Guaatupi: um ambiente para indexação e recuperação de imagens da *web* sem redundância visual[1]

Thiago Fonsêca Meneses
Departamento de Sistemas e Computação - Universidade Federal de Campina Grande - UFCG
Av Aprígio Veloso, 882, Campus Universitário
CEP 58109-970, Campina Grande, Paraíba, Brasil
+55 (83) 3310.2015
thiagofmam@gmail.com

Carlos A. F. Pimentel Filho
Departamento de Ciência da Computação - Universidade Federal de Minas Gerais - UFMG
Av. Antônio Carlos, 6627, Pampulha, CEP 31270–901, Belo Horizonte, Minas Gerais, Brasil
+55 (31) 3409.5000
fragapimentel@gmail.com

Rafael W. M. de Araujo
Departamento de Ciência da Computação - Instituto de Matemática e Estatística - Universidade de São Paulo - USP
Rua do Matão, 1010, Cidade Universitária
CEP 05508-090, São Paulo, SP, Brasil
+ 55 (11) 3091.6135
rwill@ime.usp.br

ABSTRACT
This work presents an approach for image retrieval system based on its visual features applied to World Wide Web scenario. The proposed approach presents techniques for crawling, indexing and retrieving without visual redundance image copies. In order to achieve its goals, the present work uses a spider to capture web images from Google Images and wavelets tranform in order to extract the essence of image features that well represents them. In the present work, it was indexed 310 thousand images, which were reduced to 225 thousand after image copies elimination. It represents a reduction in about 27% of redundancy and 20% faster.

RESUMO
Este artigo aborda um sistema de recuperação de imagens com base em suas características visuais voltado para *World Wide Web*. O trabalho proposto apresenta técnicas para navegação, indexação e recuperação de imagem sem cópias redundantes. Para alcançar tais objetivos, o presente trabalho utiliza uma ferramenta de *spider* para a captura de imagens do *Google Images.* A proposta também usa a transformada *wavelet* para extrair as características essências que representam cada imagem. Foram indexadas 310 mil imagens, sendo esse número reduzido a 225 mil eliminando-se as cópias. Isso representa uma redução de cerca de 27% de redundância e um incremento na velocidade das consultas em 20%.

Categories and Subject Descriptors
H.3.1 [**Information Storage and Retrieval**]: Content Analysis and Indexing – *Abstracting methods and Indexing methods*

General Terms
Management, Measurement, Performance, Experimentation.

Keywords
Indexação de imagens, CBIR, recuperação de conteúdo.

1. INTRODUÇÃO
Com a expansão da *web* e dos conteúdos multimídia, tais como imagens, vídeos e áudio, a necessidade de indexação e acesso a essa categoria de documento torna-se cada vez mais importante. No caso específico das imagens, surgem novas questões: como indexar e recuperar imagens dentre bilhões de outras imagens? Dentre as imagens indexadas como identificar cópias duplicadas?

A recuperação de imagens e o aprendizado de máquina é uma tarefa não trivial [1]. Esta recuperação pode ser feita a partir de metadados previamente anotados ou a partir de metadados obtidos por extração automática de características [3]. A primeira abordagem tem como vantagem maior riqueza semântica nos dados anotados, em contrapartida, o custo de anotação das imagens é alto e proibitivo em escalas onde o conjunto de imagens ultrapassa certos limites. A segunda possibilidade é a extração automática de características visuais. Desse modo é possível aplicar técnicas de CBIR (*Content Based Image Retrieval*) na recuperação de imagens. Nesta técnica, descritores matemáticos são usados com intuito de representar o conteúdo visual, tipicamente em informações consideradas de baixo nível e baixo valor semântico. As características visuais mais comuns são relacionadas com a cor, a forma e a textura [3].

De acordo com [1], um dos maiores desafios na computação visual e processamento de imagens é a escalabilidade, especialmente quando se considera um grande conjunto de imagens provido da *web*. Por natureza, a *web* é um ambiente onde os documentos são comumente replicados. Dentre as diversas mídias, as imagens duplicadas são um problema porque demandam grande custo de espaço. Um dos problemas

[1] Guaatupi: An approach for web image indexing and retrieval without copy redundance.

encontrados na busca e indexação de imagens é justamente esse excesso de redundância. Assim, diversos problemas são gerados: (i) o indexador re-processa e re-armazena uma grande quantidade de informação redundante; (ii) os custos de armazenamento e processamento são incrementados; (iii) a busca em uma base de dados redundante torna-se mais lenta e (iv) os usuários podem ter que navegar por diversas páginas de conteúdo repetido até encontrar o que realmente deseja. Nesse sentido, o objetivo do presente trabalho é mostrar uma abordagem para indexação e eliminação de conteúdo visual redundante nas buscas de imagens da *web*. Como vantagens, a identificação e eliminação de cópias contribui com a redução de custos financeiros e computacionais em sistemas de indexação e busca de imagens.

A abordagem proposta pelo presente trabalho implementa e avalia um ambiente de indexação e recuperação de imagens *web* chamado de *Guaatupi*. A referida ferramenta compreende as fases de *crawler*, extração de características visuais das imagens, eliminação de redundância, interface com o usuário e busca de imagens por conteúdo visual.

Como é mostrado nos resultados, a abordagem proposta consegue identificar cópias de imagens redundantes, reduzindo a base em aproximadamente 27%, melhorando por conseguinte o tempo médio de espera das buscas em 20%. Assim, proporcionando uma navegação nos resultados mais relevantes aos usuários.

O artigo está estruturado da seguinte maneira: a seção 2 apresenta os conceitos fundamentais da recuperação de informação visual, o ambiente proposto e as metodologias utilizadas nos experimentos. A seção 3 mostra os resultados e finalmente as conclusões são apresentadas na seção 4.

2. RECUPERAÇÃO DE INFORMAÇÃO VISUAL: A ABORDAGEM PROPOSTA (*GUAATUPI*)

Em 1951, Calvin Moores [2] cunhou o termo "*Information Retrieval*" ou "Recuperação de Informação", associando-o ao processo sobre o qual, um conjunto de definições pode converter uma requisição de informação num conjunto de referências de fato úteis. Este processo, em geral, inicia-se com a definição de uma estrutura de índices que é utilizada para permitir a recuperação de partes específicas do conjunto de informações armazenado.

A indexação do conteúdo de imagens da *web* é uma necessidade para a organização da informação lá contida. Suas vantagens vão desde a recuperação rápida e eficiente do seu conteúdo à identificação e descrição automática de conteúdo visual que podem melhorar o acesso a *web* de deficientes visuais. Devido à vastidão da *web*, sua indexação não tem como ser feita de modo manual, assim o estudo de técnicas de extração de características visuais visam suprir essa necessidade.

2.1 Web Crawler

Um *web crawler* também denominado: *ants*, *automatic indexers*, *bots*, *web spider*, *web robot* ou ainda w*eb scutter* é uma categoria de agente que percorre e recupera informações contidas em páginas da *web* de forma automática [4][5]. Em síntese, um *web crawler* é iniciado com uma URL (*Uniform Resource Locator*) que serve como "semente" de onde o *crawler* irá poder recuperar informações da *web* varrendo por outras URLs encontradas na semente e em seus *links* sucessores. Uma vez que a página indexada é sumarizada, seu conteúdo estará disponível para consultas [7].

2.2 Arquitetura do *Guaatupi*

O *Guaatupi* é um sistema completo de indexação e busca de imagens da *web*. O protótipo foi dividido em dois módulos: um denominado de *Tupi*, que é responsável pelo *crawler*, *download*, extração de características visuais e indexação das imagens, e outro módulo, chamado de *Guarani*, que por sua vez é responsável por oferecer uma interface de busca com a qual o usuário interage. Para implementação do protótipo, foi utilizada linguagem Java e o banco de dados *PostgreSQL*. A arquitetura do sistema é mostrada na Figura 1.

Para a construção de uma base de imagens de testes e indexação para o *Guaatupi*, foi utilizado o mecanismo de busca de imagens do *Google*. Com base em uma lista de pintores de aproximadamente 1300 nomes [8]. Foram buscadas imagens de quadros de artistas que formaram uma base de imagens de pinturas. Os resultados das primeiras páginas do *google imagens* foram utilizados como entradas para indexação, gerando assim, inicialmente, um conjunto de 310.000 imagens, incluindo-se cópias redundantes. Estas cópias são naturalmente presentes nos resultados de busca do *google imagens*.

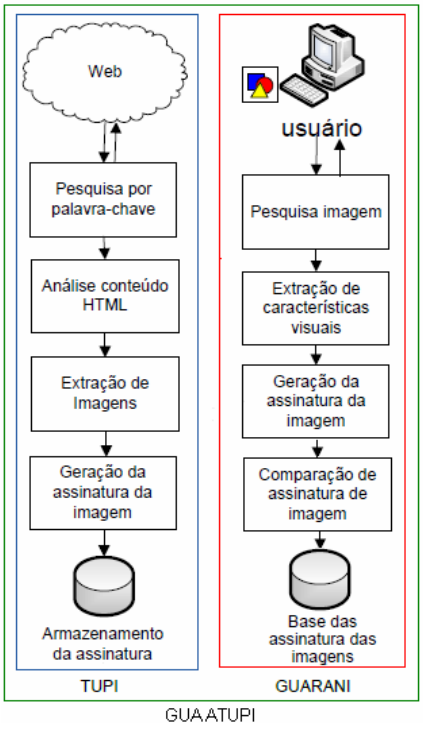

Figura 1 – Arquitetura do *Guaatupi.*

O funcionamento do *Tupi* é feito da seguinte maneira: a partir de uma palavra chave (nome de um pintor), O *Tupi* acessa o mecanismo de busca do *google imagens* passando como parâmetro o nome do artista (nada impede que outro tipo de entrada seja utilizado). A partir dos resultados do *google imagens* o *Tupi* analisa o conteúdo HTML buscando URLs que contenham imagens JPEG/JPG. Uma vez extraídas as URLs, estas são adicionadas a uma lista de URLs para que o módulo de captura e

download de imagens seja executado. A fase seguinte do *Tupi* compreende a extração das características visuais das imagens que serão usadas como parâmetro para busca e comparação de cópias repetidas. As características extraídas das imagens e o processo de comparação foi utilizado exatamente como descrito por Jacobs *et al.* [6]. Estes autores descrevem um processo onde a essência da imagem é resumida em um conjunto de dados obtidos a partir da transformada *wavelet* da imagem, os autores chamaram essa essência de: "assinatura da imagem". Uma vez que a imagem tenha sido processada e tenha sua assinatura obtida, então suas características são armazenadas no banco de dados juntamente com uma miniatura da imagem original.

2.3 Processo de aquisição do banco de imagens

Para executar a busca foi identificado o padrão utilizado pelo *google imagens* para executar as consultas e navegar nos resultados. A URL utilizada para busca e navegação no resultado a partir de uma palavra chave é mostrada a seguir:

http://images.google.com.br/images?q=Palavra-chave&start=PAG

Onde: *Palavra-chave* – Indica a palavra chave utilizada para fazer a consulta e *PAG* indica a página de navegação do resultado.

Uma vez obtida a página proveniente do *google imagens* o *Tupi* procede as buscas nos endereços das imagens para *download* das mesmas.

Após o *download* de cada imagem, as mesmas sofrem um redimensionamento em sua resolução para 128x128 *pixels* devido a uma restrição do algoritmo proposto por Jacobs *et al.* [6], onde as dimensões das imagens devem ser potência de dois. Embora outra resolução pudesse ser escolhida, 128x128 é utilizada com sucesso no trabalho [9] onde esta mesma resolução é utilizada para obtenção das assinaturas de quadros de vídeo.

2.4 Extração das características visuais

Como cita Jacobs *et al*. [6] diversas métricas foram definidas na composição do algoritmo de geração da assinatura da imagem. Essas métricas estão relacionadas com o espaço de cor utilizado, o tipo de característica extraído, a *wavelet* utilizada e outros parâmetros [11]. Para a implementação do *Guaatupi* e extração de assinaturas das imagens, foram utilizados exatamente os mesmos parâmetros e algoritmos descritos em [6]. Em síntese, os parâmetros utilizados para extração das assinaturas das imagens foram: (i) espaço de cor YIQ. (ii) *wavelet* de Haar e (iii) decomposição *wavelet* padrão. Detalhes a respeito de todo processo de extração de assinatura e comparação de similaridade entre imagens podem ser obtidos na referência [6].

2.5 Busca de imagens (*Guarani*)

De modo genérico, a recuperação de imagens pode ser feita basicamente de três maneiras: (i) **"busca por texto"** de acordo anotações previamente realizadas na base de dados; (ii) **"busca por *sketch*"**, onde o usuário fornece uma imagem de rascunho semelhante à aquela(s) que se deseja recuperar, ou ainda (iii) **"busca por exemplo"**, onde a entrada também é uma imagem semelhante a que ele procura, contudo, a imagem "exemplo" pode ser uma versão em baixa resolução, uma fotografia, ou de qualquer outra natureza que não seja um rascunho [9]. No *Guarani*, a recuperação de imagem é feita nos modos: "busca por *sketch*" e "busca por exemplo" [6].

2.6 Eliminação de cópias redundantes

Devido à grande quantidade de cópias observada na base de dados gerada. O passo seguinte para otimização do espaço de armazenamento e redução do custo computacional nas buscas foi a eliminação das cópias redundantes. Para tal tarefa, as imagens da base foram comparadas entre si utilizando a comparação das assinaturas das imagens previamente extraída pelo indexador *Tupi*. A eliminação de redundância tem um custo relativamente alto, uma vez que cada imagem precisa ser comparada com todo outro conjunto do banco de dados. Contudo, à medida que a comparação é feita, as copias já são excluídas. Isso reduz o conjunto de pesquisas nas comparações seguintes do processo. Além disso, o processo de eliminação de cópias é um procedimento realizado uma única vez, ou esporadicamente a medida que a base é incrementada com novas imagens.

O procedimento de eliminação de cópias é descrito a seguir. Seja Q uma imagem cujo conjunto de cópias C que precisa ser eliminado. Seja T o conjunto universo de todas as imagens da base e L um limiar de corte que define se uma imagem é ou não cópia de Q com base na métrica de similaridade da "busca por exemplo" de [6]. O algoritmo consiste em eliminar cada copia C de Q cujo limiar seja menor que um dado L.

2.7 Seleção de Unidades Experimentais

Para investigar o padrão de ocorrência de cópias, foram feitos diversas medições estatísticas. Dentre os estudos realizados, buscou-se encontrar um limiar que defina se duas imagens são cópias uma da outra a partir de suas assinaturas. Para encontrar esse valor, um estudo empírico foi realizado manualmente com 81 imagens. Por fim, experimentos foram feitos para avaliar qual percentual de cópias existente na base e quantas imagens foram eliminadas.

No cálculo do limiar foi necessário calcular o número de amostras utilizadas para inferir sobre a população das 310 mil imagens. Para essa amostra utilizou-se uma precisão de 90% e acurácia de 5% sobre a base de dados, calculando-se a média e o desvio padrão dos valores médios de intensidade de *pixels* do canal de cor Y do espaço de cor YIQ. Com a média da população igual a 119 e desvio padrão de 42.02, o número total de amostras necessárias calculada foi de 81 exemplares de imagens. Para maiores informações sobre o cálculo vide Raj Jain [10].

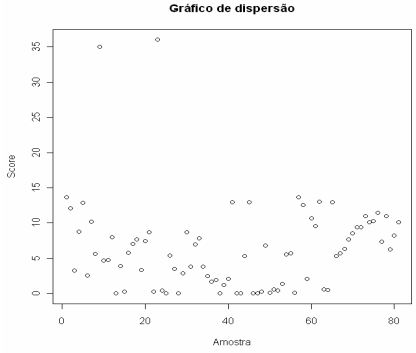

Figura 2 – Gráfico de dispersão das amostras

Conforme a Figura 2, é possível observar que apesar do *outlier* nos valores ficar acima de 35, a maioria dos valores estão compreendidos entre 0 e 15.

Um estudo mais detalhado desses valores foi necessário com o gráfico de uma *Normal Quantil* dos valores das amostras.

Como mostrado na Figura 3, os valores das amostras seguem uma *Normal Quantil* com pequenas "caldas", assim, foi utilizada a média como valor de limiar. Para uma compactação maior, esse experimento utilizou o terceiro quartil com valor médio de 9,41 para determinar a semelhança das assinaturas das imagens.

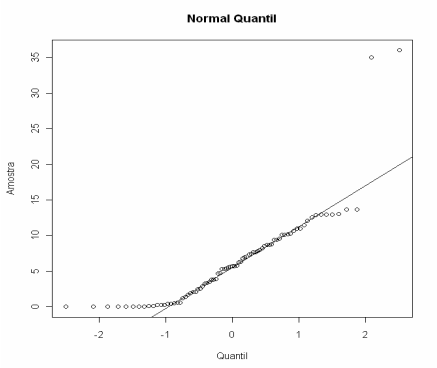

Figura 3 – Normal Quantil.

3. RESULTADOS

Após realizado o procedimento de eliminação de cópias em toda base, vinte imagens foram escolhidas aleatoriamente para averiguar se o processo havia sido executado conforme previsto. Desse modo, foi averiguado se de fato apenas cópias foram eliminadas e se alguma cópia havia sido mantida. Nestes testes, apenas duas imagens apresentaram cópias redundantes.

A base de imagens eliminou aproximadamente 85 mil cópias, o que representa 27% da base original. O tempo médio para consultas diminuiu em cerca de 90ms. A Tabela 1 mostra uma sumarização dos resultados na comparação feita na base de imagens redundante e na nova base de imagens sem cópias.

Tabela 1. Sumarização dos resultados

	Base de imagens com cópias	**Base de imagens sem cópias**
Mínimo	250.0 ms	234.0 ms
1º Quartil	339.5 ms	292.5 ms
Mediana	445.5 ms	360.0 ms
Média	472.9 ms	380.9 ms
3º Quartil	581.8 ms	438.0 ms
Máximo	843.0 ms	735.0 ms

Com base na Tabela 1, é possível comprovar que a eliminação de cópias contribui significativamente na redução do tempo de espera realizado nas buscas de imagens com base em conteúdo visual. O tempo médio de espera foi reduzido em aproximadamente 20%.

4. CONCLUSÕES

A busca de imagens da *web* com base em características visuais pode complementar o atual método de busca baseado em texto enriquecendo as possibilidades de recuperação de informação visual.

A redução de imagens redundantes em uma base de imagens é uma excelente alternativa para melhorar os resultados tanto no aspecto de custo de processamento e armazenamento, quanto na melhoria do tempo de resposta nas buscas.

Os métodos de busca de imagens mostraram-se satisfatórios tanto em termos de desempenho na qualidade das buscas, quanto em velocidade no retorno das respostas. Considerando-se a complexidade apresentada na indexação automática de conteúdo visual na *web* e a recuperação desse tipo de conteúdo, os testes realizados em um ambiente implementado em PC (configuração média comercializada em 2009) foram satisfatórios.

A redução de cópias redundantes na base testada foi de 27% e o tempo médio de resposta nas buscas ficou 20% mais rápido.

REFERÊNCIAS

[1] Rosenberg, Charles, Henry A Rowley. Clustering Billions of Images with Large Scale Nearest Neighbor Search. IEEE WACV, 2007.

[2] Moores, C.N. Datacoding applied to mechanical organization of knowledge. American Documentation 2, p. 20–32, 1951.

[3] Yong, R. *et al*. Image Retrieval: Current Techniques, Promising Directions, and Open Issues. University of Illinois - Illinois and Columbia University - New York, 1998.

[4] Pinkerton, Brian. Finding What People Want: Experiences with the WebCrawler. The Second International WWW Conference Chicago, USA, October 17-20, 1994.

[5] Pant, Gautam; et al.. Crawling the Web. Indiana, USA. Department of Management Sciences, 2004.

[6] Jacobs, Charles E.; Finkelstein, Adam; Salesin, David H. Fast Multiresolution Image Querying. University of Washington - Seattle, WA, 1995.

[7] Cho, J.; Garcia-Molina, H.. Efficient Crawling Through URL Ordering, Seventh International World-Wide Web Conference, Brisbane, Australia, 1998.

[8] Meneses, Thiago Fonsêca. *Guaatupi*: Um ambiente de indexação e recuperação de imagens da *web* com base em características visuais. Trabalho de conclusão de curso apresentado a Universidade Tiradentes, 2009.

[9] Pimentel, C. A. F. *et al*. Um ambiente para indexação e recuperação de conteúdo de vídeo baseado em características visuais. XV WEBMEDIA, 2009.

[10] R. Jain. The Art of Computer Systems Performance Analysis: Techniques for Experimental Design, Measurement, Simulation, and Modeling. Wiley- Interscience, New York, NY, April 1991, ISBN:0471503361.

[11] Stollnitz J. Eric; Derose D Tony; Salesin H. David (1995). Wavelets for computer graphics: A primer. Department of Computer Science and Engineering, University of Washington, 2009.