Cosentino, ed. McGraw-Hill do Brasil, Rio de Janeiro, 1971.
11. Manning, C., Raghavan, P., Schütze, H., An Introduction to information Retrieval. Cambridge University Press, Cambridge, England, 2007, 365pgs.
12. Salton, G. and Buckley, C. Term-weighting approaches in automatic text retrieval. Information Processing and Management 24, 513-523. 1988. Reprinted in Sparck jones, K. and Willet, P. (Eds.) Readings in Information Retrieval, 323-328. Morgan Kaufmann, 1997.
13. Qian, G. and Surat, S. and Pramanik, S. A comparative analysis of two distance measures in color image databases. ICIP02, 401-404, 2002.