UNIVERSIDADE FEDERAL DE ITAJUBÁ
PROGRAMA DE PÓS GRADUAÇÃO EM
CIÊNCIA E TECNOLOGIA DA COMPUTAÇÃO

Bruno Crivelari Sanches

# Detecção de Spam em Imagens Usando Redes Neurais Artificias

Dissertação submetida ao Programa de Pós-Graduação em Ciência e Tecnologia da Computação como parte dos requisitos para obtenção do Título de Mestre em Ciência e Tecnologia da Computação

**Área de Concentração:** Sistemas de Computação

**Orientador:** Prof. Dr. Otávio Augusto S. Carpinteiro
**Co-Orientador:** Prof. Dr. Edmilson Marmo Moreira

Agosto de 2014
Itajubá - MG

UNIVERSIDADE FEDERAL DE ITAJUBÁ
PROGRAMA DE PÓS GRADUAÇÃO EM
CIÊNCIA E TECNOLOGIA DA COMPUTAÇÃO

Bruno Crivelari Sanches

# Detecção de Spam em Imagens Usando Redes Neurais Artificias

Dissertação aprovada por banca examinadora em 8 de agosto de 2014, conferindo ao autor título de ***Mestre em Ciências em Tecnologia da Computação.***

**Banca Examinadora:**
Prof. Dr. Otávio Augusto S. Carpinteiro (Orientador)
Prof. Dr. Edmilson Marmo Moreira (Co-Orientador)
Prof. Dr. Carlos Alberto Murari Pinheiro
Prof. Dr. Carlos Henrique Quartucci Forster

Agosto de 2014
Itajubá - MG

# Resumo

O correio eletrônico ou *e-mail* é um dos meios de comunicação mais utilizados na atualidade. No entanto, sua grande popularidade e sua arquitetura tornaram-no alvo de mensagens *spam*. Mensagens *spam* carregam, usualmente, informes publicitários, conteúdos fraudulentos ou maliciosos e são enviadas de forma indiscriminada a muitos usuários sem que estes desejem recebê-las. Acarretam diversos prejuízos aos usuários do sistema de *e-mail* e desperdiçam os recursos de rede das instituições.

Para eliminar estas mensagens, foram criados diversos sistemas anti-spam que analisam o conteúdo textual das mensagens e classificam-nas. Devido ao bom desempenho destes filtros, mensagens *spam* passaram a ocorrer em imagens. Isto tornou inútil o uso de sistemas baseados apenas em análise do conteúdo textual, fomentando, assim, o desenvolvimento dos sistemas anti-spam de imagens.

O processamento de imagens é bem mais custoso computacionalmente que o processamento textual e os resultados dos sistemas anti-spam de imagens têm sido inferiores aos dos sistemas textuais. Outra dificuldade da pesquisa na área de sistemas anti-spam de imagens é devida à pouca disponibilidade de bases de dados públicas, o que dificulta a avaliação de resultados experimentais.

Este trabalho propõe um sistema anti-spam de imagens que faz uso de diversos métodos de extração de características de imagens e de um modelo neural artificial, para a classificação dos e-mails. Os métodos de extração são avaliados de forma individual e de forma combinada. O modelo neural é avaliado de forma exaustiva utilizando-se bases de dados disponíveis publicamente. A utilização destas bases de dados é descrita em detalhes, de forma a facilitar a reprodução dos resultados.

Além de se analisar a capacidade de classificação do sistema proposto, este trabalho avalia seus custos computacionais, incluindo os custos para a extração de características das imagens e para a classificação destas. Os resultados obtidos mostram-se promissores tanto em termos das taxas de classificações corretas e de falsos positivos produzidas pelo sistema anti-spam, quanto em termos de seu custo computacional.

# Abstract

The electronic mail or e-mail is nowadays one of the most frequently employed communication system. However, its popularity and its architecture made it a target for spam messages, which are messages sent indiscriminately to many users without their consent. Spam messages cause many losses to users of the mail system and waste network resources of the institutions.

In order to combat those messages a variety of efficient textual filters were created to analyse the textual context of those messages and classify them. Owing to the good performance of these filters, spammers started using image spam. The use of filters only based in analysis of textual contents became useless and that caused the development of image spam filters.

Image processing is significantly more costly than textual processing. Besides, image spam filters are not achieving the same results as textual filters. Another difficulty of the research on image spam filters is the low availability of public images databases, making it difficult the comparison of experimental results.

This work proposes an image anti-spam system which makes use of various techniques for extracting image features and of an Artificial Neural Network for classifying e-mails. The feature extraction techniques are evaluated individually and combined. The classifier is exhaustively evaluated on only public databases. The use of public databases is described in details to facilitate the reproduction of the results.

Apart from reviewing the performance of the proposed system, this work analyses its computational costs, including the costs for classifying the images and for extracting their features. The results produced by the image anti-spam system are promising, achieving good performance in spam detection and in false positive rates, keeping the computational costs low.

# Sumário

# Lista de Figuras

# Lista de Tabelas

# 1 Introdução

O correio eletrônico (ou *e-mail*) tornou-se um dos meios de comunicação mais utilizados por indivíduos, corporações e pela academia, onde nasceu. No entanto, devido a sua popularidade, surgiram meios de explorá-lo de forma indevida, acarretando prejuízos a seus usuários.

Mensagens indesejadas passaram a circular desde os anos 90 (STERN, 2008). Tais mensagens são enviadas a centenas e até milhares de recipientes sem que sejam solicitadas. Na maioria dos casos, incluem propagandas que promovem serviços, produtos, eventos, dentre outros. Receberam, por alcunha, o nome de spam, nome este que acabou por tornar-se um sinônimo de mensagens indesejadas.

## 1.1 O Termo Spam

O termo *spam* nasceu devido a uma referência ao grupo de comediantes ingleses *Monty Python*. Este grupo criou uma peça, nos anos 70, que envolvia a sigla *SPAM* (*Spiced Pork and Ham*). A peça em questão passava-se em um restaurante, do qual o cardápio consistia quase que totalmente em refeições compostas por *SPAM*. O cardápio era anunciado pela garçonete, que repetia o tempo todo a palavra *SPAM*, ao mesmo tempo em que *vikings* amantes do *SPAM* diziam ao fundo repetitivamente a palavra *SPAM* (CAMPBELL, 1994).

Tal como a garçonete e os *vikings* que repetem exaustivamente a palavra *SPAM*, incomodando as pessoas a sua volta, os usuários de *e-mail* também recebem exaustivamente as mesmas e mesmas mensagens inconvenientes, surgindo assim o termo *spam* no contexto das mensagens eletrônicas, em analogia ao programa de televisão.

O produto *SPAM* é fabricado pela Hormel Foods[1] desde 1930, empresa esta que é contra a associação de sua marca com algo mal visto e que causa tantos transtornos como o *spam*.

[1]Hormel Foods(http://www.hormel.com).

Juntamente com o termo *spam*, surgiu o termo *ham*, que, em uma tradução direta, significa presunto. O termo *ham* refere-se a *e-mails* válidos ou àqueles que o usuário deseja receber. Desta forma, neste trabalho, assim como em outros da área, o termo *spam* significa um *e-mail* indesejado pelo usuário e o termo *ham*, um *e-mail* válido ou legítimo.

## 1.2 Problemas Causados por Spam

Para muitos usuários, o *spam* pode parecer apenas uma mensagem inofensiva, que pode ser rapidamente apagada sem grandes danos. O que passa despercebido para os usuários, porém, é que o grande volume de mensagens acaba gerando problemas, como[2]:

- Não recebimento de *e-mails*: como grande parte dos provedores de Internet limitam o tamanho da caixa postal, quando um usuário recebe um grande número de *spams*, sua caixa postal pode ficar cheia, fazendo com que mensagens recebidas sejam devolvidas ao remetente.
- Aumento de custos: o usuário, de uma forma ou de outra, paga pelo uso da rede. Em alguns casos, paga também pela quantidade de dados transmitidos. Assim, *e-mails* indesejados geram tráfego de rede desnecessário e, por consequência, custo para o usuário.
- Perda de produtividade: usuários que usam *e-mails* profissionalmente gastam parte de seu tempo lendo e organizando mensagens. Com o recebimento de *spams*, usam seu tempo produtivo descartando lixo eletrônico, acarretando custos e perda de produtividade nas instituições.
- Conteúdo impróprio ou ofensivo: *Spams* são enviados de forma indiscriminada para endereços aleatórios de *e-mails*. Assim, é provável que usuários recebam mensagens que considerem ofensivas.
- Prejuízos financeiros causados por fraude: *Spams* têm sido utilizados para divulgar mensagens falsas, de forma a induzir o usuário a acessar páginas falsas na Internet ou a induzi-lo a instalar programas maliciosos, com o objetivo de furtar seus dados pessoais e financeiros. Esta técnica é conhecida como *phishing/scam*.

Não são apenas os usuários finais que são afetados pelo *spam*. Attar et al. (2013) citam três problemas causados pelo grande volume de *spams* circulando em sistemas de

[2]Comitê Gestor da Internet no Brasil - Antispam (http://www.antispam.br), acessado em 15/02/2013

rede e servidores:

- Congestionamento nos serviços de rede e falhas na comunicação de rede.
- Queda na credibilidade e confiança do serviço de *e-mail*.
- Sobrecarga no sistema de armazenamento dos servidores de *e-mail*.

## 1.3 Utilização de Imagens

Com a melhoria dos sistemas de análise textual usados no combate ao *spam*, *spammers* passaram então a utilizar cada vez mais imagens para evitar sistemas baseados em texto. Assim, passaram a inserir os textos das mensagens em imagens, tornando inúteis sistemas baseados apenas em análise do conteúdo textual.

Esta técnica teve início em 2004. Em 2005, já era responsável por 1% dos *spams* enviados e, em 2007, chegou a 65% do total de mensagens indesejadas. Em 2008 e 2009, houve uma queda, com imagens *spam* sendo responsável por aproximadamente 40% dos *spams*. Em 2010, a incidência desta técnica voltou a crescer, chegando a 55% (ATTAR; RAD; ATANI, 2013).

Além de tornar inúteis os filtros que se baseiam apenas em análise textual, imagens trazem um outro problema aos sistemas anti-spam. São computacionalmente mais caras de serem analisadas do que texto, pois são estruturas de dados maiores e mais complexas que simples textos.

Além da maior dificuldade de serem identificados por sistemas anti-spams, *spams* em imagens são também mais atrativos para os usuários. Como diz o ditado popular, "uma imagem vale mais que mil palavras". Na figura 1 pode-se observar um exemplo de imagem *spam*.

Existem diversas definições para imagens *spam*. Por exemplo, alguns autores consideram como sendo imagem *spam* quando a mensagem está em uma imagem anexada ao *e-mail*. Outros consideram quando as imagens são incluídas ou anexadas ao corpo principal da mensagem. Já alguns consideram que imagem *spam* é um tipo de *e-mail spam* onde o texto da mensagem é representado como uma foto em uma imagem. Certos autores consideram como sendo imagem *spam* quando a imagem é um *hyperlink* para uma página desconhecida. Além destas, há outras definições semelhantes às citadas (ATTAR; RAD; ATANI, 2013).


Figura 1: Exemplos de imagens *spam*.

Este estudo visa analisar as imagens contidas nos *e-mails* de forma a classificá-los como *ham* ou *spam*. Os textos contidos nos *e-mails* são desconsiderados. Assim, a análise de imagens *spam* é complementar à análise de *spam* em texto realizada pelos sistemas anti-spam atuais. O problema de como extrair as imagens dos *e-mails* também não é abordado neste estudo.

## 1.4 Métodos para detecção de imagens spam

Com o advento das imagens *spam*, surgiu o desafio de criarem-se técnicas que sejam capazes de separar imagens *spam* de imagens *ham*. A principal dificuldade é de se encontrar uma técnica que obtenha resultados satisfatórios com custo computacional baixo, pois a análise de imagens e as técnicas de processamento de imagem costumam ser mais custosas que técnicas de análise textual. As técnicas existentes podem ser categorizadas em dois grupos principais:

- Análise de características: as características das imagens são extraídas e analisadas. Parte-se aqui do pressuposto de que imagens *spam* possuem características especiais que podem ser detectadas com a análise das imagens.

- Análise OCR (*Optical Character Recognition* ou Reconhecimento Óptico de Caracteres): os textos das imagens são extraídos e analisados.

### 1.4.1 Análise de Características

A análise de características parte do princípio de que imagens *spam* são de alguma forma diferentes de imagens *ham*. Assim, os pesquisadores procuram por características que possam ser usadas para a classificação das imagens.

Um exemplo de característica é o histograma da imagem. Partindo-se do fato de que grande parte das imagens *spam* são artificiais, seus histogramas apresentam perfis diferentes daqueles de imagens naturais ou de fotografias. Dessa forma, com a análise dos histogramas, um sistema pode classificar as imagens em *spam* e *ham*.

A grande dificuldade desta abordagem está em se determinar quais características são as mais relevantes e como estas podem ser analisadas. Cabe notar que um classificador baseado na análise de características pode fazer uso de apenas uma ou de várias características, avaliando-as em conjunto ou de forma individual. A escolha de uma característica pode levar em consideração não apenas sua relevância para a classificação da imagem, mas também seu custo computacional. Assim, um sistema classificador pode usar primeiramente as características com custo computacional mais baixo, de forma a classificar a imagem mais rapidamente.

### 1.4.2 Análise OCR

A análise por OCR supõe que imagens *spam* são usadas para transportar uma mensagem textual. Assim, elas devem conter essencialmente texto que, se extraído da imagem, pode ser analisado por filtros textuais tradicionais. Em alguns casos, as imagens *spam* realmente não passam de uma imagem da versão texto da mensagem. Entretanto, os *spammers* conseguem facilmente dificultar esta análise por OCR com o uso de *captchas*[3], ruídos e outros métodos de ofuscação.

A análise por OCR possui também um custo computacional alto, que aumenta com o emprego de novos e mais sofisticados métodos de ofuscação. Métodos mais sofisticados de ofuscação dão margem, porém, ao uso de um outro método de análise por OCR. Neste método, ao invés de se tentar extrair o texto das imagens, procura-se por sinais de ofuscamento de texto, que seriam um alto indicativo de *spam*. Assim, as técnicas de ofuscamento empregadas pelos *spammers* acabam sendo usadas contra eles próprios.

[3] *Completely Automated Public Turing test to tell Computers and Humans Apart (CAPTCHA)* é um teste automatizado para distinguir entre a atuação de autômatos e a de seres humanos.

### 1.4.3 Falsos Positivos

Os falsos positivos denotam os *e-mails ham* que são erroneamente categorizados como *spam* por sistemas de classificação, ou seja, são *e-mails* legítimos, de interesse do usuário, mas que são classificados como *spam* por sistemas anti-spam.

Esta classificação errônea tem um custo alto para os usuários, posto que estes esperam que o sistema anti-spam livre-os de *e-mails* indesejados, mas não de *e-mails* do seu interesse. Quando ocorre uma classificação equivocada, o usuário vê-se obrigado a vasculhar a pasta de *e-mails spam* do seu sistema em busca de *e-mails* legítimos. Como o sistema obviamente não possui meios de informar ao usuário que isto ocorreu, este precisa criar o hábito de verificar periodicamente sua pasta de *spams* para certificar-se de que nenhum *e-mail* legítimo foi classificado como *spam*.

O problema agrava-se quando a filtragem de *spam* ocorre diretamente no servidor de *e-mails*, onde não é possível ao usuário fazer uma verificação. Assim, são eliminados falsos positivos sem que o usuário seja notificado ou tenha meios de descobrir o que foi perdido.

Inversamente aos falsos positivos, existem os falsos negativos. Os falsos negativos denotam os *e-mails spam* que são erroneamente categorizados como *ham* por sistemas de classificação de anti-spams. Os falsos negativos, quando sua ocorrência é baixa, não são, porém, tão graves, uma vez que os usuários, ao verificarem suas caixas de entrada (*mailboxes*), rapidamente conseguem identificar os *spams* e removê-los.

## 1.5 Base de dados

Outra dificuldade ao se estudar o problema das imagens *spam* é a criação de uma base de dados confiável e a comparação de resultados com outros trabalhos da área (ATTAR; RAD; ATANI, 2013).

Para a criação de uma base de dados, geralmente é fácil realizar-se a coleta de *e-mails spam* provenientes de instituições ou de usuários. Por outro lado, a coleta de *e-mails* legítimos costuma ser difícil, pois usuários e instituições restringem o acesso a eles por possuírem conteúdo particular. Em muitos casos, pesquisadores acabam por coletar imagens em bases públicas, como no *Google Images*, o que acaba gerando representações não muito fiéis às mensagens legítimas que circulam pela rede Internet.

Como consequência da dificuldade citada para montar-se uma base, surge uma outra dificuldade, que é a comparação de resultados de pesquisa. Pesquisadores acabam

formando bases próprias para suas pesquisas. São poucos os estudos reportados na literatura que disponibilizam as bases de dados empregadas e, caso sejam disponibilizadas, são mais raros ainda os estudos que reutilizam estas bases. Assim, em muitos casos, é difícil dizer se os resultados obtidos nos estudos reportados na literatura são devidos a uma técnica ou metodologia mais poderosa, ou se são devidos à base de dados utilizada.

## 1.6 Proposta de Trabalho

O método aplicado neste trabalho consiste na extração de características relevantes de imagens, com base em técnicas de Processamento Digital de Imagens, e na classificação destas imagens, utilizando-se as características extraídas e Redes Neurais Artificiais, para determinar quais imagens são *spam* e quais são *ham*.

A metodologia empregada no estudo consiste em três etapas principais. Na primeira, são selecionadas características em imagens, relevantes à classificação destas. Na segunda etapa, cada característica é empregada, de forma individual, como entrada para a rede neural. Na terceira etapa, duas ou mais características são empregadas.

O estudo fez uso de bases de *e-mails* e imagens disponibilizadas publicamente na *Internet*, possibilitando assim a comparação dos resultados, de forma mais precisa, com outros estudos na área.

### 1.6.1 Conteúdo da Dissertação

Este estudo é dividido conforme descrito a seguir. No capítulo 2, são apresentadas a estrutura de uma mensagem de *e-mail* e de um sistema de *e-mail*, são analisados diferentes tipos de imagens *spam*, e apresentados os conceitos sobre os quais este estudo se fundamenta. No capítulo 3, é apresentada uma revisão de estudos em sistemas de classificação de imagens *spam* (sistemas anti-spam). No capítulo 4, é detalhada a metodologia usada e são apresentados os resultados. Por fim, o capítulo 5 apresenta as conclusões obtidas e sugestões para pesquisas futuras.

# 2 Revisão da Teoria

Neste capítulo, é apresentada a teoria sobre a qual este estudo baseia-se.

## 2.1 Correio Eletrônico

O correio eletrônico ou *e-mail* foi criado antes mesmo da Internet. Os primeiros padrões foram propostos em 1973 com a RFC 561[1]. Com a conversão da ARPANET à Internet, nos primeiros anos da década de 1980, o *e-mail* como é conhecido hoje passou a tomar forma.

Atualmente o formato das mensagens de *e-mails* é definido pela RFC 5322[2]. Esta RFC define a divisão do *e-mail* em duas partes principais:

- cabeçalho (*header*): primeira parte do *e-mail*, organizada em campos.
- corpo (*body*): parte do *e-mail* que contém o texto da mensagem.

As mensagens dos *e-mails* são formadas apenas por textos. Tais textos utilizam caracteres, interpretados como ASCII, com valores no intervalo de 0 a 127. Os textos são divididos em linhas. O final de cada linha é marcado por uma sequência de dois caracteres: *carriage return* (CR) e *line feed* (LF), referidos pela sigla CRLF nas especificações e também ao longo deste trabalho.

O cabeçalho é separado do corpo da mensagem com o uso de uma linha em branco, ou seja, uma linha que contenha apenas os caracteres CRLF.

As linhas de texto de uma mensagem não podem exceder 998 caracteres e é recomendado que não excedam 78 caracteres. O limite de 988 caracteres foi criado de forma a manter compatibilidade com diversas implementações que não foram projetadas para

[1] Standardizing Network Mail Headers (http://tools.ietf.org/html/rfc561).
[2] Internet Message Format (http://tools.ietf.org/html/rfc5322).

lidar com linhas com mais de 1000 caracteres. Já a recomendação de manter as linhas com até 78 caracteres foi criada para que a interface de diversas aplicações não fossem prejudicadas, pois muitas aplicações exibem no máximo 80 caracteres por linha.

### 2.1.1 Cabeçalho de e-mails

O cabeçalho de um *e-mail* é formado por linhas de texto. Cada linha contém um campo. Assim, cada linha no cabeçalho é iniciada pelo nome de um campo, seguido do caractere ":" (dois pontos), do corpo do campo e do par CRLF. O nome do campo deve ser formado apenas por caracteres imprimíveis do padrão US-ASCII, mais precisamente, pelos caracteres entre os valores 33 a 126, exceto pela vírgula. O corpo de um campo pode ser formado pelos mesmos caracteres, além do espaço em branco e tabulação vertical. Na figura 2 é possível observar algumas linhas de cabeçalhos e o formato destas.

```
From: John Doe <jdoe@machine.example>
To: Mary Smith <mary@example.net>
Subject: Saying Hello
Date: Fri, 21 Nov 1997 09:55:06 -0600
Message-ID: <1234@local.machine.example>

This is a message just to say hello.
So, "Hello".
```

Figura 2: Conteúdo de um *e-mail*: seu cabeçalho e corpo da mensagem.

Existem também regras para campos desestruturados (não estruturados) e estruturados. Campos desestruturados são tratados apenas como linhas de caracteres. Campos estruturados possuem regras semânticas definidas e uma lista de *tokens* apropriados.

São definidas também regras para criação de longas linhas de texto nos cabeçalhos, que permitem dividir um longo campo de cabeçalho em múltiplas linhas.

Os campos de um cabeçalho carregam informações, tais como, remetente, destinatário, data de envio, assunto, data, dentre outras. As únicas obrigatórias são a data de origem e remetente.

### 2.1.2 Corpo de e-mails

O corpo das mensagens é simplesmente uma sequência de linhas de texto utilizando apenas caracteres do padrão US-ASCII. Existem apenas duas regras para o corpo da mensagem:

- Os caracteres CR e LF só podem ocorrer juntos, formando o par CRLF.
- As linhas de texto são limitadas a 998 caracteres, embora recomenda-se que sejam limitadas a 78 caracteres, sem contar o par CRLF.

### 2.1.3 Conteúdo não Texto

Como descrito anteriormente, o conteúdo de um *e-mail* é restrito apenas aos caracteres ASCII americanos. Este fato limita muito o conteúdo de uma mensagem de *e-mail*, devido a ausência de caracteres apropriados usados em outras línguas. Além da limitação em relação às diversas línguas existentes, a impossibilidade de se transmitir informações que não sejam texto restringem muito este meio de comunicação.

Para contornar esta limitação e ainda manter compatibilidade com os sistemas existentes, propôs-se e adotou-se o *Multipurpose Internet Mail Extensions*, mais conhecido pela sigla *MIME*. O *MIME* é definido por cinco RFCs principais:

- RFC 2045 [3]: define os diversos cabeçalhos usados em mensagens *MIME*.
- RFC 2046 [4]: define os tipos de mídia suportados.
- RFC 2047 [5]: descreve extensões que permitem o uso de caracteres que não sejam ASCII americanos nos campos de cabeçalho.
- RFC 4289 [6]: define procedimentos para registro de novas codificações de transmissão.
- RFC 6838 [7]: define métodos para o registro de novos tipos de mídia.

Além das RFCs citadas, existem outras que atualizam ou corrigem estas e outras ainda que ficaram obsoletas após a criação das novas RFCs. Por exemplo, a RFC 6838 tornou obsoleta a RFC 4288, que tinha, por sua vez, tornado obsoleta a RFC 2048.

[3]Multipurpose Internet Mail Extensions (MIME) Part One: Format of Internet Message Bodies (http://tools.ietf.org/html/rfc2045).

[4]Multipurpose Internet Mail Extensions (MIME) Part Two: Media Types (http://tools.ietf.org/html/rfc2046).

[5]Multipurpose Internet Mail Extensions (MIME) Part Three: Message Header Extensions for Non-ASCII Text (http://tools.ietf.org/html/rfc2047).

[6]Multipurpose Internet Mail Extensions (MIME) Part Four: Registration Procedures (http://tools.ietf.org/html/rfc4289)

[7]Media Type Specifications and Registration Procedures (http://tools.ietf.org/html/rfc6838)

Em resumo, pode-se dizer que o *MIME* adiciona três funcionalidades principais ao sistema de *e-mails* original:

- Definição de tipos diferentes de conteúdos (*content type*).
- Definição de codificações de transmissão, que são usadas para representar dados binários de 8 bits usando apenas os caracteres de 7 bits do conjunto de caracteres ASCII.
- Regras para codificarem-se caracteres não ASCII em campos do cabeçalho de mensagens.

Além de proporcionar novas funcionalidades, o *MIME* mantem compatibilidade com os sistemas originais. Esta compatibilidade permite, por exemplo, que um aplicativo cliente de *e-mail MIME* consiga interpretar corretamente uma mensagem não *MIME*, além de também permitir que uma mensagem *MIME* contendo textos seja processada corretamente por um cliente que não suporte o padrão *MIME*.

Para se implementar as funcionalidades citadas e manter compatibilidade com o sistema de *e-mails* original, uma mensagem *MIME* é diferenciada das demais pelo campo de cabeçalho "MIME-Version: 1.0". Este campo indica que o *e-mail* contém conteúdo *MIME* versão 1.0, versão que até o momento não foi alterada.

Após o cabeçalho, uma mensagem *MIME* pode conter o conteúdo "comum" de um *e-mail*, ou seja, um corpo de texto formado por apenas caracteres ASCII, conforme regras descritas anteriormente. Este corpo é incluído mesmo quando a mensagem possui texto em outros formatos, como, por exemplo, HTML. Isso permite que clientes de *e-mail* que não reconheçam o padrão *MIME*, possam também exibir uma mensagem legível a seus usuários.

No cabeçalho da mensagem *MIME* também é incluído o campo *Content-Type*, que define o tipo de conteúdo da mensagem. Este campo é opcional e, quando não é incluído em uma mensagem *MIME*, significa que a mensagem contém apenas texto simples, ou seja, apenas os caracteres ASCII.

O campo de cabeçalho *Content-Type* é formado por um tipo e subtipo, como, por exemplo:

- *Content-Type: text/plain*: define o tipo/subtipo como sendo texto simples.

- *Content-Type: image/jpeg*: define o tipo/subtipo como sendo imagem no formato *jpeg*

Novos tipos de conteúdos podem ser criados. Para serem usados, devem ser registrados segundo a norma definida pela RFC 4289. O registro é requisitado não para se dificultar a criação de novos tipos, mas para evitarem-se eventuais colisões de nomes.

Caso o conteúdo a ser transmitido em uma mensagem *MIME* não possa ser representado pelos caracteres ASCII, este conteúdo é então codificado por algum método, como, por exemplo, o método *base64*. Os métodos de codificação suportados definem esquemas para se codificar dados utilizando apenas caracteres ASCII suportados pelo sistema de *e-mail* original. Desta forma, é mantida total compatibilidade com os sistemas existentes. Por exemplo, para indicar que o conteúdo do *e-mail* é codificado no formato *base64*, usa-se o campo de cabeçalho "Content-Transfer-Encoding: base64".

O padrão *MIME* também permite, através da RFC 2045, a inclusão de novos métodos de codificação, que empreguem formatos privados e regras que indiquem como estes devem ser incluídos e diferenciados dos métodos padronizados. Entretanto, esta mesma RFC desencoraja seu uso.

#### 2.1.3.1 Mensagens Multipartes

O padrão *MIME* descrito até o momento resolve o problema da transmissão de informações que não podiam ser representadas pelo conjunto de caracteres ASCII. Além disto, o padrão também possibilita que uma mensagem seja formada não apenas por um conteúdo, mas por múltiplos conteúdos, onde cada um destes conteúdos encontra-se dentro de uma entidade.

Nas mensagens com múltiplos conteúdos, é possível transmitir, por exemplo, conteúdo textual e várias imagens, que podem tanto estar em arquivos anexos ao *e-mail* ou incorporadas diretamente no texto. Por exemplo, o corpo da mensagem, em formato HTML, pode conter uma entidade com o texto HTML e várias entidades com as imagens e outros componentes referenciados pelo conteúdo HTML.

Uma mensagem multiparte segue as mesmas regras, já descritas, de qualquer mensagem *MIME*. A única diferença está na forma como seu conteúdo é interpretado. Além do campo de cabeçalho *MIME-Version*, que indica que a mensagem segue um formato *MIME*, o campo *Content-Type* deve conter o valor *multipart/mixed*, que indica que a mensagem possui muitas partes. O subtipo *mixed* indica que diferentes conteúdos estão

presentes na mensagem.

No subtipo *mixed*, a ordem das entidades é importante e deve ser respeitada pelos clientes de *e-mail*. Assim, sempre que mensagem com o tipo *multipart/mixed* for exibida, o cliente deve exibir cada entidade na ordem em que esta se encontra na mensagem.

Além do subtipo *mixed*, entidades multipartes podem possuir os valores *alternative*, *digest* e *parallel*. No primeiro caso, o subtipo indica que todos os conteúdos das entidades são iguais, diferindo apenas na forma em que são codificados. Este subtipo pode ser usado para fornecer a um cliente de *e-mail* a mesma mensagem em diversos formatos, ficando a cargo do cliente exibir ao usuário aquela mensagem que for mais conveniente. O segundo subtipo, *digest*, é usado para enviar coleções de mensagens. É semanticamente idêntico ao *mixed*, diferindo-se apenas no conteúdo suposto por padrão quando este não é especificado. Por fim, o subtipo *parallel* é semelhante ao *mixed*, diferindo-se apenas pelo fato de que o cliente de *e-mail* não mais precisa respeitar a ordem das entidades.

Para permitir que múltiplas entidades sejam tratadas adequadamente, estas devem ser separadas por um valor que as delimita. Assim, sempre que um campo de cabeçalho *Content-type* multiparte é especificado, além do tipo de conteúdo, é preciso especificar qual o valor de fronteira entre as entidades. O valor de fronteira especifica o inicio e fim de cada entidade, como pode ser visto na figura 3.

Os delimitadores de entidade são gerados pelo agente que compõe a mensagem de *e-mail* e é de sua responsabilidade garantir que o valor delimitador seja único na mensagem e apareça somente como delimitador. Quando usado para separar entidades, o valor delimitador é precedido por dois hífens. Por exemplo, o delimitador da figura 3 é o texto "simple boundary" e, portanto, no corpo da mensagem é sempre exibido como "–simple boundary" (sem aspas).

Uma mensagem multiparte permite, igualmente, que cada entidade também seja multiparte, formando uma estrutura semelhante a uma árvore. Assim, é possível criar *e-mails* com conteúdos bem complexos. O delimitador de cada entidade deve, porém, continuar sendo único em todo o *e-mail*.

```
From: Nathaniel Borenstein <nsb@bellcore.com>
To: Ned Freed <ned@innosoft.com>
Date: Sun, 21 Mar 1993 23:56:48 -0800 (PST)
Subject: Sample message
MIME-Version: 1.0
Content-type: multipart/mixed; boundary="simple boundary"

This is the preamble.  It is to be ignored, though it
is a handy place for composition agents to include an
explanatory note to non-MIME conformant readers.

--simple boundary

This is implicitly typed plain US-ASCII text.
It does NOT end with a linebreak.
--simple boundary
Content-type: text/plain; charset=us-ascii

This is explicitly typed plain US-ASCII text.
It DOES end with a linebreak.

--simple boundary--

This is the epilogue.  It is also to be ignored.
```

Figura 3: Conteúdo de um *e-mail* multipartes, mostrando seus cabeçalhos e corpo da mensagem com dois conteúdos textuais.

### 2.1.4 Transmissão de emails

A transmissão de *e-mails* é feita através do protocolo *Simple Mail Transfer Protocol* (SMTP), definido pela RFC 5321[8]. Usuários geralmente utilizam um aplicativo, denominado cliente, como, por exemplo, o *Outlook Express*, para compor e transmitir suas mensagens. O cliente faz uso do protocolo SMTP para entregar a mensagem a um servidor de *e-mail*, comumente o servidor onde o usuário possui sua caixa postal (*mailbox*). Este servidor, ao receber a mensagem do seu usuário, utiliza o protocolo para transmitir a mensagem ao servidor (ou servidores) de destino da mensagem.

O protocolo *SMTP* se encarrega apenas da transmissão de *e-mails*. Estes ficam armazenados em caixas postais até serem "recuperados" pelo usuário. O destinatário para acessar suas mensagens utiliza um cliente, que acessa a caixa postal e recupera as mensagens, utilizando o protocolo *Post Office Protocol* (POP), definido pela RFC 1939[9], ou o protocolo *Internet Message Access Protocol* (IMAP), definido pela RFC 5301[10].

[8]Simple Mail Transport Protocol (http://tools.ietf.org/html/rfc5321).
[9]Post Office Protocol - Version 3 (http://tools.ietf.org/html/rfc1939)
[10]Internet Message Access Protocol - Version4rev1 (http://tools.ietf.org/html/rfc3501)

#### 2.1.4.1 Protocolo SMTP

O protocolo *SMTP* é utilizado para transporte de *e-mails*. O protocolo define comandos em formato texto, utilizando caracteres ASCII, e é, em grade parte, responsável pelas limitações do formato de mensagens de *e-mail*, como, por exemplo, pelo limitado número de caracteres por linha da mensagem.

Um servidor *SMTP* age tanto como um receptor de mensagens, quando este é o destinatário de um *e-mail*, quanto como um transmissor. O servidor age como um transmissor nos casos em que a mensagem recebida não o tem como destinatário. Neste caso, o servidor encarrega-se de transmitir a mensagem para o servidor de destino, ou para outro servidor transmissor, que vai se encarregar de encaminhar a mensagem para seu destinatário, ou outro transmissor. O processo se repete até que a mensagem atinja seu destino.

Um dos grandes problemas deste protocolo é a ausência de mecanismos fortes de autenticação. Isto permite que usuários mal intencionados possam falsificar a origem de mensagens, que consiste em uma das técnicas comuns para geração de *spam*, como descrito por Stern (STERN, 2008).

#### 2.1.4.2 Submissão de e-mails

Com o constante aumento de dispositivos maliciosos na rede e, consequentemente, de *worms*, *spams*, vírus, e outras formas de conteúdo malicioso, muitos servidores proibiram o envio de *e-mails* através da porta 25 (porta padrão do protocolo *SMTP*). Com o aumento desses dispositivos maliciosos, exige-se dos servidores de envio de mensagens, atualmente, que sejam responsáveis pelo tráfego que geram, exigindo-lhes que o envio de mensagens passe por um processo de autenticação.

Assim, como consequência, o processo de submissão de *e-mails* aos servidores separa-se do processo de transmissão de *e-mails* entre servidores. Esta separação entre submissão e transmissão de *e-mails* é definida pela RFC 6409 [11].

A RFC define que, antes de encaminhar uma mensagem, qualquer servidor deve certificar-se de que todos os domínios encontrados na mensagem sejam qualificados. Caso não o sejam, deve rejeitá-la. O servidor também deve exigir que seja autenticada a conexão pela qual recebeu a mensagem a ser encaminhada. A RFC define formas de autenticação. Além disto, a RFC também sugere que o servidor verifique a sintaxe dos endereços da mensagem, faça um registro (*log*) de erros de configuração do cliente e tenha menor

[11] Message Submission for Mail (http://tools.ietf.org/html/rfc6409)

tolerância para *timeouts*.

#### 2.1.4.3 Protocolo POP

O *Post Office Protocol* (*POP*) foi desenvolvido para permitir a um usuário acessar sua caixa postal (*mailbox*) em um servidor de *e-mail*. Assim, um servidor *SMTP* encarrega-se de receber mensagens ininterruptamente e o protocolo *POP* provê mecanismos para que o usuário transfira, sempre que lhe for conveniente, os *e-mails* de sua caixa postal para sua estação de trabalho. *POP* não foi concebido para manipular grandes volumes de mensagens. Normalmente, é usado apenas para transferir suas mensagens do servidor para sua estação de trabalho. Após a transferência, suas mensagens são removidas do servidor.

Assim como o protocolo *SMTP*, o protocolo *POP* baseia-se em comandos textuais, compostos por caracteres da tabela ASCII e encerrados pelo par CRLF. Seu funcionamento pode ser representado por uma máquina de estados, onde o estado inicial, que se inicia assim que a conexão é aberta, é chamado de autenticação (*authentication*). Neste estado, o servidor envia ao cliente uma "saudação" e, em seguida, aguarda sua autenticação. Ao receber uma autenticação válida, o servidor permite o acesso exclusivo à caixa postal solicitada e transita para o estado de *transação* (*transaction*).

No estado de transação, o cliente pode requisitar informações sobre as mensagens no servidor, requisitar a transferência das mesmas e requisitar que estas sejam removidas. É importante ressaltar que, quando um cliente solicita a remoção de uma mensagem, o servidor apenas marca a mensagem como removida. A remoção propriamente dita ocorre apenas quando o cliente encerra, de forma normal, a conexão.

Quando, por fim, o cliente envia o comando "QUIT", o servidor transita para o estado *atualizar* (*update*). Neste estado, o servidor procede com a remoção efetiva das mensagens e encerra o acesso à caixa postal. Se a conexão for encerrada por outro método que não através do comando "QUIT", o servidor não procede com a remoção das mensagens.

#### 2.1.4.4 Protocolo IMAP

O *Internet Message Access Protocol* (*IMAP*) é um protocolo para acesso a caixas postais em servidores, tal como o protocolo *POP*. Diferentemente deste, porém, o protocolo *IMAP* remove as mensagens do servidor apenas quando o usuário explicitamente solicita.

Além disto, enquanto um cliente *POP* copia todas as mensagens do servidor para a

estação de trabalho do usuário, um cliente *IMAP* copia apenas as mensagens que o usuário solicita, o que é vantajoso no caso de usuários que recebem muitas mensagens. Uma outra diferença ainda em relação ao protocolo *POP* é que o *IMAP* permite que vários clientes estejam conectados e acessando concorrentemente uma mesma caixa postal.

O protocolo *IMAP* possibilita também, no caso de mensagens *MIME*, que o usuário acesse apenas as partes da mensagem que desejar. Por exemplo, ao receber uma mensagem com arquivos anexos, o usuário pode apenas acessar o conteúdo texto, não precisando copiar toda a mensagem para sua estação de trabalho, economizando tempo e recursos computacionais.

O *IMAP* possibilita, igualmente, que as mensagens no servidor tenham informações de estado, como, por exemplo, se foram lidas, respondidas ou apagadas. O estado de cada *e-mail* geralmente é mantido por clientes de *e-mail*. O armazenamento do estado de cada *e-mail* no servidor permite a diferentes clientes de *e-mail* fazerem uso e tirarem proveito desta informação.

Usuários do protocolo *IMAP* também têm a possibilidade de criar múltiplas caixas postais no servidor, renomeá-las ou removê-las facilmente, posto que são apresentadas ao usuário como se fossem pastas. Além de manipular suas caixas postais, é possível ao usuário transferir mensagens entre elas.

Além destas facilidades, o protocolo *IMAP* também oferece um sistema com vários critérios de busca. Possibilita, igualmente, a criação de extensões ao próprio protocolo.

Contudo, há também desvantagens deste protocolo em relação ao protocolo *POP*. Por exemplo, a implementação computacional mais complexa, maior consumo de recursos computacionais de servidores, dentre outras.

Assim como os protocolos *SMTP* e *POP*, o protocolo *IMAP* baseia-se em comandos textuais, compostos por caracteres da tabela ASCII e encerrados pelo par CRLF.

### 2.1.5 Transmissão de Imagens em e-mails

A transmissão de imagens em *e-mails* pode ser realizada por diferentes formas, devido à variedade de recursos oferecidos pelo padrão *MIME*. Usualmente, contudo, imagens são enviadas em arquivos anexos ou incluídas diretamente no corpo das mensagens.

Independentemente da forma utilizada para transporte de imagens, este trabalho parte do pressuposto de que um sistema anti-spam usual, utilizado para a detecção de *spam* tex-

tual, identifica *e-mails* que possuam imagens. Assim, envia-as a um sistema especializado em reconhecimento de imagens *spam* para análise. O resultado da análise é então enviado ao sistema classificador original, que toma as decisões cabíveis sobre o que fazer com a mensagem.

Assim, o sistema especializado em reconhecimento de imagens *spam* aqui proposto atua como um aplicativo que estende a funcionalidade dos sistemas anti-spam textuais já existentes. Desta forma, este trabalho tem como escopo analisar apenas a questão da classificação de imagens, não abordando a integração deste classificador aos diversos sistemas anti-spam textuais existentes.

## 2.2 Tipos de Imagens spam

Com a precisão cada vez maior das técnicas de detecção de *spam* voltadas para a análise textual dos *e-mails*, *spammers* passaram a utilizar imagens para transportarem suas informações e assim evitar os filtros textuais existentes.

Imagens *spam* possuem diversos conteúdos de propaganda, como descrito abaixo (ATTAR; RAD; ATANI, 2013).

- adulto: imagens oferecendo serviços para público adulto, como pornografia, serviços de encontros, anúncios pessoais, dentre outros;
- financeiro: imagens com ofertas para serviços financeiros, mercado de ações, oportunidades de investimentos. Tais imagens podem estar incluídas, junto com mensagens fraudulentas, em *e-mails* que almejam conseguir senhas ou informações bancárias de seus destinatários;
- produtos: imagens que oferecem os mais variados produtos e serviços.
- Internet: imagens que ofertam serviços de computação ou da Internet, como, por exemplo, registro de domínios, hospedagem e serviços de *marketing*;
- lazer: ofertas de prêmios, viagens, cassinos, dentre outras;
- saúde: ofertas de suplementos alimentares, de produtos esportivos, farmacêuticos (figura 4), de produtos para prevenção de doenças, serviços médicos, dentre outras;
- político: propagandas política, de produtos relacionados a partidos políticos, solicitações de doação financeira a políticos e partidos;

- educação: ofertas de cursos em instituições de ensino privadas;
- religião: propaganda de serviços religiosos, de entidades religiosas, mensagens de evangelização.

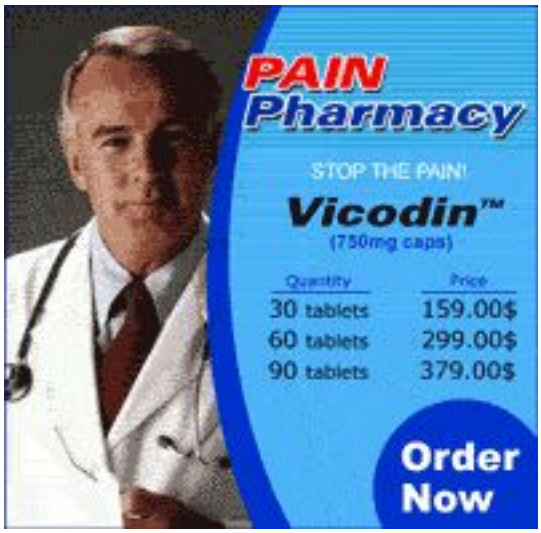


Figura 4: Exemplo de imagem *spam* oferecendo produto farmacêutico.

### 2.2.1 Artimanhas usadas em imagens spam

*Spammers* utilizam as mais variadas técnicas para confundir os sistemas de detecção de imagens *spam*. Algumas destas técnicas são descritas a seguir (ATTAR; RAD; ATANI, 2013).

#### 2.2.1.1 Ofuscação

Ofuscação é uma das técnicas mais antigas e conhecidas. Consiste em usar palavras com erros de ortografia, girar levemente textos, adicionar sombreamentos, ocultar bordas e adicionar ruídos aleatórios. Na figura 5, é mostrado um exemplo de imagem utilizando esta técnica.

#### 2.2.1.2 Conteúdo Textual

Esta técnica emprega imagens compostas apenas por texto, como, por exemplo, na figura 6. Apesar de seu alto custo computacional, técnicas de OCR podem ser empregadas para o reconhecimento de *spam* nestas imagens.

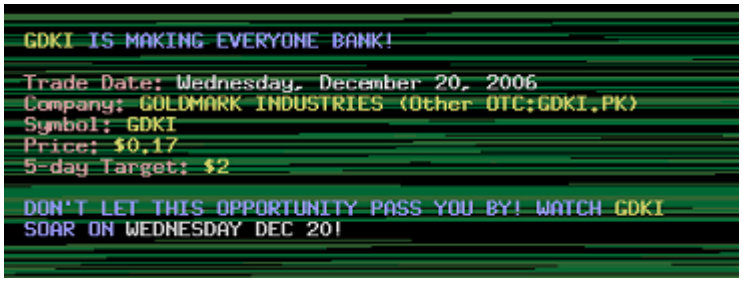


Figura 5: Exemplo de imagem *spam* com ofuscação.

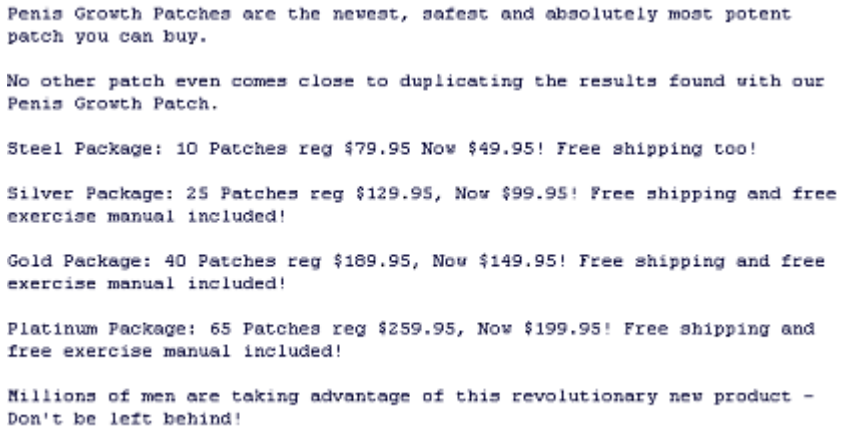


Figura 6: Exemplo de imagem *spam* apenas com textos.

#### 2.2.1.3 Seções Múltiplas

Nesta técnica, a imagem é dividida em múltiplas seções, o que dificulta sua análise. O cliente de *e-mail*, porém, remonta as seções e exibe a imagem corretamente. A imagem pode ser seccionada de forma aleatória. Na figura 7, é mostrado um exemplo da técnica.

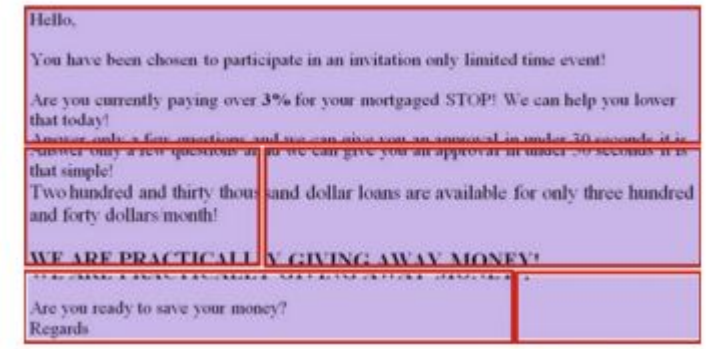


Figura 7: Exemplo de imagem *spam* dividida em múltiplas seções.

#### 2.2.1.4 Imagens com escrita cursiva

O uso de escrita cursiva, como pode ser observado na figura 8, é uma técnica utilizada para confundir sistemas baseados em OCR, pois a escrita cursiva não é semelhante a nenhuma fonte comumente usada.


Figura 8: Exemplo de imagem *spam* com escrita manual.

#### 2.2.1.5 Fundo "Selvagem"

Esta técnica consiste na adição de fundos vagos e estranhos às imagens. São utilizadas figuras geométricas e grande variação de cores. A técnica é usada para confundir técnicas de OCR, pois estas buscam nas imagens padrões geométricos similares a letras. A figura 9 é um exemplo de *spam* utilizando esta técnica.


Figura 9: Exemplo de imagem *spam* com uso de fundo "selvagem".

#### 2.2.1.6 Variação Geométrica

Esta técnica consiste na alteração da imagem somente, mas não da mensagem que ela contém. A figura 10 apresenta um exemplo desta técnica.

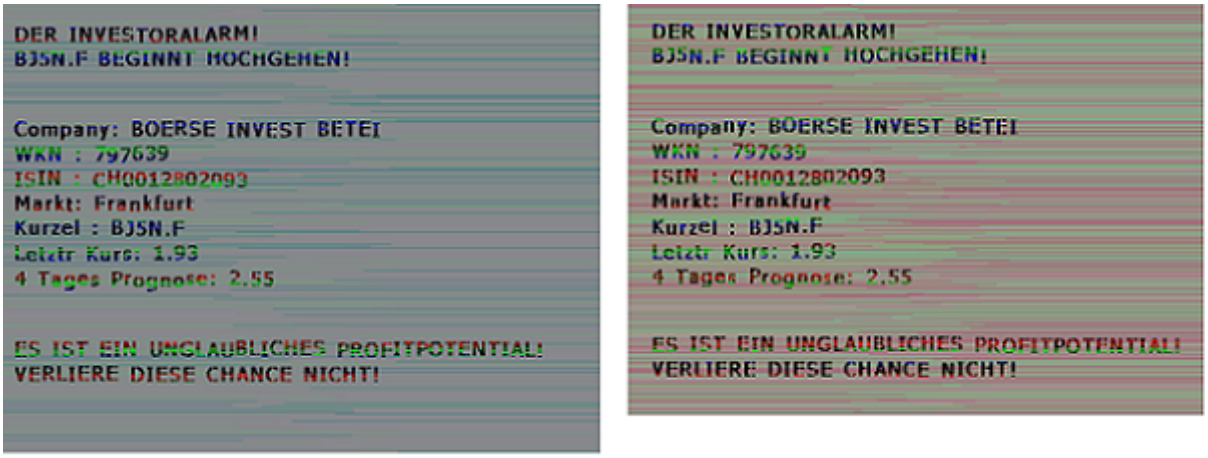


Figura 10: Exemplo de imagem *spam* com variação geométrica. As mensagens são idênticas, mas usam diferentes fundos e tamanhos.

#### 2.2.1.7 Imagens Gif Animadas

O formato GIF (*Graphics Interchange Format*) é muito usado, na Internet, tanto para codificar imagens estáticas, quanto para exibir imagens animadas. A animação é utilizada pelos *spammers* para dificultar a análise e classificação das imagens. A figura 11 apresenta um exemplo de imagem animada.

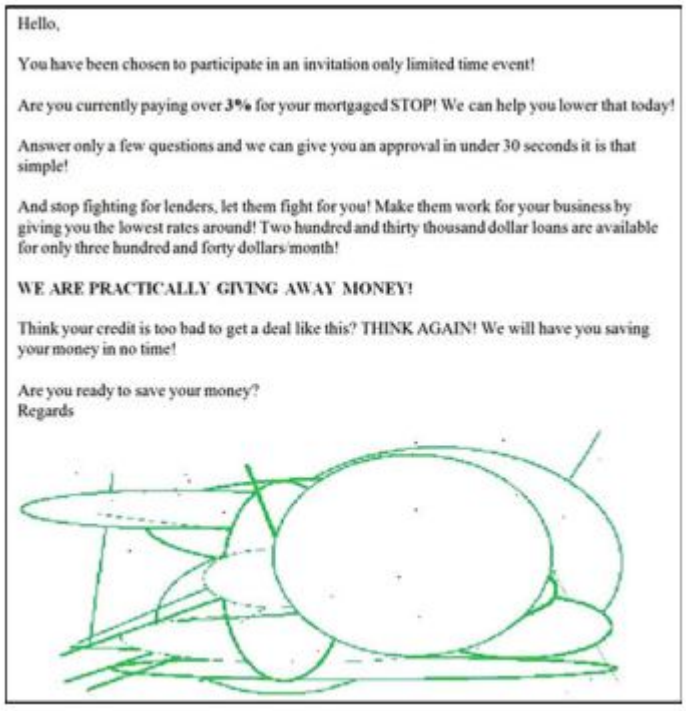


Figura 11: Exemplo de imagem *spam* utilizando um GIF animado.

#### 2.2.1.8 Desenho Animado ou Cartoon

Esta técnica consiste no uso de fontes estilo *cartoon*, como mostrado na figura 12. Este estilo de fonte produz uma imagem bonita e atrativa para seres humanos mas de difícil análise para sistemas computacionais.


Figura 12: Exemplo de imagem *spam* utilizando imagem com fontes estilo de "desenho animado".

#### 2.2.1.9 Recolorir

Esta técnica consiste na variação dos atributos de uma imagem, como, por exemplo, de suas cores. Esta técnica pode iludir diversas técnicas de análise de imagens, tais como análise de histogramas, de *pixels*, e de cor média. A figura 13 apresenta um exemplo desta técnica.


Figura 13: Exemplo de duas imagens *spam* idênticas, mas usando cores diferentes.

## 2.3 Extração de características

Para a extração de características, foi desenvolvido um aplicativo, em linguagem Java, que processa grupos de imagens, gerando um arquivo, no formato CSV,[12] com as características de cada imagem. Neste trabalho, o aplicativo foi empregado sobre cada uma das bases de imagens para gerar os arquivos correspondentes de cada base.

Um arquivo CSV é formado por campos separados por vírgulas. Cada linha do arquivo representa um registro e, opcionalmente, o arquivo pode possuir um cabeçalho. Caso exista, o cabeçalho, apresentado na primeira linha do arquivo, contém os nomes de cada um dos campos.

O aplicativo desenvolvido gera os arquivos CSV sempre em um mesmo formato. O número de campos, porém, é variável, pois depende da quantidade de características que cada método de processamento digital de imagens gera. O primeiro campo é o nome do arquivo que contém a imagem. Após este campo, existe um campo para cada valor de cada característica. Por fim, o último campo de cada registro contém a classe — *ham* ou *spam*— da imagem. A inclusão do campo com a classe da imagem é necessária tanto para o treinamento da rede neural MLP quanto para a avalição das classificações por ela

[12]Common Format and MIME Type for Comma-Separated Values (CSV) Files (http://tools.ietf.org/html/rfc4180)

realizadas.

A figura 14 apresenta um exemplo de arquivo CSV gerado pelo aplicativo desenvolvido. A primeira linha contém o cabeçalho, indicando o nome de cada campo. Depois, cada linha representa uma imagem. A primeira coluna contém o nome do arquivo que contém a imagem. A segunda, terceira e quarta colunas contêm os valores de três características — os valores médios dos canais RGB de cor. Por fim, a última coluna indica a classe da imagem.

```
filename,R,G,B,class
trec\ham\inmail.10293.0.png,0.15459333333333247,0.12501764705882298,0.13325647058823475,HAM
trec\ham\inmail.10293.1.png,0.0,0.0,0.0,HAM
trec\ham\inmail.10293.2.png,0.800000000000127,0.20000000000003176,0.4000000000000635,HAM
trec\ham\inmail.10293.3.png,0.8160552941177726,0.43549529411761734,0.5661705882353137,HAM
trec\ham\inmail.10293.4.png,0.35786352941175587,0.31822705882351976,0.26909490196076546,HAM
trec\ham\inmail.10293.5.png,0.9701960784313914,0.868365098039218,0.902205098039266,HAM
trec\ham\inmail.10293.6.png,0.5197768627450957,0.45255019607840696,0.46885333333330925,HAM
trec\ham\inmail.10293.7.png,0.4605996078431324,0.338519607843116l,0.25920941176469564,HAM
trec\ham\inmail.10293.8.png,0.9931447058823535,0.9912772549019612,0.9919019607843136,HAM
trec\ham\inmail.10295.0.png,0.7977776470588833,0.8204454901961219,0.8553376470588457,HAM
trec\ham\inmail.10295.1.png,0.9250674509804488,0.9487474509804342,0.9728427450980612,HAM
trec\ham\inmail.10295.2.png,0.8309996078431855,0.8392588235294594,0.8516301960784581,HAM
trec\ham\inmail.1031.0.png,0.09748823529411722,0.09817529411764674,0.11080078431372381,HAM
```

Figura 14: Trecho de um arquivo CSV usado durante este estudo.

O aplicativo extrai as características das imagens segundo cada um dos métodos de processamento digital de imagens listados a seguir.

- Cor Média: cálculo dos valores médios dos canais RGB das imagens;
- Histograma: geração do histograma em tons de cinza das imagens;
- Histograma Colorido: geração do histograma colorido das imagens;
- Momento de Cor: cálculo dos momentos de cor das imagens;
- Vetor de Coerência de Cor: geração do vetor de coerência de cor das imagens.

O aplicativo desenvolvido faz uso de algumas funções da biblioteca ImageJ[13] para a extração das características das imagens.

### 2.3.1 Imagens Digitais

Como imagens digitais possuem forma retangular, podem ser representadas como uma matriz de duas dimensões, onde cada elemento desta matriz contém o valor numérico da

[13]ImageJ (http://rsbweb.nih.gov/ij/)

cor de um ponto da imagem. Cada ponto da imagem é chamado *picture element* e comumente abreviado por *pixel*.

Conforme descrito por Burger e Burge (2008) uma imagem $I$ é, formalmente, uma função de duas dimensões, com coordenadas $\mathbb{N} \times \mathbb{N}$, que mapeia um intervalo de valores $\mathbb{P}$, tal que

$$I(u, v) \in \mathbb{P} \text{ e } u, v \in \mathbb{N} \tag{2.1}$$

O tamanho de uma imagem é definido pela largura $M$ (número de colunas de *pixels*) e altura $N$ (número de linhas de *pixels*) da matriz $I$.

A resolução de uma imagem indica seu nível de detalhe. A resolução é definida pelo número de *pixels* por unidade de medida, como, por exemplo, pontos por polegada (*dots per inch* ou dpi). Na maioria dos casos, a resolução usada é a mesma tanto para linhas, quanto para colunas da imagem.

#### 2.3.1.1 Sistema de Coordenadas em Imagens Digitais

Para se saber em qual posição na imagem situa-se um *pixel*, é necessário definir um sistema de coordenadas para imagens. Ao contrário das coordenadas cartesianas, o sistema de coordenadas para processamento digital de imagens usa o eixo $y$ invertido, ou seja, a coordenada $y$ vai do topo da imagem ao seu rodapé. A origem do sistema de coordenadas de uma imagem encontra-se no canto superior esquerdo.

#### 2.3.1.2 Valores de Pixels

Um *pixel* contém uma informação da imagem na localização específica onde situa-se. Esta informação que o *pixel* contém pode consistir tanto em um valor numérico quanto em um conjunto de valores numéricos. Em imagens binárias ou em imagens em tons de cinza, por exemplo, a informação é um simples número binário de tamanho $k$, definindo uma faixa de $2^k$ diferentes valores.

Em imagens em cores RGB, a informação consiste em um conjunto de três números binários de tamanho $k$ que indicam a intensidade dos três canais de cores do *pixel*. O canal $R$ representa a cor vermelha (*red*), o canal $G$ representa a cor verde (*green*) e o canal $B$ representa a cor azul (*blue*). A mistura dessas três cores resulta na cor do *pixel*.

#### 2.3.1.3 Imagens Binárias

Em imagens binárias, cada *pixel* contém apenas um *bit* de informação. Assim, cada *pixel* representa ou a cor branca ou preta. Imagens binárias são usadas para representar gráficos simples, documentos arquivados, codificação de transmissões de fax, dentre outros usos.

#### 2.3.1.4 Imagens em Tons de Cinza

Imagens em tons de cinza possuem um único canal de cor. Usualmente, cada *pixel* de uma imagem em tons de cinza contém $k = 8$ bits (1 byte), definindo valores de intensidade no intervalo [0...255]. O valor *0* representa o brilho mínimo (cor preta) e o valor *255*, o brilho máximo (cor branca). Um exemplo de imagem em tons de cinza é apresentado na figura 15.


Figura 15: Exemplo de imagem colorida e imagem em tons de cinza.

#### 2.3.1.5 Imagens Coloridas

Imagens coloridas são usualmente compostas pelas cores vermelho, verde e azul, que formam o sistema RGB (do inglês *Red Green Blue*). A intensidade de cada cor é indicada por um valor inteiro de 8 bits. Assim, cada *pixel* contém $3 \times 8 = 24$ bits para se armazenar os valores das intensidades dos três componentes de cor. Cada componente possui valores no intervalo [0...255].

As imagens coloridas são majoritariamente representadas pelo formato RGB. Existem, porém, outros formatos, tais como o CMYK. O formato CMYK possui quatro canais

de cor — ciano, magenta, amarelo e preto (do inglês *Cyan-Mangenta-Yellow-black*). É comumente usado em sistemas de impressão.

### 2.3.2 Armazenamento de Imagens

Existem diferentes formatos para arquivos de imagens digitais. Alguns destes formatos permitem a compressão das imagens.

A compressão de imagens pode ser feita com algoritmos que gerem perda de dados. Com o uso destes algoritmos, não é possível restaurar as imagens originais. No entanto, um algoritmo bem elaborado é capaz de comprimir a imagem de forma que as perdas sejam imperceptíveis ou quase imperceptíveis visualmente.

Além dos *pixels*, arquivos de imagens digitais contêm, igualmente, um cabeçalho. No cabeçalho, localizado geralmente no início do arquivo, são encontradas as informações básicas sobre a imagem, tais como, dimensões, formato de cores, tipo de compressão, dentre outras.

Os formatos mais populares de imagens são TIFF, GIF, PNG e JPEG. Os três primeiros não perdem informação pelo processo de compressão. Os formatos são descritos a seguir.

- *Tagged Image File Format (TIFF)*: Este formato de arquivo armazena tanto imagens em tons de cinza quanto coloridas. Permite diversos esquemas de compressão e espaços de cores. É bem flexível e comumente usado com aplicações profissionais de imagens.

- *Graphics Interchange Format (GIF)*: Foi desenvolvido em 1986 para ser usado em transmissões *dial-up* de *BBS (Bulletin Board System)*. É um dos formatos mais usados para armazenamento de imagens na Internet. Provê suporte para animações simples.

- *Portable Network Graphics (PNG)*: Foi criado para substituir o formato GIF devido a problemas com a licença do algoritmo de compressão LZW, usado pelo GIF. O formato permite apenas compressão sem perdas. Imagens em formato PNG têm funcionalidades equivalentes ou superiores ao GIF, com exceção apenas de que o formato PNG não provê suporte a animações.

- *JPEG*: Padronizado pelo padrão ISO IS-10918, é o formato de imagem mais utilizado. A compressão permitida por este formato pode alcançar a taxa de 1:25. O

formato JPEG define apenas o método de compressão, pois o formato de imagem é definido por outros padrões, tais como o JFIF e o *Exchangeable Image File Format* (EXIF), desenvolvido especialmente para câmeras digitais.

Além dos formatos citados, existem outros formatos de imagens. Contudo, aplicações que lidam com imagens preocupam-se com os detalhes de cada formato apenas ao ler ou gravar imagens em arquivos. O aplicativo lê a imagem do arquivo, descomprime-a e seus *pixels* são postos em uma área de memória. Além dos *pixels*, o aplicativo armazena outras informações, tais como, dimensões da imagem, formato de cores usado, tamanho da informação contida nos *pixels*, dentre outras. Assim, um aplicativo de processamento digital de imagens, ao manipular imagens em memória, utiliza uma estrutura de dados genérica, que não depende do formato utilizado pelo arquivo da imagem.

### 2.3.3 Processamento Digital de Imagens

São utilizados métodos de processamento digital para a extração das características das imagens. Os métodos de processamento digital de imagens empregados neste trabalho são descritos a seguir.

#### 2.3.3.1 Cor Média

A cor média consiste no cálculo da média de cada canal de cor. Como são somente utilizadas imagens em formato RGB neste trabalho, o cálculo da cor média consiste no cálculo da média de cada um dos três canais de cor do formato. O cálculo da média de cada canal de cor é descrito como

$$E_i = \sum_{j=1}^{N} \frac{1}{N} p_{ij} \tag{2.2}$$

onde $E_i$ é a média do canal de cor $i$ e $N$, o número total de *pixels* da imagem, calculado através da multiplicação do número de linhas da imagem pelo número de colunas. O valor $p_{ij}$ indica o valor do canal $i$ do *pixel* $j$.

#### 2.3.3.2 Histograma de imagens em tons de cinza

Histogramas são utilizados para mostrar informações estatísticas de imagens, em um formato visual de fácil interpretação. Informações obtidas através do histograma podem

ser usadas para melhorar a aparência de uma imagem.

Um histograma é uma distribuição de frequências. Portanto, o histograma de uma imagem descreve a frequência da intensidade de seus *pixels*. Como mencionado anteriormente, a intensidade de um *pixel* é caracterizada por um valor numérico. O histograma $h$ de uma imagem em tons de cinza $I$, com valores de *pixels* no intervalo $I(u, v) \in [0, K-1]$, contém $K$ valores (BURGER; BURGE, 2008). Assim, uma típica imagem em tons de cinza com valores de *pixel* com tamanho de 8 bits contém $K = 2^8 = 256$ valores.

O histograma de uma imagem informa, portanto, a frequência com que ocorre, nos *pixels* da imagem, cada valor possível de *pixel*. Formalmente, para cada valor de *pixel* $i$,

$$h(i) = card\{(u, v) \mid I(u, v) = i\} \tag{2.3}$$

onde *card* significa o número de elementos (cardinalidade) do conjunto. Assim, o valor de $h(0)$ é o número de *pixels* com valor 0, $h(1)$, o número de *pixels* com valor 1 e, assim sucessivamente. Por fim, o valor de $h(255)$ indica o número de *pixels* brancos, ou seja, com valor de intensidade máximo $255 = K - 1$. O resultado do cálculo do histograma é um vetor de comprimento $K$. A figura 16 mostra uma imagem *spam* e seu histograma (em tons de cinza), calculado pela biblioteca *ImageJ*.

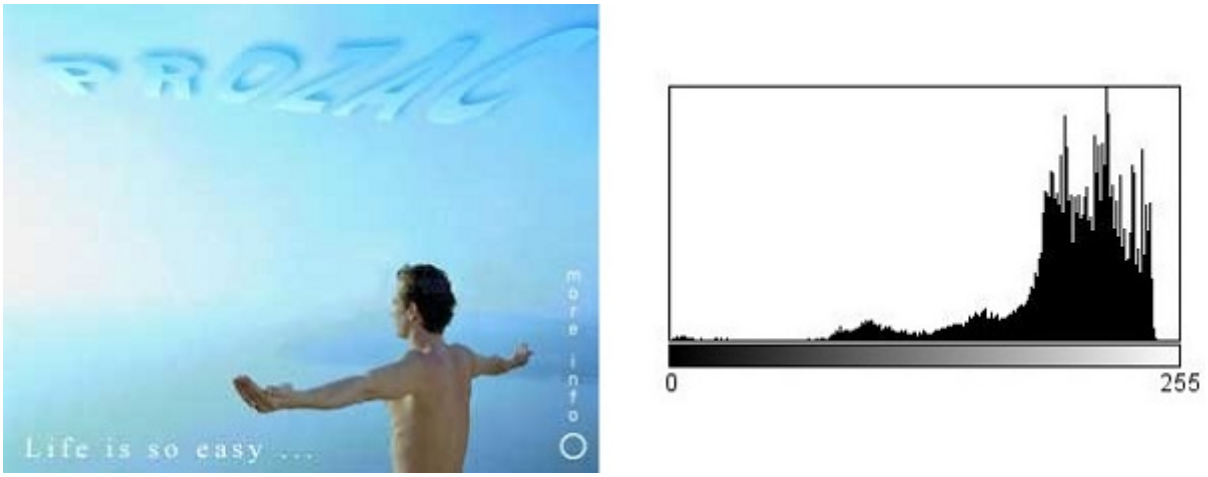


Figura 16: Exemplo de imagem *spam* e seu histograma, calculado pela biblioteca *ImageJ*.

Um histograma não fornece informações espaciais a respeito da localização das cores na imagem, ou seja, não descreve onde as intensidades estão localizadas, apenas a distribuição estatística destas.

O histograma em tons de cinza, também conhecido como histograma de intensidades é referenciado neste trabalho apenas como "histograma". No caso de um histograma não

ser em tons de cinza, este possuirá outra denominação, como, por exemplo, “histograma colorido”.

#### 2.3.3.3 Histograma de imagens coloridas

O cálculo do histograma de uma imagem colorida é similar ao do histograma de uma imagem em tons de cinza. Contudo, a quantidade de valores que um histograma de uma imagem colorida contém é usualmente bem maior. Por exemplo, o histograma de uma imagem no formato RGB, com tamanho de *pixel* de 8 bits por canal de cor, contém $2^{24}$ valores.

Qu e Zhang (2009) reduzem a dimensão do histograma ao reduzir a quantidade de cores da imagem. Segundo os autores, a redução de cores da imagem traz adicionalmente a vantagem de ser eficaz contra ruídos.

A redução de cores é realizada ao descartarem-se os bits menos significativos de cada valor de *pixel* da imagem (PASS; ZABIH; MILLER, 1996). Neste trabalho, os histogramas coloridos foram calculados sobre imagens reduzidas para apenas 512 cores. Esta redução foi realizada através do descarte dos 5 bits menos significativos de cada canal de cor dos *pixels* de uma imagem RGB. Assim, o tamanho dos valores dos *pixels* de uma imagem RGB é de 3 bits por canal de cor e seu histograma contém 512 valores. A figura 17 mostra a redução, para 512 cores, da imagem original mostrada na figura 16.

Figura 17: Exemplo de imagem *spam* reduzida para 512 cores.

#### 2.3.3.4 Momento de Cor

O uso do momento de cor tem por base a suposição de que a distribuição de cores em uma imagem segue uma distribuição probabilística. A teoria descreve que uma distribuição probabilística é caracterizada por seus momentos centrais. Stricker e Orengo (1995) propõe o uso do primeiro, segundo e terceiro momentos centrais de cada canal de cor, para a obtenção do momento de cor das imagens.

O primeiro momento consiste na média de cada canal de cor. O cálculo da média de cada canal de cor é realizado por meio da equação 2.2, descrita na seção 2.3.3.1. O segundo momento consiste no desvio padrão de cada canal de cor e o terceiro, na obliquidade (assimetria) de cada canal de cor.

O segundo momento é definido como

$$\sigma_i = \sqrt{(\frac{1}{N}\sum_{j=1}^{N}(p_{ij} - E_i)^2)} \tag{2.4}$$

onde $\sigma_i$ é o desvio padrão para o canal de cor $i$.

O terceiro momento é definido como

$$s_i = \sqrt[3]{(\frac{1}{N}\sum_{j=1}^{N}(p_{ij} - E_i)^3)} \tag{2.5}$$

onde $s_i$ representa a obliquidade do canal de cor $i$.

Devido à utilização do formato de cor RGB neste trabalho, o cálculo do momento de cor de cada imagem gera nove valores, pois são gerados a média, desvio padrão e obliquidade para cada canal de cor.

#### 2.3.3.5 Vetor de Coerência de Cores

O vetor de coerência indica o nível de agrupamento de *pixels* de uma mesma cor em uma região da imagem. Estas regiões com *pixels* de cores iguais são chamadas de coerentes. O vetor de coerência, ao contrário do histograma, leva em consideração a distribuição espacial dos *pixels*. Fornece, portanto, informações adicionais às do histograma.

O cálculo do vetor de coerência é realizado em dois passos (PASS; ZABIH; MILLER, 1996). No primeiro, é feita uma redução da imagem para $n$ cores distintas, em um

processo idêntico ao descrito na seção 2.3.3.3 para a obtenção de seu histograma colorido.

No segundo passo, é verificado se os *pixels* de cada cor são coerentes ou incoerentes. Um *pixel* coerente compartilha de um grande grupo de *pixels* com a mesma cor, enquanto um *pixel* incoerente não. Os grupos de *pixels* são identificados através do cálculo dos componentes conectados.

Um componente conectado $C$ é um conjunto com o maior número de *pixels* da mesma cor onde, para quaisquer dois pixels $p, p' \in C$, existe um caminho em $C$ entre $p$ e $p'$. Um caminho em $C$ é uma sequência de *pixels* $p = p_1, p_2, \ldots, p_n = p'$ na qual cada *pixel* $p_i$ está contido em $C$ e quaisquer dois *pixels* sequenciais $p_i$,$p_{i+1}$ são adjacentes um ao outro.

Após o cálculo dos componentes conectados ser finalizado, cada *pixel* pertencerá a exatamente um componente conectado. Assim, é possível classificar-se cada *pixel* da imagem como coerente ou incoerente, dependendo do tamanho, em *pixels*, de seu componente conectado. Um *pixel* é coerente se seu componente conectado tem tamanho superior a um valor fixo $\tau$. Caso contrário, é incoerente.

Uma imagem possui, usualmente, *pixels* coerentes e incoerentes para qualquer uma de suas cores. O par $(\alpha_j, \beta_j)$, no qual $\alpha_j$ e $\beta_j$ são, respectivamente, o número de *pixels* coerentes e incoerentes da cor $j$, é chamado de par coerente. Portanto, o número total de *pixels* de uma determinada cor $j$ é dado por $\alpha_j + \beta_j$.

O vetor de coerência de cores consiste em um vetor onde suas coordenadas são os pares coerentes de cada cor possível definida pelo tamanho em bits dos *pixels* da imagem. Assim, o vetor de coerência de cores $VC$ é dado por $VC = \langle(\alpha_1, \beta_1), \ldots, (\alpha_n, \beta_n)\rangle$.

Neste trabalho, tal como no estudo realizado por Pass et al. (1996), são utilizadas 64 cores $n = 64$ e o valor de $\tau$ é fixado em 25. Com estes valores de $n$ e $\tau$, o vetor de coerência de cores tem, como coordenadas, 64 pares coerentes, no formato $(\alpha_j, \beta_j)$. A figura 18 mostra uma imagem e seus componentes conectados realçados.

Figura 18: Imagem (à esquerda) e seus componentes conectados (à direita).

### 2.3.4 ImageJ

O pacote de *software* *ImageJ* foi utilizado neste trabalho. O *ImageJ* foi criado por Wayne Hasband e está disponível em seu sítio oficial[14]. Foi desenvolvido no U.S. National Institute of Health (NIH) e é usado por pesquisadores em diversos laboratórios ao redor do mundo, especialmente para processamento de imagens relacionadas à biologia e medicina (BURGER; BURGE, 2008).

O pacote *ImageJ* possui uma biblioteca, escrita em linguagem Java, e uma interface gráfica (figura 19) para acesso direto aos métodos da biblioteca. Os métodos da biblioteca podem, igualmente, ser acessados através de programas escritos em linguagem Java. A *ImageJ* fornece diversos recursos para processamento digital de imagens, bem como permite que seja facilmente estendida com métodos adicionais.

Apesar de suas facilidades para o rápido desenvolvimento de aplicações e realização de experimentos, a biblioteca deixa a desejar quanto ao desempenho. Durante os estudos realizados, observou-se que o custo para o carregamento de cada imagem é muito alto. Este alto custo parece ser causado pelo fato de que o *ImageJ*, ao carregar uma imagem, gera diversas informações sobre a mesma, tais como histograma e outras estatísticas.

[14]ImageJ (http://rsb.info.nih.gov/ij/)

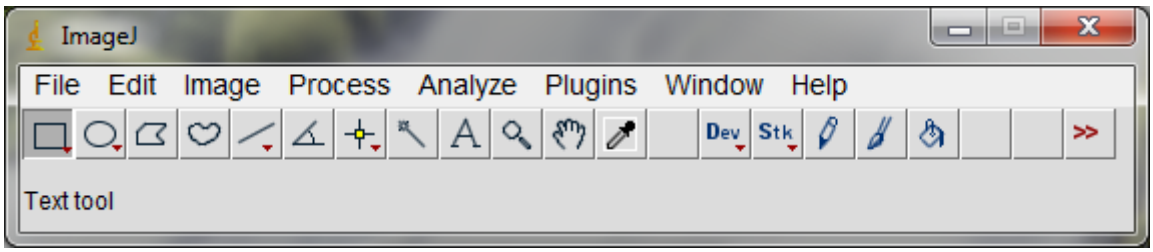


Figura 19: A interface gráfica do pacote ImageJ.

O alto custo computacional não impede a realização de experimentos e a avaliação dos métodos de processamento de imagens. Entretanto, este custo pode impedir um uso do pacote *ImageJ* em aplicações de tempo real. Por exemplo, o alto custo computacional do pacote pode inviabilizar seu uso em um sistema anti-spam de imagens executando em um servidor de *e-mails*.

## 2.4 Redes Neurais Artificiais

Uma Rede Neural Artificial é um modelo matemático cujo comportamento procura simular a operação de neurônios biológicos. A Rede Neural Artificial é composta por um conjunto de neurônios artificiais, interconectados de forma similar à interconexão dos neurônios de um sistema nervoso. As Redes Neurais Artificiais originaram-se em 1943, quando Warren McCulloch e Walter Pitts publicaram um trabalho sobre neurônios artificiais.

Um neurônio biológico possui um corpo celular, chamado de Soma. A partir do Soma, surgem diversos filamentos, chamados de dentritos, e uma estrutura longa, chamada de axônio. Na extremidade do axônio, existem conexões aos dentritos de outros neurônios. Estas conexões são chamadas de sinapses. Através das sinapses, o axônio de um neurônio envia sinais elétricos a outros neurônios. A figura 20 apresenta um modelo simplificado de um neurônio.

Sinais são transmitidos de um neurônio a outro através de reações físico-químicas. Substâncias químicas transmissoras são liberadas nas sinapses e entram nos dentritos, aumentando ou reduzindo o potencial elétrico do corpo celular. Quando o potencial elétrico chega a um valor limiar, o neurônio dispara um pulso elétrico por seu axônio. Este pulso atinge outras sinapses e, através destas, outros neurônios. Sinapses sofrem mudanças de estado em resposta a padrões de estímulos recebidos. Acredita-se que formem a base para o aprendizado de um cérebro (RUSSEL; NORVIG, 1995).

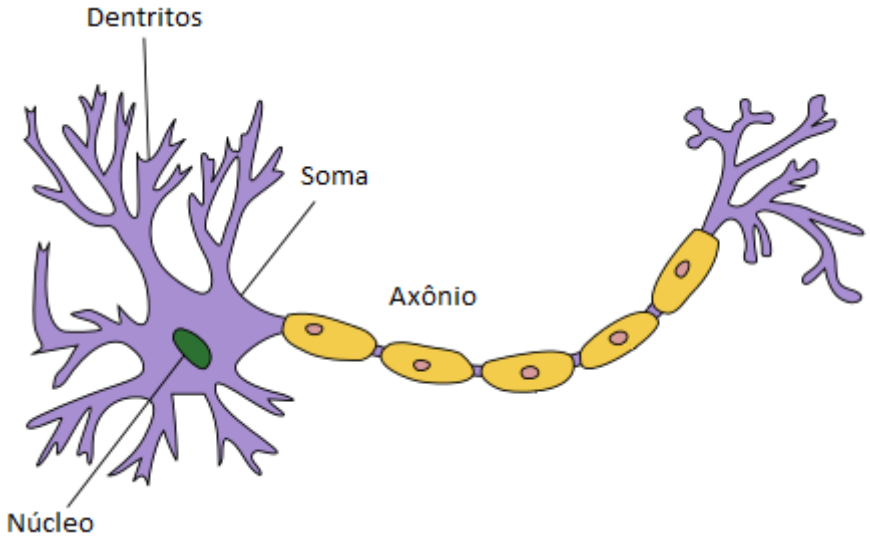


Figura 20: Modelo simplificado de um neurônio.

### 2.4.1 Neurônio Artificial

Em uma Rede Neural Artificial, a unidade fundamental de processamento é o neurônio artificial, também conhecido como *nó*. Os nós são conectados uns aos outros através de conexões ponderadas. O aprendizado de uma Rede Neural Artificial realiza-se através do ajuste dos pesos destas conexões.

Cada nó possui um nível de ativação, conexões de entrada e conexões de saída. Através de suas conexões de entrada, um nó recebe os níveis de ativação de outros nós e, de suas conexões de saída, transmite seu nível de ativação a outros nós. Cada nó computa seu nível de ativação, tendo por base as conexões com os axônios de seus vizinhos, sem que seja necessário um controle global sobre todo o conjunto de nós da rede (RUSSEL; NORVIG, 1995). A figura 21 mostra a estrutura de um neurônio artificial.

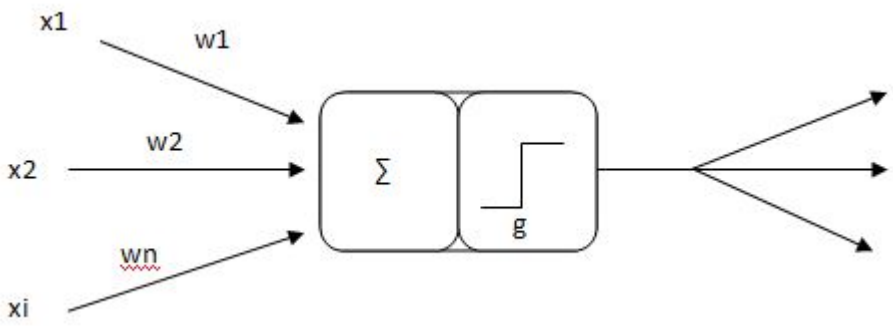


Figura 21: Neurônio Artificial.

O nível de ativação $a_k$ de um nó $k$ é dado por

$$a_k = g(\sum_{i=1}^{n} x_i w_{ki} + b_k) \tag{2.6}$$

onde $g$ é a função de ativação, $x_i$ é o nível de ativação recebido do nó de entrada $i$, $w_{ki}$ é o peso da conexão entre os nós $i$ e $k$, e $b_k$, o *bias* do nó $k$.

Diferentes funções de ativação podem ser usadas nos nós da rede. As principais são as funções degrau, linear, sigmóide e tangente hiperbólica. Estas funções são definidas, respectivamente, pelas equações 2.7, 2.8, 2.9, 2.10, e apresentadas nas figuras 22, 23, 24 e 25.

$$f(x) = \begin{cases} 1, x \geq t \\ 0, x < t \end{cases} \tag{2.7}$$

$$f(x) = ax \tag{2.8}$$

$$f(x) = \frac{1}{1 + e^{-x}} \tag{2.9}$$

$$f(x) = \frac{2}{1 + e^{-2x}} - 1 \tag{2.10}$$

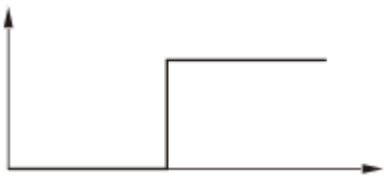

Figura 22: Gráfico da função degrau.

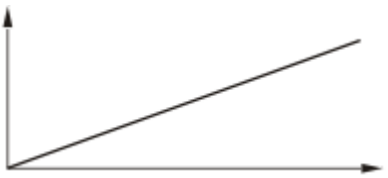

Figura 23: Gráfico da função linear.

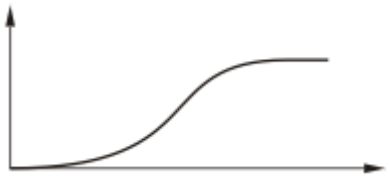

Figura 24: Gráfico da função sigmóide.

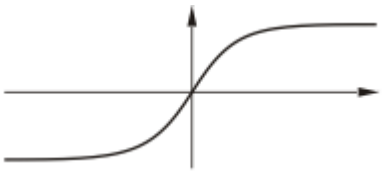

Figura 25: Gráfico da função tangente hiperbólica.

### 2.4.2 Redes MLP

Redes *multi-layer-perceptron* (MLP) são Redes Neurais Artificiais que possuem neurônios dispostos em camadas — uma camada de entrada, uma ou mais camadas intermediárias, chamadas de camadas escondidas, e uma camada de saída. A figura 26 apresenta a arquitetura de um MLP.

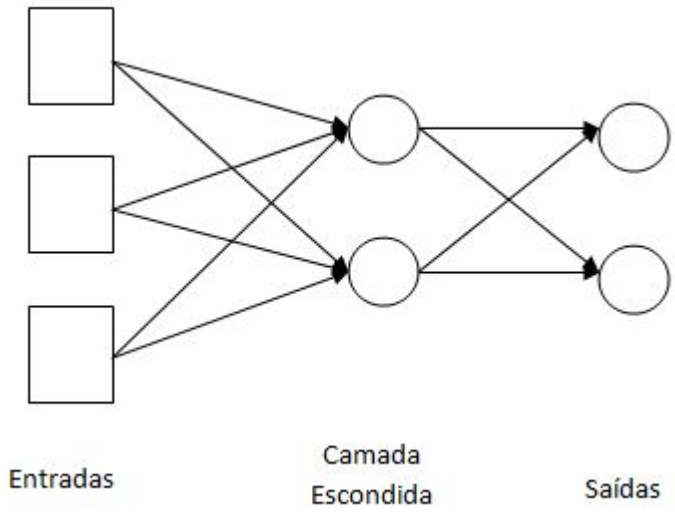


Figura 26: Arquitetura de rede MLP, com uma camada escondida.

*Back-propagation* é o mais popular algoritmo utilizado para treinamento de um MLP. O treinamento é feito de forma supervisionada. O algoritmo possui dois passos — um para a frente e outro para trás. No primeiro passo, é apresentado à rede MLP um padrão de treinamento. A saída produzida pela rede é então comparada com a saída esperada para aquele padrão. No segundo passo, é calculado o erro da rede. O erro é passado,

de forma retroativa, para a correção dos pesos das conexões da rede. O algoritmo é chamado de *back-propagation* devido ao fato de que, durante o treinamento, a correção dos pesos ocorre de forma retroativa, ou seja, dos neurônios de saída para os de entrada (RUMELHART; HINTON; WILLIAMS, 1985).

# 3 Revisão Bibliográfica

Este capítulo apresenta uma revisão da literatura existente sobre detecção de imagens *spam*.

Wu et al. (2005) utilizam uma base específica que contém imagens reais. As imagens *spam* foram extraídas da base SpamArchive e as imagens *ham* foram obtidas de voluntários. Com as imagens desta base, constroem dois conjuntos — treinamento e teste. O conjunto de treinamento possui 8500 *e-mails spam*. O conjunto de teste possui 1500 *e-mails spam* e 428 *e-mails ham*. Os autores empregam três métodos para identificar e selecionar características das imagens. O primeiro, consiste na detecção de textos na imagem. O segundo método consiste na detecção de *banners* e gráficos. O terceiro método detecta se a imagem é anexada ao *e-mail* ou apenas referenciada nele através de um *link*. Para a classificação dos *e-mails*, os autores utilizam uma SVM (*Support Vector Machine*) de uma classe, que apenas classifica *e-mails spam*. O classificador SVM, em seu melhor resultado, obteve uma taxa de 84,6% de acertos na classificação de *spams*, com o uso combinado dos três métodos de seleção de características. Embora os autores afirmem que o custo computacional do classificador é baixo, por ser executado no cliente de *e-mail*, o custo computacional é, em verdade, alto, pois o classificador gasta entre 2 a 3 segundos para classificar um *e-mail* com quatro imagens.

Aradhye et al. (2005) fazem uso de três métodos de seleção de características. O primeiro, consiste na comparação do tamanho das regiões de texto em relação ao tamanho da imagem. O segundo, consiste no cálculo da saturação de cor, conforme definido por Swain, Frankel e Athitsos (1996), nas regiões que incluem ou não textos. O terceiro método consiste no cálculo da homogeneidade da cor nas regiões que incluem ou não textos. Os autores utilizam SVMs para a classificação das imagens. A base de imagens utilizada é composta por 1742 imagens *spam*, coletadas de *e-mails* locais, e 1486 imagens *ham*, coletadas no *Google Image*. As imagens são divididas em quatro categorias temáticas — fotografia, bebês, gráficos/mapas e *screenshots*. Os resultados são divididos de acordo com cada categoria de imagem, variando de 1% a 27% e 73% a 87%, respectivamente, nas

taxas de falso positivo e de detecção de *spam*.

Fumera et al. (2006) propõem o uso de técnicas de OCR para a extração de textos em imagens. Devido ao custo computacional destas técnicas, os autores propõem seu uso, porém, como última alternativa para a classificação das imagens. A possibilidade de ofuscação dos textos, utilizada para confundir as técnicas de OCR, é ignorada pelos autores. Após sua extração, o texto é classificado por um sistema anti-spam textual tradicional. Os autores utilizam duas bases de *e-mails*— uma pessoal e outra do *SpamArchive*. Nos estudos realizados, os autores obtiveram taxas de falsos positivos e negativos inferiores a 2% e 20%, respectivamente.

Drezde et al. (2007) propõem a análise e seleção incremental das características das imagens. O classificador usa, em ordem sequencial, cada característica selecionada até que consiga estabelecer a classificação da imagem. A sequência das características, definida pelos autores, segue um critério de custo computacional. Assim, as primeiras e últimas características na sequência são as que possuem, respectivamente, o menor e maior custo computacional para serem extraídas das imagens. Com isto, os autores esperam reduzir o custo da classificação. Algumas das características selecionadas pelos autores são o tamanho e formato da imagem, cor média, saturação de cor, arestas, cor predominante e cores aleatórias. A base de *e-mails* empregada nos estudos é composta por *e-mails* pessoais e por *e-mails* de uma base do *SpamArchive*. São analisados três classificadores: *Maximum Entropy*, naive Bayes e uma árvore de decisão ID3. Para treinamento e teste de cada sistema classificador são usados, respectivamente, 80% e 20% dos *e-mails* da base. A precisão dos resultados obtidos varia de 88% a 98%.

Qu e Zhang (2009) propõem um sistema que utiliza a classificação prévia de outros sistemas. Se o sistema classificar uma imagem como *spam*, suas características são extraídas, agrupadas em um vetor de características e armazenadas em uma base com vetores de características de imagens *spam*. Caso contrário, o sistema compara o vetor de características da imagem com os vetores armazenados na base. Se for similar a algum vetor da base, o sistema classifica a imagem como *spam*. Os vetores são compostos por três tipos de características — histograma colorido, Haar wavelet e histograma orientado. A base de imagens utilizada nos estudos é formada por 1071 imagens *spam*, coletadas de contas particulares de *e-mail*, por 100.000 imagens *ham*, coletadas de *web sites* e CDs de fotos, e por outras 10 milhões de imagens *ham*, coletadas do *Flickr*[1]. O sistema proposto pelos autores detecta *spams* com uma taxa de 96,7% de acertos e produz uma taxa de

[1]Flickr - Photo Sharing (https://www.flickr.com/).

falso positivos de 0,006%.

Byun et al. (2007) propõem a separação dos *e-mails* em cinco categorias — texto simples, texto complexo, sem texto, não sintética com conteúdo sexual e não sintética sem conteúdo sexual —, de acordo com seus conteúdos. Quatro características das imagens foram selecionadas e analisadas — momento de cor, heterogeneidade de cor, conspicuidade (*conspicuousness*) e auto similaridade (*self-similarity*). Os autores relatam a dificuldade que encontraram para construir uma base com *e-mails* legítimos e para definir uma base com *e-mails spam* que fosse utilizada como padrão nos estudos reportados na literatura. Decidem, por fim, usar, para treinamento do classificador *MFoM-based*, uma base com 2461 imagens, composta por imagens *spam* extraídas do *SpamArchive* e por imagens *ham* extraídas de CDs do *Corel Draw* e do *Google Images*. O classificador possui regras específicas para determinar em qual das cinco categorias situa-se cada um dos *e-mails*. Para teste do classificador, os autores utilizam a base TREC 2005 (CORMACK; LYNAM, 2005), composta por 1249 imagens *spam* e 288 *ham*. Sem a separação dos *e-mails* em categorias, o classificador foi capaz de classificar corretamente 84,6% dos *e-mails*, com uma taxa de 25% de falso positivos. Com a separação dos *e-mails* em categorias, o classificador obteve taxas de 86,6% e 19,1% de classificações corretas e falso positivos, respectivamente.

Mehta et al. (2008) propõem duas soluções para a detecção de imagens *spam*. A primeira consiste no uso de uma SVM para classificação das imagens. A segunda solução consiste no uso de *Gaussian mixture models* para detectar, de forma probabilística, imagens similares. Os autores partem do pressuposto, portanto, de que uma imagem *spam* pode pertencer a um *cluster* de imagens com características similares. As características selecionadas das imagens são o RGB médio, histograma de cor, momento de cor, vetor de coerência de cor, auto correlação, frequência de arestas, comprimento de primitivas, matriz de co-ocorrências, momento geométrico, excentricidade, momentos de *Legendre* e *Zernique* e a existência de caracteres na imagem, detectada através de OCR. Uma base com 13000 imagens foi empregada nos estudos. Esta base é composta por imagens provenientes do *SpamArchive*, da base criada por Drezde (DREDZE; GEVARYAHU; ELIAS-BACHRACH, 2007), e por imagens pessoais. As taxas de classificação correta das imagens, nos estudos realizados, variam de 95% a 99,6%, sendo esta última taxa obtida com o redimensionamento da resolução das imagens para 400x400 *pixels*.

Hayati e Potdar (2008) analisam técnicas de detecção de *spam* textual e em imagens. Separam estas técnicas em categorias e discutem seus prós e contras. Contudo, a qualidade e a eficácia destas técnicas não são avaliadas. Além das técnicas, abordam também os

motivos que induzem *spammers* a gerarem *spam*, tais como geração de renda, promoção de produtos e serviços, roubo de informações e *phishing*.

Biggio et al. (2008) propõem uma técnica para detecção de ofuscamento em imagens com texto, baseada no nível de ruído das imagens. Através do estudo de imagens, observaram que imagens *ham* possuem nível de ruído maior que o de imagens *spam*. Observaram, igualmente, que os níveis de ruído de imagens *spam* e *ham* situam-se em intervalos distintos. Três características são selecionadas das imagens com texto — presença de pequenos fragmentos, de grandes fragmentos, e de grandes formas de fundo que interceptam caracteres de texto na imagem. Como classificadores de *e-mails*, usaram SVMs e árvores de decisão. Duas bases de imagens são usadas. A primeira, contendo 2.006 *hams* e 3.297 *spams*, é composta por *e-mails* pessoais. A segunda base contém as mesmas imagens *ham* da primeira base e 8.549 imagens *spam*, coletadas pelos autores. Seus resultados são comparados com os dos estudos de Aradhye et al. (2005) e de Dredze et al. (2007). Os autores concluem que a técnica proposta merece ser mais detalhadamente avaliada.

Stern (2008) analisa, detalhadamente, três aplicações para geração de *spam*— *Dark Mailer*, *Send Safe* e *Reactor Mailer*. Estas fazem uso de padrões para geração do conteúdo e cabeçalho dos *e-mails*. As três aplicações geram *e-mails spam* com cabeçalhos idênticos aos de clientes de *e-mail* populares, bem como geram-nos com mensagens diferentes, dificultando seu reconhecimento pelos sistemas anti-spam. A tecnologia por trás destas aplicações inclui o uso de *zumbis* e de procedimentos para iludir servidores SMTP.

Zhen et al. (2009) desenvolvem um sistema de classificadores de imagens que faz uso do *Analytic Hierarchy Process* (AHP). Selecionam diversas características das imagens, tais como o formato dos arquivos que as contêm, tamanho, quantidade de cores, heterogeneidade, momento de cor, saturação, RGB médio, dentre outras. A base de imagens usada nos estudos é a mesma usada por Biggio et al. (2008). Cada classificador do sistema é treinado com 80% das imagens da base, em diferentes permutações. A precisão dos classificadores variou de 74% a 97%, dependendo das imagens usadas no treinamento e do classificador.

Liu et al. (2010) propõem um sistema de três camadas para detecção de *spam*. Na primeira camada, são analisados, através de um filtro Bayesiano, os cabeçalhos dos *e-mails*. *E-mails* suspeitos de serem *spam* são encaminhados para a segunda camada. Na segunda camada, é usado um classificador SVM, que classifica os campos do cabeçalho das imagens (seção 2.3.2). Dependendo dos resultados destas duas camadas, o sistema ou decide a classe do *e-mail* ou encaminha-o para a terceira camada. Na última camada,

são extraídos o histograma colorido e o momento de cor da imagem. Uma SVM classifica então o *e-mail*, baseando-se nestas duas características. Os autores empregam quatro bases de *e-mails* nos estudos — TREC (versões 2005 e 2007), Sansone (formada apenas por *e-mails spam*), Drezde (descrita em (DREDZE; GEVARYAHU; ELIAS-BACHRACH, 2007)) e *Image Spam Hunter*, que contém imagens *spam* e imagens do Flickr[2] como imagens *ham*. Os treinamentos são realizados com 50% dos *e-mails* das bases. As taxas de falsos positivos variam de 17,06% a 18,01%, dependendo da combinação das bases empregadas no estudo. A detecção de *spam* varia de 94,93% a 96,8%. Os resultados também indicam que 93,7% dos *e-mails* são classificados na primeira camada, 4,7% na segunda camada, e apenas 1,6% na última camada.

Cheng et al. (2010) propõem um sistema anti-spam que aborda a subjetividade de classificação das *gray images*. *Gray images* são imagens que podem ser classificadas tanto como *ham* quanto como *spam*, dependendo da avaliação subjetiva de usuários de *e-mails*. O sistema é composto por dois estágios. No primeiro, as imagens são classificadas como *ham* ou *spam*, sem considerar as preferências dos usuários. No segundo estágio, as etiquetas de cada imagem *spam* são determinadas. Após este estágio, as imagens *spam* que possuem etiquetas são enviadas aos usuários interessados em recebê-las. Para representar as imagens, os autores empregam a *high-order local autocorrelation* (CHENG et al., 2008). A classificação, nos dois estágios, é realizada pelo algoritmo *k-nearest neighbor* (KNN). A base de imagens é a mesma utilizada por Drezde et al. (2007) em seus estudos. A acurácia do sistema na classificação dos *e-mails* alcança 96,31%. Utilizando as etiquetas, o sistema alcança acurácia de 89,42%.

Soranamageswari e Meena (2010) utilizam uma Rede Neural Artificial para classificar as imagens. Usam duas características — histograma colorido e média de blocos — para representar as imagens. Utilizam, respectivamente, 4000 e 1000 imagens do SpamArchive no treinamento e teste da Rede Neural Artificial. A Rede Neural Artificial alcança uma acurácia de 92,82%, com o uso do histograma colorido, e 89,39%, com a média de blocos.

Wang et al. (2010) selecionaram metadados e características visuais das imagens para representá-las. Dentre os metadados e características selecionados, encontram-se a dimensão, formato de cor, tipo de arquivo, momento de cor, número de cores, número de diferentes cores, cores primárias, saturação de cor e textura. A textura foi obtida por intermédio do momento do histograma em tons de cinza. Uma SVM é utilizada para a classificação das imagens. A base de imagens contém 12.483 imagens *spam* e 3.171 imagens

[2]Flickr Photo Sharing (http://www.flickr.com/)

*ham*, extraídas da base usada por Drezde em seus estudos (DREDZE; GEVARYAHU; ELIAS-BACHRACH, 2007), do *SpamArchive*, do Google Images[3], bem como de uma base pessoal. A proporção de imagens usadas no treinamento e teste da SVM não é reportada pelos autores. A precisão da SVM na detecção de *spam* varia de 95% a 97%. A taxa de falsos positivos varia de 1% a 3%, dependendo da base de imagens usada.

Attar et al. (2013) faz uma revisão dos tipos de imagem *spam*, dos métodos usados para confundir sistemas anti-spam, dos modelos usados para detectar imagens *spam*, e das bases de imagens, descritos na literatura. Destaca as dificuldades encontradas para construção de uma base de imagens pública e confiável, tais como, manter a privacidade das informações e ser representativa do universo de *e-mails* que circulam na Internet. O autor conclui que a detecção de imagens *spam* requer o uso de modelos flexíveis, capazes de adaptarem-se às constantes variações nas características das imagens, realizadas pelos *spammers* com o intuito de confundir os sistemas anti-spam existentes.

Al-Duwairi et al. (2011) analisam o desempenho de alguns classificadores, tais como, árvore de decisão, *support vector machine*, *Naïve Bayes* e *random forest*. Selecionam características visuais, tais como, histograma, gradiente, matriz de co-ocorrências, transformação *wavelet*, dentre outras, para representação das imagens. Os autores utilizam *principal component analysis* para reduzir a dimensionalidade gerada pelo uso de muitas características. Utilizam três bases de imagens — a base usada por Drezde em seus estudos (DREDZE; GEVARYAHU; ELIAS-BACHRACH, 2007), a *Image Spam Hunter*, e uma base pessoal, contendo 810 imagens *ham*, selecionadas aleatoriamente do Flicker[4], e 926 imagens *spam*, extraídas de *e-mails*. Os resultados obtidos variam com a base de imagens e o classificador usados, situando-se entre 93% a 99% e 0% a 2% para as taxas de detecção de *spams* e falsos positivos, respectivamente.

Soranamageswari e Meena (SORANAMAGESWARI; MEENA, 2011) propõem o uso do histograma do gradiente para representação de imagens. Antes do cálculo de seu gradiente, cada imagem é processada através de um filtro, para remoção de ruídos, e redimensionada para 256x256 *pixels*. O histograma é dividido em 5 valores. Os autores empregam uma Rede Neural Artificial como classificador. A base de imagens possui 3.209 imagens *spam* e 1878 imagens *ham*, extraídas de *e-mails* do Spam Archive[5]. O treinamento é realizado com o uso de 80% das imagens. A Rede Neural Artificial alcança uma acurácia de 93,7%.

Hazza e Aziz (2012) propõem um modelo classificador com treinamento não supervi-

[3]Google Images (http://www.google.com/imghp)
[4]Flicker (www.flicker.com)
[5]Spam Archive (http://spamarchive.org/)

sionado. Utilizam o espaço de cor HSL e os picos de luminância para representação das imagens. Destacam a importância do componente L (luminância) pois consideram que imagens criadas em computador são mais luminosas que imagens naturais. Os picos de luminância das imagens são identificados através de seus histogramas. A base de imagens é formada por 500 imagens naturais, obtidas na Internet, e 350 imagens *spam*. Os resultados indicam uma taxa de falsos positivos e de detecção de imagens *spam* de 7% e 94,86%, respectivamente. O processamento computacional de cada imagem leva, em média, 0,67 segundos.

Dhanaraj e Karthikeyani (2013) fazem uma revisão dos tipos de imagem *spam*, dos métodos de ofuscação e dos modelos e métodos usados em sistemas anti-spam, descritos na literatura. Abordam a insegurança dos protocolos da Internet, a rápida reatividade dos *spammers*, bem como aspectos legais que possam coibir a disseminação de *spams*. Abordam, igualmente, de forma superficial, mecanismos para avaliação de desempenho de sistemas anti-spam, tais como o uso de bases de dados comuns e de métricas padronizadas.

## 3.1 Considerações

Na literatura estudada observa-se que a utilização de OCR tem se tornado algo incomum, mas seu uso ainda existe, sendo que é mais frequentemente utilizado de forma "indireta". Este uso "indireto" é observado em situações onde procura-se a presença de texto nas imagens, mas não se tenta fazer a interpretação do mesmo.

Percebe-se também que não existe um consenso entre as formas de se avaliar a qualidade de um classificador, pois alguns pesquisadores se baseiam em taxa de detecção, outros na precisão, alguns na taxa de falsos positivos e etc. Isso torna difícil a comparação de resultados. Os trabalhos aqui analisados, são descritos utilizando-se as mesmas métricas empregadas por seus autores.

Sobre a qualidade dos resultados dos classificadores é possível observar que alguns trabalhos possuem excelentes níveis de classificação, mas em muitos casos são ausentes informações precisas sobre desempenho computacional. Em outros casos observa-se que a proporção entre as bases de testes e de treinamento são mais favoráveis ou menos favoráveis ao classificador.

Sobre as bases de dados, os estudos analisados deixam claro que ainda existe uma grande dificuldade em se construir bases de dados confiáveis, especialmente bases de *e-mails ham*. Nos trabalhos observa-se que é comum o uso de *e-mails* pessoais e de parti-

culares envolvidos na instituição de pesquisa e o compartilhamento destes é incomum.

Os trabalhos aqui analisados foram selecionados com base na sua relevância em relação a este trabalho de pesquisa e principalmente para que fosse possível entender e avaliar as formas mais utilizadas de detecção de *spam* em imagens.

Este estudo procura evitar alguns dos problemas descritos anteriormente. Primeiramente são utilizadas apenas bases de *e-mails* públicas (disponibilizadas na internet), de forma a facilitar a reprodução dos resultados. Com o mesmo intuito, o uso das bases de dados, treinamento do sistema e coleta dos resultados são descritos de forma detalhada. Outro diferencial é que a proporção entre base de testes e treinamento não é tão favorável ao classificador, o que possibilita avaliar a capacidade de generalização do mesmo. Por fim, é feita uma avaliação do desempenho computacional do sistema de forma precisa, que permite avaliar se seu uso em um ambiente de produção é viável.

# 4 Estudos Realizados

Este capítulo descreve a base de dados utilizada, os estudos realizados, bem como os resultados obtidos.

## 4.1 Base de Dados

São utilizadas quatro bases de dados neste trabalho. Três destas bases — Drezde, Prag, Trec —, provenientes de diferentes estudos, foram convenientemente agrupadas em um único local na Internet por Fumera et al. (2006). A quarta base — Combo — é formada pela combinação das três outras bases.

A base Drezde (DREDZE; GEVARYAHU; ELIAS-BACHRACH, 2007) é composta por 2006 *e-mails ham* e 3297 *e-mails spam*. É utilizada nos estudos realizados por Mehta et al. (2008), Liu et al. (2010), Cheng et al. (2010), Wang et al. (2010) e Al-Duwairi et al. (2011).

A base Prag possui 8549 *e-mails spam* somente. Assim, adotando o mesmo procedimento empregado por Fumera et al. (2006), foram incluídos, nesta base, todos os *e-mails ham* da base *Drezde*.

A base Trec é formada por 1219 *e-mails ham* e 7365 *e-mails spam*, extraídos da base original Trec (CORMACK; LYNAM, 2005), versão 2007. A versão 2005 desta base foi utilizada por Byun et al. (2007) em seus estudos.

A tabela 1 descreve o tamanho das bases de dados utilizadas. Percebe-se, claramente, que há grande discrepância entre a quantidade de *e-mails ham* e *spam* nas bases de dados. Para evitar tendências no treinamento das Redes Neurais Artificiais, cada base foi incrementada com cópias de seus *e-mails ham*, para que a quantidade de *e-mails* de ambas as classes (*ham* e *spam*) fossem iguais. Drezde et al. (2007) observam que a base *SpamArchive* possui diversas duplicatas. Afirmam que este fato não constitui um problema, mas ao contrário, descreve melhor o mundo real, onde é comum usuários

receberem, por diversas vezes, as mesmas imagens ou imagens bem semelhantes.

Tabela 1: Tamanhos das Bases de Dados Utilizadas

| | ***ham*** | ***spam*** |
|---|---|---|
| *Drezde* | 2.006 | 3.297 |
| *Prag* | 0 | 8.549 |
| *Trec* | 1.219 | 7.365 |
| **Total** | 3.225 | 19.211 |

O desbalanceamento na quantidade de *e-mails spam* e *ham* presentes nas bases de dados, a pequena quantidade de bases disponível publicamente, e a falta de reúso das mesmas contribuem para dificultar a comparação de resultados obtidos por diferentes pesquisadores (BYUN et al., 2007). Neste trabalho, o uso de bases públicas já utilizadas por alguns pesquisadores teve por objetivo permitir que os resultados obtidos possam ser comparados com os resultados obtidos em outros estudos reportados na literatura.

## 4.2 Extração de Características

Como descrito na seção 2.3, as características RGB médio, histograma, histograma colorido, momento de cor e vetor de coerência de cor são extraídas das imagens de cada base de dados. gerando um arquivo CSV. Neste arquivo, cada linha contém as características de uma imagem e a indicação se esta é *spam* ou *ham*.

A extração de características é realizada por duas vezes. Na segunda extração, as imagens são previamente reduzidas para a resolução de 100x100 *pixels*[1], conforme descrito por Mehta et al. (2008). Estes autores afirmam que a redução das imagens ajuda a eliminar *pixels* aleatórios e métodos simples de ofuscação, tais como translação e escala. Assim, a extração de cada uma das características das imagens de uma base gera dois arquivos CSV — um para as imagens originais da base e outro para as imagens reduzidas.

## 4.3 Redes Neurais Artificiais

Cada característica extraída de uma imagem compõe um vetor que representa a imagem. Estes vetores de características são os vetores de entrada na Rede Neural Artificial. São realizados estudos preliminares de forma a se encontrar a melhor arquitetura de Rede Neural Artificial para cada modelo de característica. Após a determinação de sua arquite-

[1] A redução de resolução afeta a proporção (*aspect ratio*) da imagem.

tura, a rede é treinada e avaliada com cada base de dados. Os resultados são apresentados nas seções a seguir.

### 4.3.1 Treinamento

Conjuntos de vetores são criados para cada uma das bases de dados. Cada conjunto é composto por vetores de uma determinada característica extraída das imagens de uma mesma base de dados. Cada conjunto de vetores é dividido aleatoriamente em três partes — conjunto de treinamento, conjunto de validação e conjunto de testes, com as proporções de 40%, 20% e 40% respectivamente.

O treinamento da Rede Neural Artificial envolve o uso tanto do conjunto de treinamento quanto do de validação. Seu treinamento é dividido em passos de 64 épocas. A rede é treinada, em um passo, com o conjunto de treinamento. O treinamento é parado e é calculado o erro produzido pela rede sobre o conjunto de validação. Caso seja menor que o erro produzido no passo anterior, o processo de treinamento e validação é repetido até que a taxa de erro sobre o conjunto de validação pare de decrescer. Quando isso acontece, é encerrado o treinamento. Esta forma de treinamento é conhecida, na literatura, como validação cruzada (*cross-validation*).

As Redes Neurais Artificiais são treinadas por, pelo menos, dez vezes sobre cada conjunto de treinamento e validação. Em cada treinamento, são utilizados pesos iniciais diferentes, gerados aleatoriamente. O conjunto de testes nunca é usado durante o treinamento da rede.

## 4.4 Análise dos Resultados

A avaliação do desempenho das Redes Neurais Artificiais é feita sobre os conjuntos de teste. Para determinar se uma imagem, representada por seu vetor de características, em um conjunto de testes foi classificada corretamente ou não, são analisados os valores de ativação dos dois neurônios de saída da rede.

Cada Rede Neural Artificial utiliza dois neurônios na camada de saída. Um indica se a imagem é *ham* e o outro se a imagem é *spam*. Um resultado é considerado válido apenas quando o valor de ativação de um neurônio é maior que 0,6 e do outro é inferior a 0,4. Caso os valores estejam fora dessa faixa, o resultado é considerado como sendo um erro de classificação.

A tabela 2 mostra alguns exemplos de como, através dos valores de ativação dos neurônios, a classificação é determinada. A imagem *0001.jpg* foi classificada incorretamente como *spam*. A imagem *0002.jpg* foi classificada corretamente, pois o neurônio que representa *ham* tem ativação acima de 0,6 e o neurônio que representa *spam* tem ativação abaixo de 0,4. As imagens *0003.jpg* e *0004.jpg* foram classificadas incorretamente, pois os valores de ativação dos dois neurônios situaram-se fora dos valores de faixa permitidos.

Tabela 2: Exemplos de como são classificadas as imagens

| **Imagem** | **Neurônio *ham*** | **Neurônio *spam*** | **Classe** | **Classe Obtida** |
|---|---|---|---|---|
| *0001.jpg* | 0,02 | 0,98 | *ham* | *spam* (erro) |
| *0002.jpg* | 0,93 | 0,18 | *ham* | *ham* (acerto) |
| *0003.jpg* | 0,5 | 0,99 | *spam* | (erro) |
| *0004.jpg* | 0,01 | 0,55 | *spam* | (erro) |

O tempo médio de treinamento, o tempo médio de classificação e a porcentagem de classificações corretas são usados como métricas para a avaliação do desempenho da Rede Neural Artificial em cada estudo realizado. Nas tabelas que apresentam, nas seções seguintes, os resultados dos estudos, as porcentagens de classificações corretas em termos de *e-mails spam* e *ham* são descritas nas colunas "Detecção" e "Falso Positivo", respectivamente.

## 4.5 Análise de cada Característica

Os primeiros estudos tiveram por objetivo avaliar o comportamento individual de cada característica, de forma a verificar sua capacidade de representação de uma imagem.

### 4.5.1 Estudo 1 - RGB Médio

O estudo 1 utiliza, para a representação das imagens, a característica RGB médio, descrita na seção 2.3.3.1. Utiliza uma Rede Neural Artificial formada por duas camadas escondidas, uma de entrada e outra de saída. A camada de entrada possui três neurônios — um para cada componente de cor. Cada camada escondida possui 32 neurônios, dotados de função de ativação tangente hiperbólica (equação 2.10). Os dois neurônios da camada de saída possuem função de ativação linear (equação 2.8).

A tabela 3 apresenta os resultados do estudo. Com exceção da base *Trec*, as taxas de falsos positivos são significativamente altas, sugerindo que o RGB médio, usado de forma isolada, não produz informação suficiente para a classificação correta das imagens *ham*.

Tabela 3: Resultados do estudo com RGB Médio

| | **Detecção de Spam (%)** | **Falso Positivo (%)** | **Tempo do Treino (min)** |
|---|---|---|---|
| *Drezde* | 80,98 +/- 0,20 | 34,18 +/- 0,12 | 0,65 +/- 0,02 |
| *Prag* | 73,42 +/- 0,82 | 40,01 +/- 1,17 | 1,65 +/- 0,08 |
| *Trec* | 82,75 +/- 2,97 | 9,55 +/- 3,38 | 2,11 +/- 0,42 |
| *Combo* | 72,48 +/- 2,49 | 20,74 +/- 0,99 | 6,61 +/- 1,04 |

Drezde et al. (2007), Mehta et al. (2008) e Zhen et al. (2009) utilizam a característica RGB médio em conjunto com outras características. Assim, não é possível uma comparação direta dos resultados destes estudos com os obtidos no estudo 1.

#### 4.5.1.1 Estudo 1.1 - RGB Médio com Imagens Reduzidas

Neste estudo, as imagens foram reduzidas para resolução de 100x100 *pixels* antes da extração do RGB médio. A tabela 4 mostra os resultados. A Rede Neural Artificial apresenta uma melhoria discreta na taxa de falsos positivos sobre a base *Drezde*, enquanto, nas demais bases, há uma queda em sua capacidade de classificação de imagens. A base *Trec* foi a mais afetada pela redução da resolução, pois a taxa de falsos positivos obtida pela rede sobre suas imagens subiu de 9,55% para 16,96%.

Tabela 4: Resultados do estudo com RGB Médio e imagens reduzidas

| | **Detecção de Spam (%)** | **Falso Positivo (%)** | **Tempo do Treino (min)** |
|---|---|---|---|
| *Drezde* | 85,15 +/- 3,31 | 31,65 +/- 2,59 | 0,82 +/- 0,18 |
| *Prag* | 69,76 +/- 1,39 | 45,12 +/- 4,66 | 0,92 +/- 0,59 |
| *Trec* | 79,08% +/- 3,34 | 16,96% +/- 5,00 | 2,16 +/- 0,39 |
| *Combo* | 70,00% +/- 5,66 | 24,26% +/- 3,87 | 4,81 +/- 1,75 |

Mehta et al. (2008) afirmam que a redução da resolução das imagens ajuda a eliminar ruídos. Os resultados do estudo 1.1 sugerem, porém, que a informação dada pela característica RGB Médio é muito pouco afetada pela presença ou não de ruídos nas imagens.

### 4.5.2 Estudo 2 - Histograma

Este estudo usa o histograma em tons de cinza, descrito na seção 2.3.3.2, para representação das imagens. Usa, igualmente, uma rede composta por duas camadas escondidas, além das camadas de entrada e saída. A camada de entrada possui 256 neurônios — um para cada valor do histograma. Cada camada escondida possui 16 neurônios com função

de ativação tangente hiperbólica (equação 2.10). A camada de saída possui dois neurônios com função de ativação linear (equação 2.8).

A tabela 5 mostra os resultados obtidos. O histograma provê informação suficiente para que a rede classifique satisfatoriamente as imagens. Com a base *Trec*, por exemplo, a taxa de falsos positivos é de 1,75%, um valor consideravelmente baixo. O pior desempenho da rede, com taxa de falsos positivos de 13,49%, foi produzido na classificação das imagens da base *Prag*. Além de prover informação suficiente para a classificação correta das imagens, o histograma é uma característica facilmente extraível das imagens.

Tabela 5: Resultados do estudo com Histograma

| | **Detecção de Spam (%)** | **Falso Positivo (%)** | **Tempo do Treino (min)** |
|---|---|---|---|
| *Drezde* | 87,67 +/- 0,46 | 9,16 +/- 2,27 | 2,00 +/- 0,44 |
| *Prag* | 82,47 +/- 0,75 | 13,49 +/- 1,88 | 6,01 +/- 0,92 |
| *Trec* | 96,21 +/- 0,37 | 1,75 +/- 1,75 | 4,06 +/- 0,88 |
| *Combo* | 87,88 +/- 0,86 | 9,62 +/- 0,72 | 10,95 +/- 2,50 |

#### 4.5.2.1 Estudo 2.1 - Histograma com Imagens Reduzidas

Neste estudo, as imagens foram reduzidas para resolução de 100x100 *pixels* antes da extração do histograma. Os resultados do estudo estão apresentados na tabela 6. Com a redução da resolução, a Rede Neural Artificial alcança uma pequena melhora na taxa de falsos positivos sobre a base *Drezde*. Sobre as demais bases, entretanto, perde acurácia na classificação dos *e-mails spam*.

Tabela 6: Resultados do estudo com Histograma e imagens reduzidas

| | **Detecção de Spam (%)** | **Falso Positivo (%)** | **Tempo do Treino (min)** |
|---|---|---|---|
| *Drezde* | 92,23 +/- 0,47 | 6,05 +/- 0,53 | 1,70 +/- 0,28 |
| *Prag* | 79,45 +/- 0,80 | 13,31 +/- 2,33 | 5,28 +/- 1,23 |
| *Trec* | 97,34 +/- 0,07 | 2,47 +/- 0,18 | 4,18 +/- 0,79 |
| *Combo* | 87,13 +/- 0,72 | 10,21 +/- 0,36 | 11,15 +/- 3,21 |

Os resultados sugerem que a redução na resolução prejudica a representação das imagens através do histograma. Sugerem, igualmente, que a redução na resolução não só elimina ruídos das imagens, como descrito por Mehta et al. (2008), mas também elimina informações relevantes para suas representações através de histogramas.

### 4.5.3 Estudo 3 - Histograma Colorido

Este estudo emprega o histograma colorido para representação das imagens. Como mencionado na seção 2.3.3.3, o histograma colorido é calculado por meio do agrupamento, em 512 grupos, das cores da imagem.

Neste estudo, a rede é formada por duas camadas escondidas, além das camadas de entrada e saída. A camada de entrada contém 512 neurônios — um neurônio para cada valor do histograma. Cada camada escondida contém 16 neurônios. Na primeira camada, os neurônios possuem função de ativação tangente hiperbólica (equação 2.10) e, na segunda, função sigmóide (equação 2.9). A camada de saída contém dois neurônios, com função de ativação linear (equação 2.8). A tabela 7 apresenta os resultados do estudo.

Tabela 7: Resultados do estudo com Histograma Colorido

| | **Detecção de Spam (%)** | **Falso Positivo (%)** | **Tempo do Treino (min)** |
|---|---|---|---|
| *Drezde* | 91,31 +/- 0,68 | 6,75 +/- 1,95 | 5,96 +/- 1,10 |
| *Prag* | 87,97 +/- 0,28 | 9,18 +/- 0,61 | 18,50 +/- 2,68 |
| *Trec* | 98,08 +/- 0,05 | 1,69 +/- 0,05 | 11,57 +/- 2,28 |
| *Combo* | 91,28 +/- 0,22 | 5,08 +/- 0,30 | 28,20 +/- 2,50 |

Qu e Zhang (2009) também utilizam o histograma colorido para representação das imagens. No entanto, procuram detectar *spam* através da busca por imagens duplicadas. Além disto, utilizam uma base de imagens pessoal, o que dificulta a comparação de seus resultados com os do estudo 3.

Soranamageswari e Meena (2010) também usam o histograma colorido para representação das imagens e Redes Neurais Artificiais para classificá-las como *ham* ou *spam*. Utilizam, porém, 80% das imagens da base no treinamento da rede, o que favorece sobremodo a qualidade dos resultados que obtêm. Reportam uma taxa de classificação correta de *e-mails spam* de 92,82%. Não reportam a taxa de falsos positivos.

Os resultados obtidos neste estudo 3 sugerem que o histograma colorido representa satisfatoriamente as imagens das quatro bases. A Rede Neural Artificial alcança, sobre a base *Trec*, uma taxa de falsos positivos menor que sobre as demais bases de imagens.

#### 4.5.3.1 Estudo 3.1 - Histograma Colorido com Imagens Reduzidas

Neste estudo, as imagens foram previamente reduzidas para resolução de 100x100 *pixels* antes da extração do histograma colorido. A tabela 8 apresenta os resultados obtidos no estudo.

Tabela 8: Resultados do estudo com Histograma e imagens reduzidas

| | **Detecção de Spam (%)** | **Falso Positivo (%)** | **Tempo do Treino (min)** |
|---|---|---|---|
| *Drezde* | 93,76 +/- 0,57 | 7,04 +/- 0,27 | 5,73 +/- 1,45 |
| *Prag* | 88,08 +/- 0,31 | 9,35 +/- 0,49 | 15,91 +/- 3,06 |
| *Trec* | 98,85 +/- 0,15 | 1,18 +/- 0,08 | 13,06 +/- 1,88 |
| *Combo* | 91,99 +/- 0,54 | 6,04 +/- 0,39 | 27,51 +/- 5,92 |

Com a redução na resolução das imagens, há uma discreta melhoria na taxa de detecção de *spam*, mas há também uma discreta piora na taxa de falsos positivos. A *Trec* foi a única base com melhorias em ambos os indicadores. A taxa de detecção e de falsos positivos sobre esta base passou de 98,08% para 98,85% e de 1,69% para 1,18%, respectivamente, com a redução na resolução. Assim, os resultados obtidos no estudo 3.1 sugerem que a redução de resolução das imagens pouco afeta a informação contida em seus histogramas coloridos.

Mehta et al. (2008) utilizam o histograma colorido em seus estudos (juntamento com outras características) e imagens reduzidas a resolução de 100x100 *pixels*. Com seu uso, reportam taxas médias de classificações corretas de 95% a 99,6%, sendo esta última taxa obtida com o redimensionamento das imagens e uma proporção mais favorável entre a quantidade de *e-mails* da base de treinamento e a de testes. Faz-se necessário reportar, no entanto, o fato de que a base de dados usada por Mehta et al. possui os *e-mails* da base *Drezde*, *e-mails* de outras bases, além de *e-mails* pessoais, o que torna difícil uma comparação precisa de resultados.

### 4.5.4 Estudo 4 - Momento de Cor

Este estudo usa o momento de cor para representação das imagens. Como descrito na seção 2.3.3.4, nove valores descrevem o momento de cor da imagem. Estes nove valores consistem na média, desvio padrão e obliquidade de cada canal de cor no espaço RGB.

A Rede Neural Artificial empregada neste estudo é formada por duas camadas escondidas, uma camada de entrada e outra de saída. A camada de entrada possui 9 neurônios — um neurônio para cada valor da característica momento de cor.

As duas camadas escondidas possuem 16 neurônios cada. Os neurônios da primeira e segunda camadas possuem, respectivamente, funções de ativação tangente hiperbólica (equação 2.10) e sigmóide (equação 2.9). Os dois neurônios da camada de saída possuem função de ativação linear (equação 2.8). A tabela 9 mostra os resultados do estudo.

Tabela 9: Resultados do estudo com Momento de Cor

| | **Detecção de Spam (%)** | **Falso Positivo (%)** | **Tempo do Treino (min)** |
|---|---|---|---|
| *Drezde* | 86,86 +/- 0,63 | 14,16 +/- 0,82 | 0,34 +/- 0,04 |
| *Prag* | 77,91 +/- 0,62 | 16,79 +/- 0,81 | 1,45 +/- 0,13 |
| *Trec* | 96,11 +/- 0,33 | 5,48 +/- 0,69 | 0,90 +/- 0,10 |
| *Combo* | 83,52 +/- 1,20 | 14,91 +/- 0,58 | 2,34 +/- 0,32 |

Com a base *Trec*, a rede alcançou resultados superiores aos alcançados com as demais bases de imagens. Os resultados sugerem que o momento de cor é capaz de representar satisfatoriamente as imagens. Além disto, a informação fornecida pelo momento de cor é compacta, disposta em nove valores e, por consequência, o treinamento da Rede Neural Artificial é rápido.

Byun et al. (2007) também fizeram uso do momento de cor combinado com heterogeneidade de cor, conspicuidade e auto similaridade. Utilizaram imagens provenientes de várias bases e fontes públicas. Os autores reportam resultados com taxas de falsos positivos e de classificação de *spam* de até 19,1% e 86,6%, respectivamente.

Mehta et al. (2008) usam o momento de cor no espaço de cor HSV, combinado com outras características das imagens. Fazem uso de imagens provenientes da base *Drezde*, de *e-mails* do *SpamArchive* e de imagens pessoais. Reportam resultados com taxas de detecção de *spam* de até 98%, sobre as imagens da base *Drezde*.

Wang et al. (2010) utilizam, igualmente, o momento de cor. Em seus estudos, utilizam a base *Drezde*, combinada com outras bases de dados, e a característica momento de cor, combinada com outras características visuais e metadados, o que dificulta a comparação de seus resultados com os do estudo 4.

#### 4.5.4.1 Estudo 4.1 - Momento de Cor com Imagens Reduzidas

Neste estudo, as imagens foram reduzidas para resolução de 100x100 *pixels* antes da extração do momento de cor. A tabela 10 mostra os resultados obtidos.

Tabela 10: Resultados do estudo com Momento de Cor e imagens reduzidas

| | **Detecção de Spam (%)** | **Falso Positivo (%)** | **Tempo do Treino (min)** |
|---|---|---|---|
| *Drezde* | 85,31 +/- 0,48 | 16,31 +/- 1,50 | 0,43 +/- 0,05 |
| *Prag* | 75,94 +/- 0,94 | 22,92 +/- 1,95 | 1,29 +/- 0,19 |
| *Trec* | 95,98 +/- 0,60 | 6,57 +/- 1,06 | 0,88 +/- 0,13 |
| *Combo* | 83,90 +/- 0,93 | 15,87 +/- 0,62 | 2,45 +/- 0,32 |

Com a redução na resolução das imagens, houve, igualmente, uma redução na ca-

pacidade de classificação da Rede Neural Artificial. O maior aumento na taxa de falsos positivos, que saltou de 16,79% para 22,92%, ocorreu com a base *Prag*. Os resultados do estudo 4.1 sugerem, portanto, que a redução na resolução das imagens reduz a qualidade da informação contida no momento de cor.

### 4.5.5 Estudo 5 - Vetor de Coerência de Cor

O estudo 5 emprega o vetor de coerência de cor para a representação das imagens. Como descrito na seção 2.3.3.3, as cores das imagens são reduzidas para 64 cores. O vetor de características possui 128 valores — dois para cada cor — que indicam o nível de coerência e incoerência de cada cor.

A Rede Neural Artificial utilizada neste estudo é formada por duas camadas escondidas, além das camadas de entrada e saída. A camada de entrada possui 128 neurônios — um para cada valor da característica. As duas camadas escondidas possuem 64 neurônios cada. Os neurônios da primeira camada têm função de ativação tangente hiperbólica (equação 2.10) e os da segunda, função sigmóide (equação 2.9). Os dois neurônios da camada de saída têm função de ativação linear (equação 2.8). A tabela 11 apresenta os resultados obtidos.

Tabela 11: Resultados do estudo com Vetor de Coerência de Cores

| | **Detecção de Spam (%)** | **Falso Positivo (%)** | **Tempo do Treino (min)** |
|---|---|---|---|
| *Drezde* | 93,44 +/- 0,36 | 7,19 +/- 0,35 | 11,86 +/- 5,60 |
| *Prag* | 84,46 +/- 0,83 | 12,5 +/- 0,95 | 28,02 +/- 3,12 |
| *Trec* | 98,71 +/- 0,16 | 1,94 +/- 0,06 | 20,39 +/- 3,65 |
| *Combo* | 91,02 +/- 0,57 | 8,95 +/- 1,20 | 57,70 +/- 13,37 |

A rede classificou corretamente aproximadamente 98% das imagens *spam*. Somada à qualidade da informação fornecida por esta característica, faz-se necessário também ressaltar a simplicidade do processo para sua extração. A menor taxa de falsos positivos foi alcançada pela rede sobre a base *Trec*. Os resultados sugerem que o vetor de coerência de cor possui grande potencial para a representação das imagens.

#### 4.5.5.1 Estudo 5.1 - Vetor de Coerência de Cor com Imagens Reduzidas

Neste estudo, as imagens foram reduzidas para resolução de 100x100 *pixels* antes da extração do vetor de coerência de cor. A tabela 12 apresenta os resultados.

Os resultados indicam uma melhoria na taxa de falsos positivos sobre as bases *Drezde*,

Tabela 12: Resultados do estudo com Vetor de Coerência de Cores sobre imagens reduzidas

| | **Detecção de Spam (%)** | **Falso Positivo (%)** | **Tempo do Treino (min)** |
|---|---|---|---|
| *Drezde* | 94,49 +/- 0,20 | 5,41 +/- 0,30 | 9,68 +/- 1,41 |
| *Prag* | 86,30 +/- 0,74 | 9,78 +/- 0,73 | 26,84 +/- 2,02 |
| *Trec* | 89,96 +/- 0,12 | 11,52 +/- 0,07 | 12,61 +/- 3,98 |
| *Combo* | 91,57 +/- 0,74 | 5,70 +/- 0,21 | 59,56 +/- 9,55 |

*Prag* e *Combo*. Sobre a base *Trec*, porém, a rede teve desempenho inferior — a taxa de detecção de *spam* caiu de 98,71% para 89,96% e a taxa de falsos positivos aumentou de 1,94% para 11,52%. Os resultados indicam, assim, que nem sempre a redução de resolução das imagens aumenta a qualidade da informação fornecida por esta característica.

Mehta et al. (2008) também fazem uso da característica vetor de coerência de cor juntamente com outras características visuais. Além disto, reduzem as imagens para resolução de 100x100 *pixels*, de forma a eliminar ruídos. Com o uso do vetor de coerência de cor em conjunto com as outras características visuais, os autores alcançam uma taxa de 98%, para detecção de *spam*, sobre a base *Drezde*. Não reportam, porém, as taxas de falsos positivos.

### 4.5.6 Análise dos Resultados de cada Característica

Todas as características empregadas demonstram fornecer informação suficiente para a detecção de imagens *spam*. RGB médio é a característica que provê menor qualidade de informação e vetor de coerência de cor e histograma colorido, as que proveem maior qualidade de informação para a detecção de *spam*. Estes resultados podem ser observados na figura 27.

Em relação à taxa de falsos positivos, diferença na qualidade de informação fornecida pelas quatro características é mais significativa. RGB médio e histograma colorido são as características que proveem, respectivamente, a menor e maior qualidade de informação. Estes resultados podem ser observados na figura 28.

Os gráficos apresentados nas figuras 27 e 28 indicam que o histograma colorido é a característica com melhor capacidade para a representação das imagens. Supre, assim, o classificador neural com informação de maior qualidade para a classificação das imagens nas classes *ham* e *spam*.

As taxas de detecção de *spam* obtidas pela Rede Neural Artificial sobre imagens reduzidas à resolução de 100x100 *pixels* são semelhantes às taxas obtidas sobre imagens

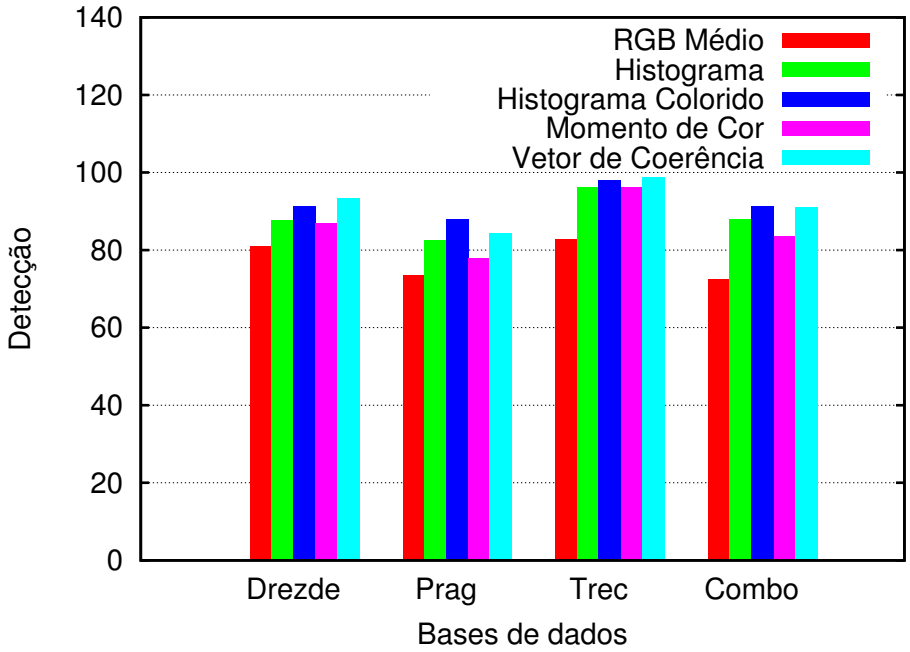


Figura 27: Comparativo da taxa de detecção da Rede Neural Artificial com o uso das características.

normais, como pode ser observado na figura 29. Observa-se, igualmente, nesta figura, que, com a redução prévia na resolução, a característica vetor de coerência de cor perde informação relevante para a detecção de *spam* sobre as imagens da base *Trec*.

Igualmente, as taxas de falsos positivos, ou seja, de detecção de *ham*, obtidas pela Rede Neural Artificial sobre imagens reduzidas à resolução de 100x100 *pixels* são semelhantes às taxas obtidas sobre imagens normais, como pode ser observado na figura 30. Nesta figura, observa-se que, com a redução prévia na resolução, a característica vetor de coerência de cor perde informação relevante para a detecção de *ham* sobre as imagens da base *Trec*. Sobre as imagens das bases *Prag* e *Combo*, porém, este comportamento inverte-se.

Os gráficos apresentados nas figuras 29 e 30 sugerem que o histograma colorido é a característica com melhor capacidade para a representação de imagens com resolução reduzida. Sugerem, igualmente, que a redução na resolução das imagens tanto pode ser benéfica quanto prejudicial, pois, ao reduzir ruídos, a redução na resolução parece também reduzir informações relevantes das imagens.

A figura 31 faz uma comparação dos resultados obtidos pela característica RGB médio sobre as imagens originais e reduzidas. No gráfico da figura, pode ser observado que a redução na resolução das imagens retira da característica informação relevante para a representação das imagens, reduzindo a capacidade de classificação da Rede Neural

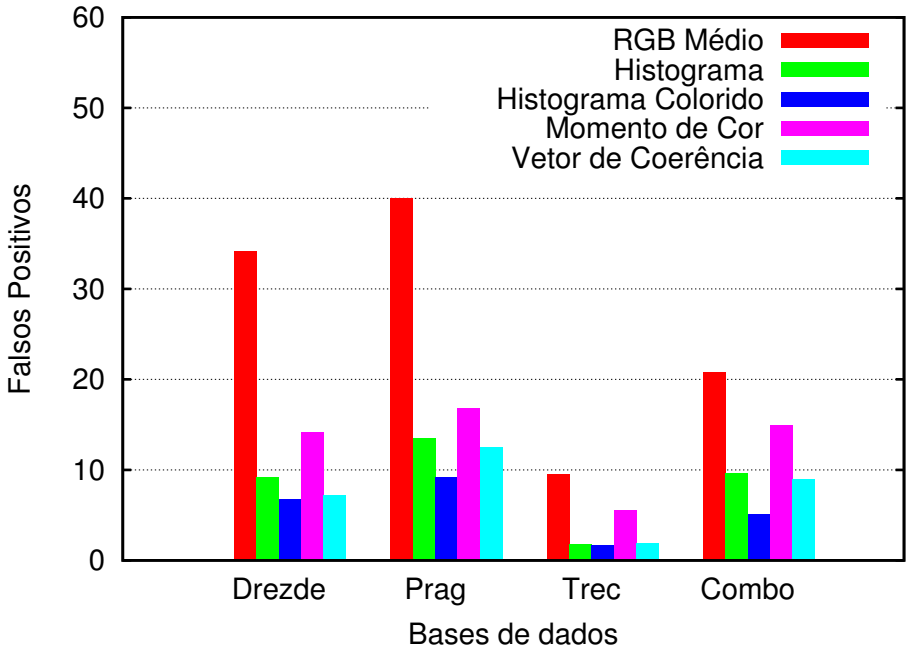


Figura 28: Comparativo da taxa de falsos positivos entre as técnicas.

Artificial.

Os resultados apresentados na figura 32, por sua vez, sugerem que a redução na resolução das imagens pouco afeta a capacidade de representação do histograma. Há uma pequena diferença entre os resultados com as imagens originais e os com as imagens reduzidas das bases *Drezde* e *Prag*. Com as imagens das bases *Trec* e *Combo*, porém, os resultados são próximos.

Com o uso do histograma colorido, o desempenho da Rede Neural Artificial sobre as imagens originais e reduzidas é similar ao desempenho com o uso do histograma em tons de cinza, como pode ser visto na figura 33. Apenas com base *Drezde*, a rede apresenta um discreto aumento na taxa de detecção de *spam* sobre imagens reduzidas. Com as demais bases, a rede apresenta pequenas diferenças de desempenho sobre imagens originais e reduzidas.

A figura 34 mostra o desempenho da Rede Neural Artificial sobre as imagens originais e reduzidas, representadas pela característica momento de cor. É possível observar que a redução na resolução das imagens reduziu levemente a capacidade de representação, por meio desta característica, das imagens da base *Prag*.

A figura 35 apresenta o desempenho da Rede Neural Artificial sobre as imagens originais e reduzidas, representadas pelo vetor de coerência de cor. Com exceção da base *Trec*, onde a rede apresentou queda de desempenho na classificação das imagens reduzidas, a

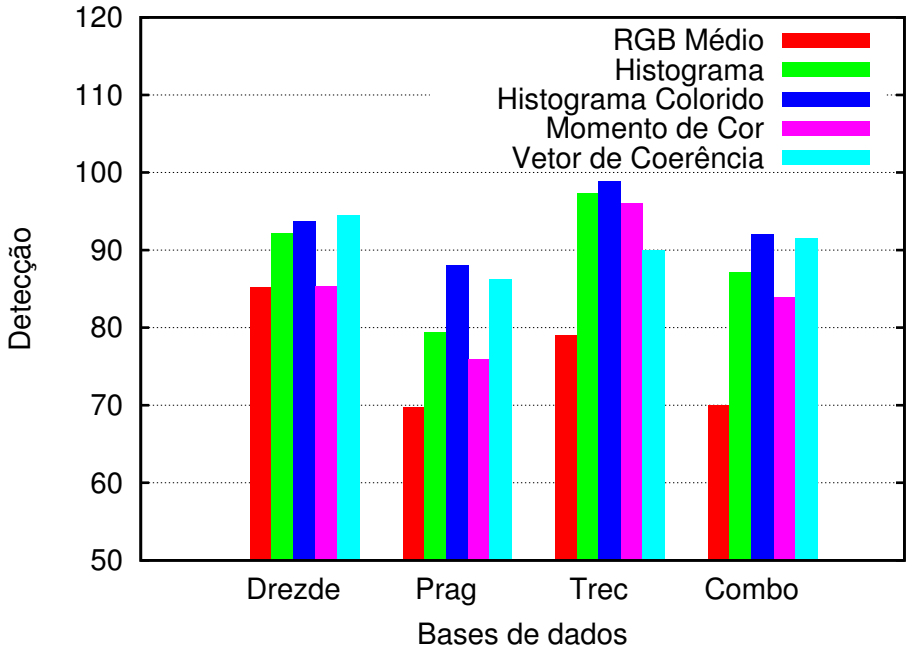


Figura 29: Comparativo entre as taxas de detecção das técnicas com o uso de imagens reduzidas.

rede apresentou desempenho similar sobre as imagens originais e reduzidas das demais bases.

A comparação do uso das cinco características sobre as imagens originais e reduzidas sugere que a redução de resolução das imagens deve ser usada com critério, pois pode ser prejudicial em algumas situações. Contudo, com características cuja extração possui alto custo computacional, como, por exemplo, vetor de coerência de cor, a redução da resolução pode ser vantajosa.

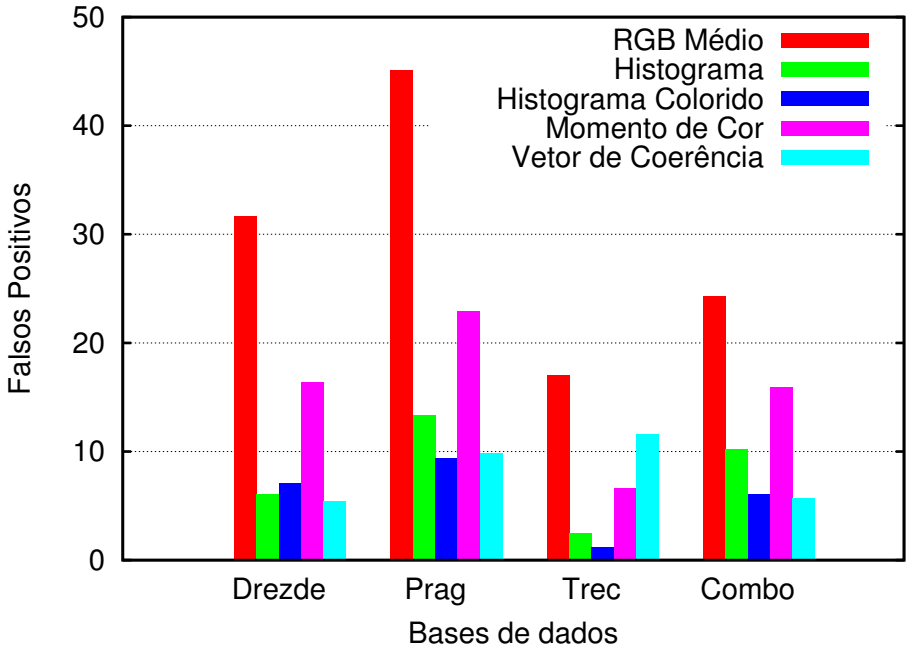


Figura 30: Comparativo entre as taxas de falsos positivos com o uso de imagens reduzidas.

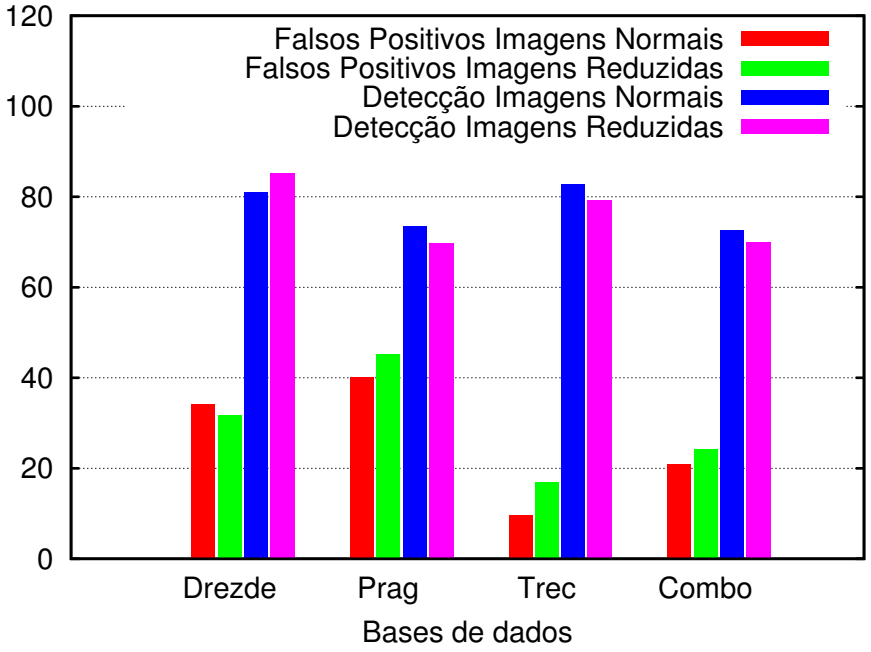


Figura 31: Comparação do uso do RGB médio sobre as imagens originais e reduzidas.

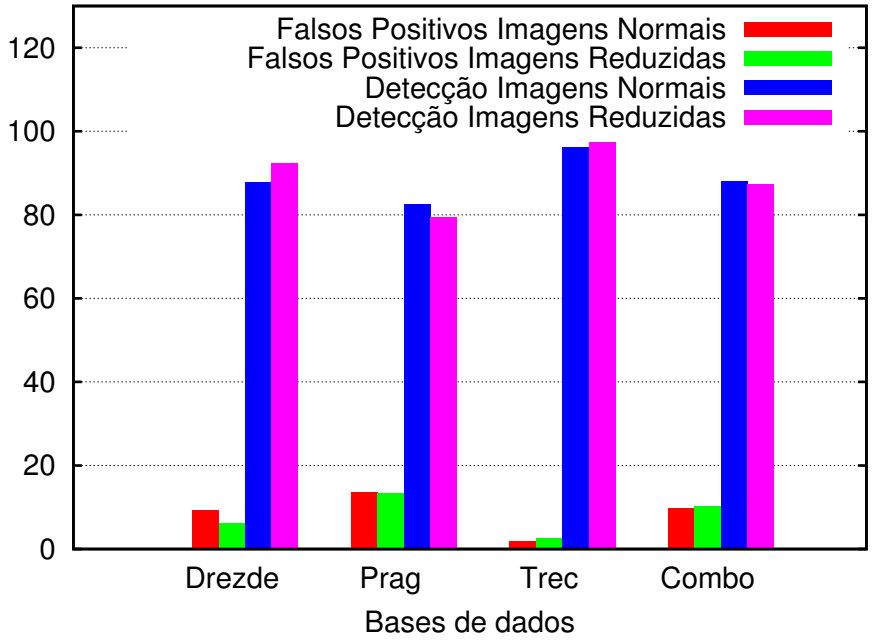


Figura 32: Comparação do uso do histograma sobre as imagens originais e reduzidas.

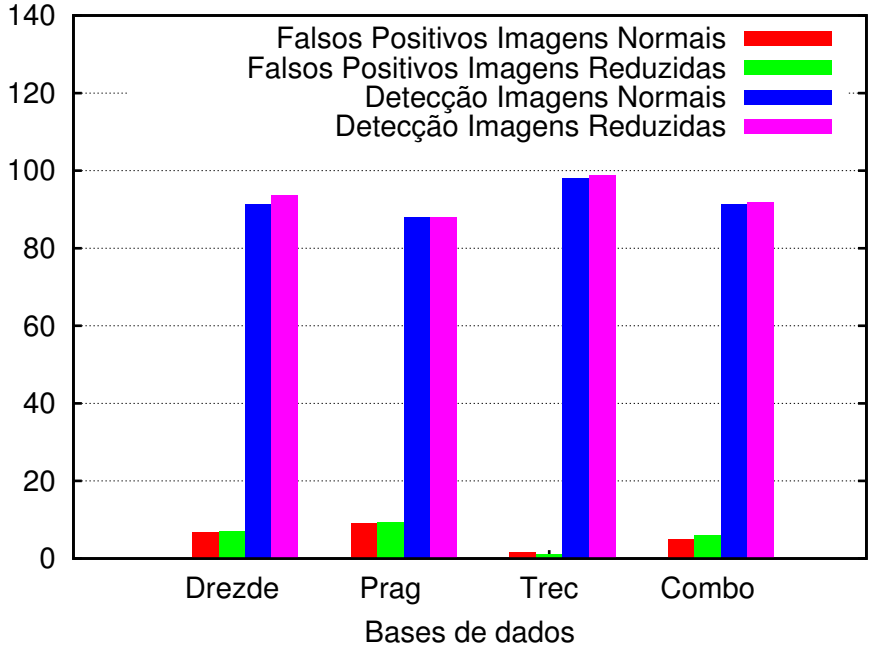


Figura 33: Comparação do uso do histograma colorido sobre as imagens originais e reduzidas.

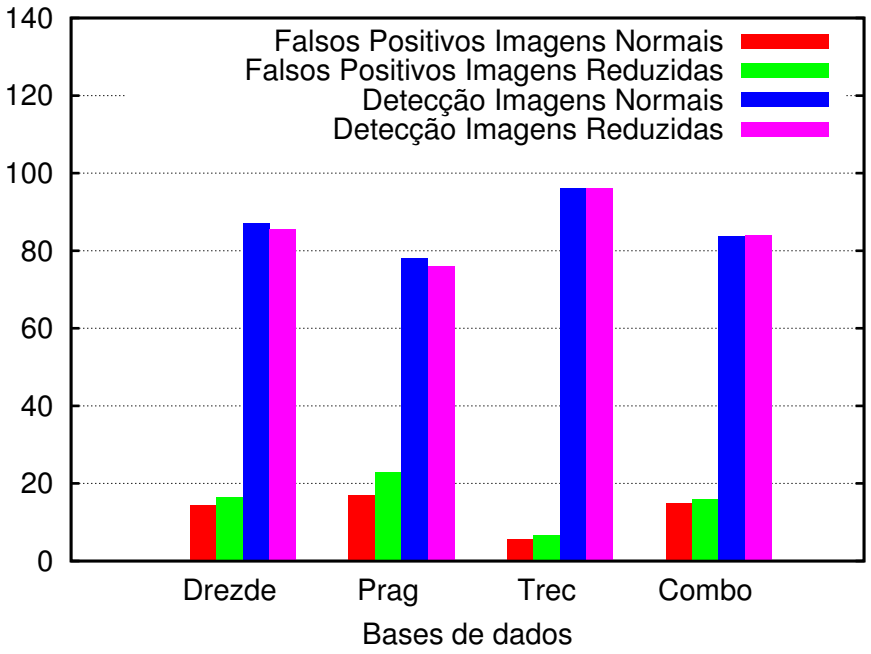


Figura 34: Comparação do uso do momento de cor sobre as imagens originais e reduzidas.

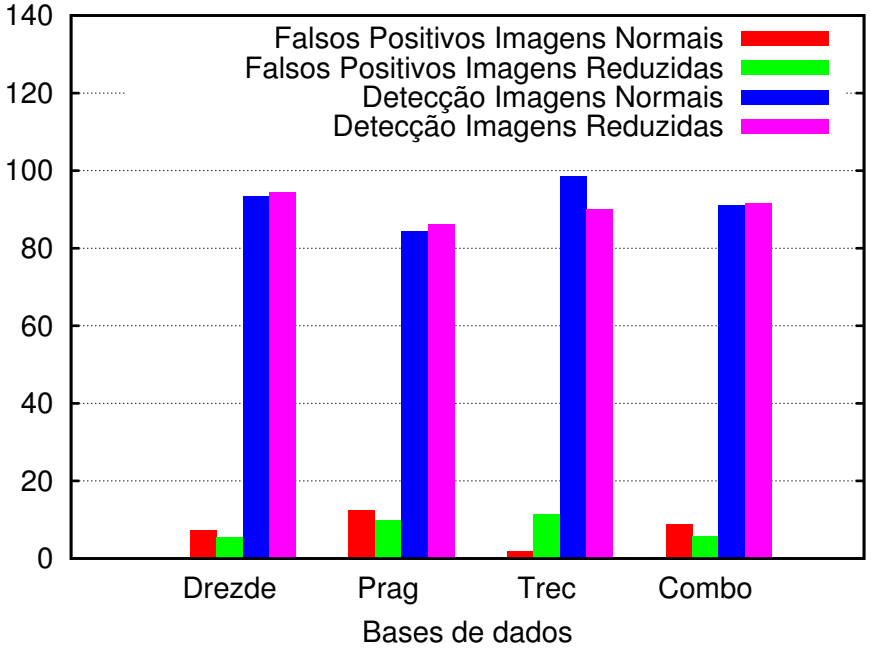


Figura 35: Comparação do uso do vetor de coerência sobre as imagens originais e reduzidas.

## 4.6 Análise da Combinação das Características

Os estudos reportados nesta seção tiveram por objetivo avaliar o comportamento combinado das características, de forma a verificar a capacidade das combinações na representação das imagens. A característica RGB médio não foi utilizada em nenhuma combinação, por ser um caso particular da característica momento de cor e ter apresentado, como descrito na seção 4.5.6, desempenho inferior ao das demais características.

### 4.6.1 Estudo 6 - Vetor de Coerência de Cor e Histograma

No estudo 6, as características vetor de coerência de cor e histograma foram combinadas para representar as imagens das bases. Com o vetor de coerência, as imagens são representadas por um vetor de características com 128 valores — dois para cada cor —, que indicam o nível de coerência e incoerência, respectivamente, de cada cor. Com o histograma, as imagens são representadas em tons de cinza.

A Rede Neural Artificial empregada no estudo é formada por duas camadas escondidas, além das camadas de entrada e saída. A camada de entrada possui 348 neurônios — um para cada valor de cada uma das características. As duas camadas escondidas possuem 32 neurônios cada. Os neurônios da primeira e segunda camadas possuem funções de ativação tangente (equação 2.10) e sigmóide (equação 2.9), respectivamente. A camada de saída é composta por dois neurônios com função de ativação linear (equação 2.8). A tabela 13 mostra os resultados do estudo.

Tabela 13: Resultados do estudo com vetor de coerência e histograma

| | **Detecção de Spam (%)** | **Falso Positivo (%)** | **Tempo do Treino (min)** |
|---|---|---|---|
| *Drezde* | 94,63 +/- 0,53 | 3,61 +/- 0,25 | 11,06 +/- 2,90 |
| *Prag* | 89,25 +/- 0,29 | 5,71 +/- 0,30 | 24,06 +/- 2,33 |
| *Trec* | 97,97 +/- 0,22 | 1,21 +/- 0,08 | 21,15 +/- 5,10 |
| *Combo* | 93,26 +/- 0,24 | 2,83 +/- 0,15 | 60,07 +/- 7,99 |

Ao compararem-se os resultados do estudo 6 com os do estudo 5 (seção 4.5.5), que faz uso apenas do vetor de coerência de cor, pode-se notar uma considerável redução na taxa de falsos positivos sobre todas as bases. A taxa de detecção de *spam* sobre as imagens da base *Trec* reduziu-se em aproximadamente 1%. Nas demais bases, houve uma discreta melhora.

Na comparação do estudo 6 com o estudo 2 (seção 4.5.2), que faz uso apenas do histograma, observam-se resultados semelhantes. As taxas de falsos positivos reduziram-

se significativamente enquanto que as taxas de detecção de *spam* apresentaram discreta melhora.

Os resultados sugerem, portanto, que a combinação destas duas características aumenta a qualidade da representação das imagens. Com a combinação das duas características, observa-se que a taxa de detecção de *spam* sobre a base *Drezde* aproxima-se de 98%, valor obtido por Mehta et al. (2008) em seus estudos. Infelizmente, os autores não reportam as taxas de falsos positivos obtidas.

#### 4.6.1.1 Estudo 6.1 - Vetor de Coerência de Cor e Histograma com Imagens Reduzidas

Neste estudo, as imagens foram reduzidas para resolução de 100x100 *pixels* antes da extração das duas características — vetor de coerência de cor e histograma. A tabela 14 apresenta os resultados.

Tabela 14: Resultados do estudo com Vetor de Coerência de Cores e Histograma sobre imagens reduzidas

| | **Detecção de Spam (%)** | **Falso Positivo (%)** | **Tempo do Treino (min)** |
|---|---|---|---|
| *Drezde* | 95,37 +/- 0,61 | 3,35 +/- 0,28 | 10,81 +/- 2,24 |
| *Prag* | 88,82 +/- 0,42 | 5,81 +/- 0,61 | 28,01 +/- 5,38 |
| *Trec* | 97,21 +/- 0,02 | 1,48 +/- 0,02 | 22,56 +/- 4,47 |
| *Combo* | 93,31 +/- 0,18 | 3,45 +/- 0,35 | 51,46 +/- 8,41 |

De acordo com a tabela, a taxa de falsos positivos sobre a base *Drezde* reduziu-se, mas, sobre as demais bases, houve um discreto aumento. A taxa de detecção de *spam* reduziu-se sobre todas as bases, com exceção sobre a base *Drezde*, onde a taxa apresentou uma ligeira melhora.

### 4.6.2 Estudo 7 - Vetor de Coerência de Cor e Histograma Colorido

Neste estudo, as características vetor de coerência de cor e histograma colorido foram combinadas para representar as imagens das bases. Com o vetor de coerência, tal como realizado no estudo 6, as imagens são representadas por um vetor de características com 128 valores — dois para cada cor —, que indicam o nível de coerência e incoerência, respectivamente, de cada cor. Com o histograma colorido, as imagens são representadas por 512 valores — um para cada cor.

A Rede Neural Artificial utilizada é formada por duas camadas escondidas, além das camadas de entrada e saída. A camada de entrada possui 640 neurônios — um

para cada valor de cada uma das características. As duas camadas escondidas possuem 32 neurônios cada. Os neurônios da primeira e segunda camadas possuem funções de ativação tangente hiperbólica (equação 2.10) e sigmóide (equação 2.9), respectivamente. Os dois neurônios da camada de saída possuem função de ativação linear (equação 2.8). A tabela 15 apresenta os resultados do estudo.

Tabela 15: Resultados do estudo com Vetor de Coerência e Histograma Colorido

| | **Detecção de Spam (%)** | **Falso Positivo (%)** | **Tempo do Treino (min)** |
|---|---|---|---|
| *Drezde* | 95,71 +/- 0,30 | 4,67 +/- 0,27 | 21,90 +/- 5,10 |
| *Prag* | 89,66 +/- 0,40 | 5,25 +/- 0,28 | 38,09 +/- 3,36 |
| *Trec* | 97,83 +/- 0,71 | 5,18 +/- 3,50 | 34,19 +/- 9,41 |
| *Combo* | 93,86 +/- 0,30 | 2,87 +/- 0,26 | 80,75 +/- 13,86 |

Ao compararem-se os resultados do estudo 7 com os do estudo 5 (seção 4.5.5), que faz uso apenas do vetor de coerência de cor, percebe-se que houve uma redução significativa na taxa de falsos positivos sobre as bases *Drezde*, *Prag* e *Combo*. Sobre a base *Trec*, a menor taxa de falsos positivos obtida foi de 1,68%, valor superior ao obtido no estudo 5, com o uso apenas do vetor de coerência de cor.

Em relação à taxa de detecção de *spam*, os resultados do estudo 7 são ligeiramente superiores aos do estudo 5, sobre as bases *Drezde*, *Prag* e *Combo*. Sobre a base *Trec*, porém, há uma discreta redução na taxa.

Na comparação do estudo 7 com o estudo 3 (seção 4.5.3), que faz uso apenas do histograma colorido, observam-se resultados semelhantes. Com exceção da base *Trec*, há uma significativa redução na taxa de falsos positivos e uma discreta melhora na taxa de detecção de *spam*.

Os resultados do estudo 7 são semelhantes aos do estudo 6 (seção 4.6.1). Tal como no estudo 6, houve uma redução significativa na taxa de falsos positivos com a combinação das duas características. Os resultados sugerem, assim, que a combinação das duas características aumenta a qualidade da representação das imagens.

#### 4.6.2.1 Estudo 7.1 - Vetor de Coerência de Cor e Histograma Colorido com Imagens Reduzidas

Neste estudo, as imagens foram reduzidas para resolução de 100x100 *pixels* antes da extração das duas características — vetor de coerência de cor e histograma colorido. A tabela 16 apresenta os resultados do estudo.

Os resultados indicam um aumento na taxa de falsos positivos sobre as imagens das

Tabela 16: Resultados do estudo com Vetor de Coerência de Cores e Histograma Colorido com imagens reduzidas

| | **Detecção de Spam (%)** | **Falso Positivo (%)** | **Tempo do Treino (min)** |
|---|---|---|---|
| *Drezde* | 95,85 +/- 0,34 | 3,96 +/- 0,25 | 19,39 +/- 4,21 |
| *Prag* | 89,94 +/- 0,37 | 7,09 +/- 0,39 | 65,64 +/- 27,65 |
| *Trec* | 98,62 +/- 0,11 | 1,29 +/- 0,06 | 31,64 +/- 3,16 |
| *Combo* | 93,55 +/- 0,17 | 3,67 +/- 0,16 | 75,37 +/- 8,09 |

bases *Prag* e *Combo*. Sobre as imagens das demais bases, porém, há uma redução nesta taxa. A taxa de detecção de *spam* exibe uma discreta melhora em quase todas as bases, excetuando-se a base *Combo*.

### 4.6.3 Estudo 8 - Vetor de Coerência de Cor e Momento de Cor

Neste estudo, as características vetor de coerência de cor e momento de cor foram combinadas para representar as imagens das bases. O vetor de coerência das imagens, tal como realizado no estudo 6, possui 128 valores — dois para cada cor —, que indicam o nível de coerência e incoerência, respectivamente, de cada cor. O momento de cor das imagens é calculado no espaço RGB. Possui nove valores — média, desvio padrão e obliquidade para cada canal de cor.

A Rede Neural Artificial é formada por duas camadas escondidas, uma de entrada e outra de saída. A camada de entrada possui 137 neurônios — um para cada valor de cada uma das características. Cada uma das duas camadas escondidas possui 16 neurônios. Os neurônios da primeira e segunda camadas possuem funções de ativação tangente hiperbólica (equação 2.10) e sigmóide (equação 2.9), respectivamente. Os neurônios da camada de saída possuem função de ativação linear (equação 2.8). A tabela 17 apresenta os resultados do estudo.

Tabela 17: Resultados do estudo com Vetor de Coerência e Momento de Cor

| | **Detecção de Spam (%)** | **Falso Positivo (%)** | **Tempo do Treino (min)** |
|---|---|---|---|
| *Drezde* | 94,97 +/- 0,30 | 5,17 +/- 0,22 | 1,04 +/- 0,11 |
| *Prag* | 88,78 +/- 0,50 | 9,11 +/- 0,48 | 3,04 +/- 0,32 |
| *Trec* | 98,35 +/- 0,24 | 1,77 +/- 0,09 | 2,55 +/- 0,40 |
| *Combo* | 92,09 +/- 0,25 | 6,28 +/- 0,40 | 5,49 +/- 0,80 |

Ao compararem-se os resultados do estudo 8 com os do estudo 5 (seção 4.5.5), que faz uso apenas do vetor de coerência de cor, percebe-se que houve uma melhora nas taxas de falsos positivos e de detecção de *spam* sobre as bases *Drezde*, *Prag* e *Combo*. Sobre a base *Trec*, porém, houve uma pequena redução na taxa de detecção de *spam*.

Por ter uma quantidade bem menor de neurônios nas duas camadas escondidas, a Rede Neural Artificial empregada no estudo 8 é consideravelmente menor que a rede empregada no estudo 5. Assim, apesar dos resultados do estudo 8 serem pouco superiores aos do estudo 5, a adição da característica momento de cor na representação das imagens é recomendada, por reduzir os tempos de treinamento e de classificação de imagens consumidos pela rede.

#### 4.6.3.1 Estudo 8.1 - Vetor de Coerência de Cor e Momento de Cor com Imagens Reduzidas

Neste estudo, as imagens foram reduzidas para resolução de 100x100 *pixels* antes da extração das duas características — vetor de coerência de cor e momento de cor. A tabela 18 apresenta os resultados do estudo.

Tabela 18: Resultados do estudo com Vetor de Coerência e Momento de Cor com Imagens Reduzidas

| | **Detecção de Spam (%)** | **Falso Positivo (%)** | **Tempo do Treino (min)** |
|---|---|---|---|
| *Drezde* | 94,51 +/- 0,22 | 4,62 +/- 0,11 | 1,31 +/- 0,09 |
| *Prag* | 87,14 +/- 3,75 | 10,49 +/- 1,28 | 3,81 +/- 0,79 |
| *Trec* | 98,00 +/- 0,27 | 2,13 +/- 0,10 | 2,98 +/- 0,30 |
| *Combo* | 91,78 +/- 0,43 | 6,24 +/- 0,73 | 6,97 +/- 0,65 |

Há pequenas diferenças entre os resultados do estudo 8 e 8.1. Apesar do pequeno aumento na taxa de falsos positivos obtidas no estudo 8.1, a redução da resolução das imagens reduzidas pode ser vantajosa, devido ao alto custo computacional para a geração do vetor de coerência, custo este dependente das resoluções das imagens.

### 4.6.4 Estudo 9 - Histograma e Momento de Cor

Neste estudo, as características histograma e momento de cor foram combinadas para representar as imagens das bases. O histograma possui 256 valores — um para cada tom de cinza. O momento de cor é calculado no espaço RGB. Possui nove valores — média, desvio padrão e obliquidade para cada canal de cor.

A Rede Neural Artificial utilizada é formada por duas camadas escondidas, além das camadas de entrada e saída. A camada de entrada tem 265 neurônios — um para cada valor de cada uma das características. Cada camada escondida tem 32 neurônios. Os neurônios da primeira e segunda camadas possuem funções de ativação tangente hiperbólica (equação 2.10) e sigmóide (equação 2.9), respectivamente. Os dois neurônios da

camada de saída possuem função de ativação linear (equação 2.8). A tabela 19 apresenta os resultados do estudo.

Tabela 19: Resultados do estudo com Histograma e Momento de Cor

| | **Detecção de Spam (%)** | **Falso Positivo (%)** | **Tempo do Treino (min)** |
|---|---|---|---|
| *Drezde* | 94,77 +/- 0,56 | 4,55 +/- 0,60 | 8,41 +/- 2,61 |
| *Prag* | 86,37 +/- 0,46 | 8,49 +/- 0,55 | 19,06 +/- 4,42 |
| *Trec* | 97,07 +/- 0,23 | 1,32 +/- 0,09 | 14,37 +/- 3,20 |
| *Combo* | 91,99 +/- 0,38 | 6,08 +/- 0,37 | 35,71 +/- 4,26 |

Ao compararem-se os resultados do estudo 9 com os do estudo 2 (seção 4.5.2), que faz uso apenas do histograma, percebe-se que houve uma pequena melhora na taxa de detecção de *spam* e uma expressiva melhora na taxa de falsos positivos. Apesar do momento de cor não apresentar resultados tão bons quanto aos das outras características quando usadas isoladamente, os resultados sugerem que seu uso é vantajoso quando combinado com o histograma.

#### 4.6.4.1 Estudo 9.1 - Histograma e Momento de Cor com Imagens Reduzidas

Neste estudo, as imagens foram reduzidas para resolução de 100x100 *pixels* antes da extração das duas características — histograma e momento de cor. A tabela 20 apresenta os resultados do estudo.

Tabela 20: Resultados do estudo com Histograma e Momento de Cor com Imagens Reduzidas

| | **Detecção de Spam (%)** | **Falso Positivo (%)** | **Tempo do Treino (min)** |
|---|---|---|---|
| *Drezde* | 94,26 +/- 0,69 | 5,12 +/- 0,62 | 7,59 +/- 1,66 |
| *Prag* | 84,23 +/- 0,92 | 10,64 +/- 0,96 | 19,35 +/- 4,14 |
| *Trec* | 97,98 +/- 0,25 | 1,96 +/- 0,18 | 14,21 +/- 2,76 |
| *Combo* | 90,70 +/- 0,39 | 6,60 +/- 0,66 | 38,91 +/- 4,52 |

Com a redução das imagens, observa-se a redução em algumas das taxas de detecção de *spam* e o aumento nas taxas de falsos positivos. É interessante ressaltar que os resultados dos estudos 2 (seção 4.5.2) e 4 (seção 4.5.4), que fazem uso isolado do histograma e do momento de cor, respectivamente, apresentam, igualmente, um aumento nas taxas de falsos positivos. Os resultados sugerem, assim, que a redução na resolução das imagens reduz a qualidade da informação fornecida por estas duas características combinadas.

### 4.6.5 Estudo 10 - Histograma Colorido e Momento de Cor

Neste estudo, as características histograma colorido e momento de cor foram combinadas para representar as imagens das bases. O histograma colorido possui 512 valores — um para cada cor. O momento de cor é calculado no espaço RGB. Possui nove valores — média, desvio padrão e obliquidade para cada canal de cor.

A Rede Neural Artificial utilizada é formada por duas camadas escondidas, além das camadas de entrada e saída. A camada de entrada possui 512 neurônios — um para cada valor de cada uma das características. Cada camada escondida possui 32 neurônios. Os neurônios da primeira e segunda camadas possuem funções de ativação tangente hiperbólica (equação 2.10) e sigmóide (equação 2.9), respectivamente. Os dois neurônios da camada de saída possuem função de ativação linear (equação 2.8). A tabela 21 apresenta os resultados do estudo.

Tabela 21: Resultados do estudo com Histograma Colorido e Momento de Cor

| | **Detecção de Spam (%)** | **Falso Positivo (%)** | **Tempo do Treino (min)** |
|---|---|---|---|
| *Drezde* | 93,38 +/- 0,45 | 4,51 +/- 0,23 | 15,84 +/- 2,64 |
| *Prag* | 89,02 +/- 1,01 | 9,14 +/- 1,43 | 36,18 +/- 6,65 |
| *Trec* | 98,30 +/- 0,06 | 1,56 +/- 0,06 | 33,96 +/- 7,84 |
| *Combo* | 91,88 +/- 0,37 | 5,66 +/- 0,18 | 66,96 +/- 10,00 |

Ao compararem-se os resultados do estudo 10 com os do estudo 3 (seção 4.5.3), que faz uso apenas do histograma colorido, percebe-se que houve uma melhora na taxa de detecção de *spam*. Em relação à taxa de falsos positivos, há uma discreta melhora sobre as imagens de quase todas as bases, com exceção apenas da base *Combo*. A melhor taxa de falsos positivos foi obtida sobre as imagens da base *Drezde*. Os resultados sugerem, portanto, que é vantajoso o uso do histograma colorido em conjunto com o momento de cor para a representação das imagens.

Liu et al. (2010) propõem um sistema para classificação de imagens *spam* composto por três camadas. Caso uma camada não demonstre um determinado grau de certeza na classificação, a imagem é encaminhada para a camada seguinte. Na terceira e última camada, as imagens são representadas por meio do histograma colorido e do momento de cor e são classificadas. Os autores não detalham como as cores são agrupadas no histograma colorido, mas mencionam que, na extração do momento de cor, são considerados apenas os dois primeiros momentos. Utilizam as bases *Trec* e *Drezde*, além de duas outras bases de dados privadas. As taxas de falsos positivos e de detecção de *spam* obtidas por eles variam, respectivamente, de 17,06% a 18,01% e de 94,93% a 96,8%.

Segundo a tabela 21, é possível perceber que as taxas de detecção de *spam* obtidas no estudo 10 são superiores às obtidas por Liu et al. É possível perceber, igualmente, que as taxas de falsos positivos obtidas no estudo 10 são significativamente superiores.

#### 4.6.5.1 Estudo 10.1 - Histograma Colorido e Momento de Cor com Imagens Reduzidas

Neste estudo, as imagens foram reduzidas para resolução de 100x100 *pixels* antes da extração das duas características — histograma colorido e momento de cor. A tabela 22 apresenta os resultados do estudo.

Tabela 22: Resultados do estudo com Histograma Colorido e Momento de Cor com Imagens Reduzidas

| | **Detecção de Spam (%)** | **Falso Positivo (%)** | **Tempo do Treino (min)** |
|---|---|---|---|
| *Drezde* | 94,28 +/- 0,31 | 5,72 +/- 0,28 | 12,72 min +/- 2,03 |
| *Prag* | 88,67 +/- 0,62 | 10,25 +/- 0,71 | 34,30 +/- 4,93 |
| *Trec* | 98,65 +/- 0,17 | 1,45 +/- 0,09 | 32,46 +/- 8,17 |
| *Combo* | 91,90 +/- 0,63 | 6,89 +/- 0,38 | 62,83 +/- 7,83 |

Com a redução na resolução das imagens, observa-se que ocorre um aumento na taxa de falsos positivos sobre as imagens de quase todas as bases, com exceção apenas da base *Trec*. Contudo, as taxas de detecção de *spam* apresentaram uma discreta melhora sobre as imagens das bases *Drezde*, *Trec* e *Combo*. Sobre as imagens da base *Prag*, a taxa de detecção de *spam* apresentou uma ligeira redução.

### 4.6.6 Análise dos Resultados da Combinação das Características

Com as combinações das características descritas nos estudos 6 a 10, a Rede Neural Artificial produziu melhores resultados, tanto em termos da taxa de detecção de *spam* quanto em termos da taxa de falsos positivos. Isto sugere que a combinação de duas características possui qualidade representacional maior que a de cada uma das duas características isoladamente. A figura 36 apresenta as taxas de detecção de *spam* obtidas com as combinações das características das imagens originais.

Na figura 36, é possível perceber-se que as taxas de detecção de *spam* estão bem próximas umas das outras. A melhor taxa de detecção é obtida com a combinação das características vetor de coerência de cor e histograma colorido. A pior taxa é obtida com a combinação das características histograma e momento de cor, sobre as imagens das bases *Prag*, *Trec* e *Combo*.

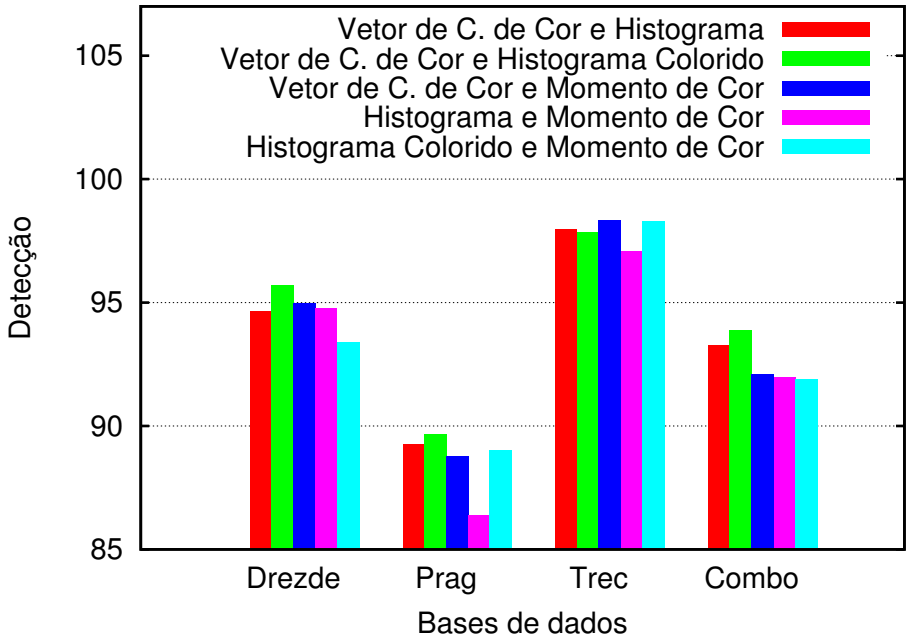


Figura 36: Comparação das taxas de detecção de *spam* com as características combinadas.

A figura 37 apresenta as taxas de falsos positivos obtidas com as combinações das características das imagens originais. A melhor taxa de falsos positivos é alcançada com a combinação das características vetor de coerência de cor e histograma.

Os resultados apresentados nas figuras 36 e 37 indicam que as características vetor de coerência de cor e histograma formam a melhor combinação para a representação das imagens originais. As características vetor de coerência de cor e histograma colorido formam a segunda melhor combinação.

A figura 38 apresenta as taxas de detecção de *spam* obtidas com as combinações das características das imagens reduzidas para a resolução de 100x100 *pixels*. As taxas são similares às obtidas com as imagens originais. Uma vez mais, a melhor taxa de detecção é obtida com a combinação das características vetor de coerência de cor e histograma colorido e a pior taxa, com a combinação das características histograma e momento de cor.

A figura 39 apresenta as taxas de falsos positivos obtidas com as combinações das características das imagens reduzidas para a resolução de 100x100 *pixels*. As taxas são similares às obtidas com as imagens originais. Uma vez mais, a melhor taxa de falsos positivos é alcançada com a combinação das características vetor de coerência de cor e histograma. Ao contrário do resultado obtido com imagens originais, porém, a taxa de falsos positivos obtida com a combinação das características vetor de coerência de cor e

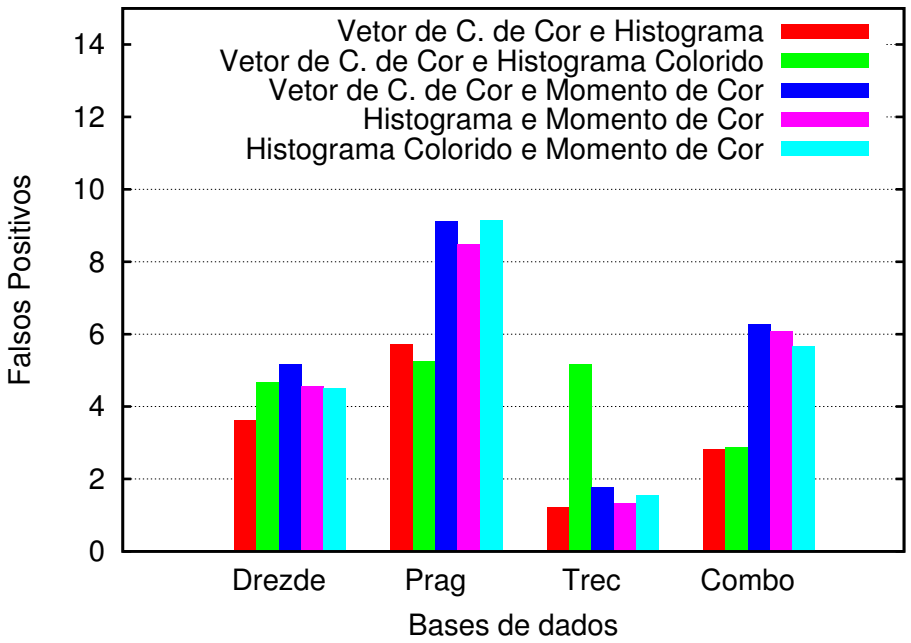


Figura 37: Comparação das taxas de falsos positivos com as características combinadas.

momento de cor é similar às taxas obtidas com as demais combinações das características.

Os resultados apresentados nas figuras 38 e 39 indicam que as características vetor de coerência de cor e histograma formam a melhor combinação para a representação das imagens reduzidas para a resolução de 100x100 *pixels*. As características vetor de coerência de cor e histograma colorido formam, novamente, a segunda melhor combinação.

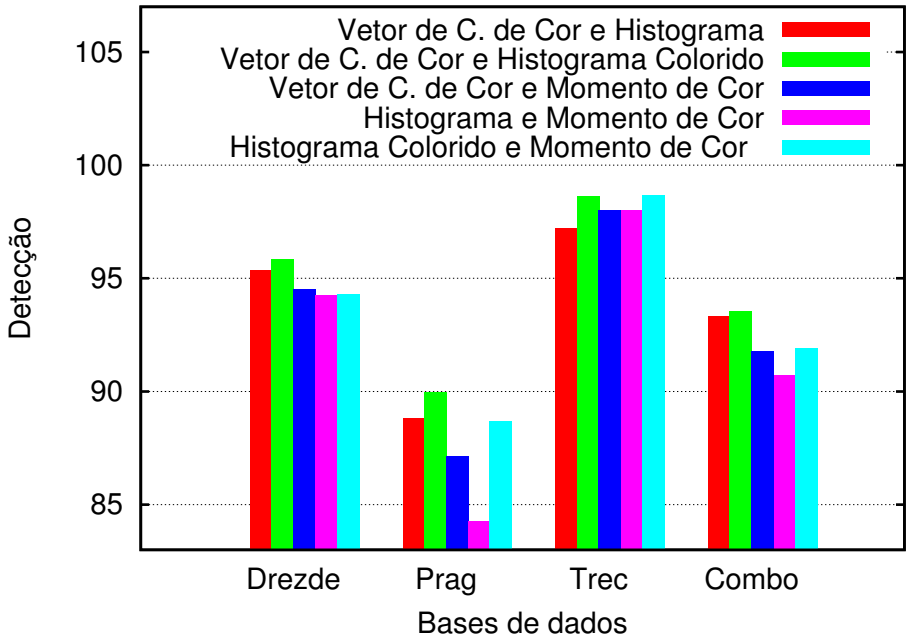


Figura 38: Comparação das taxas de detecção de *spam* obtidas com as combinações das características das imagens reduzidas.

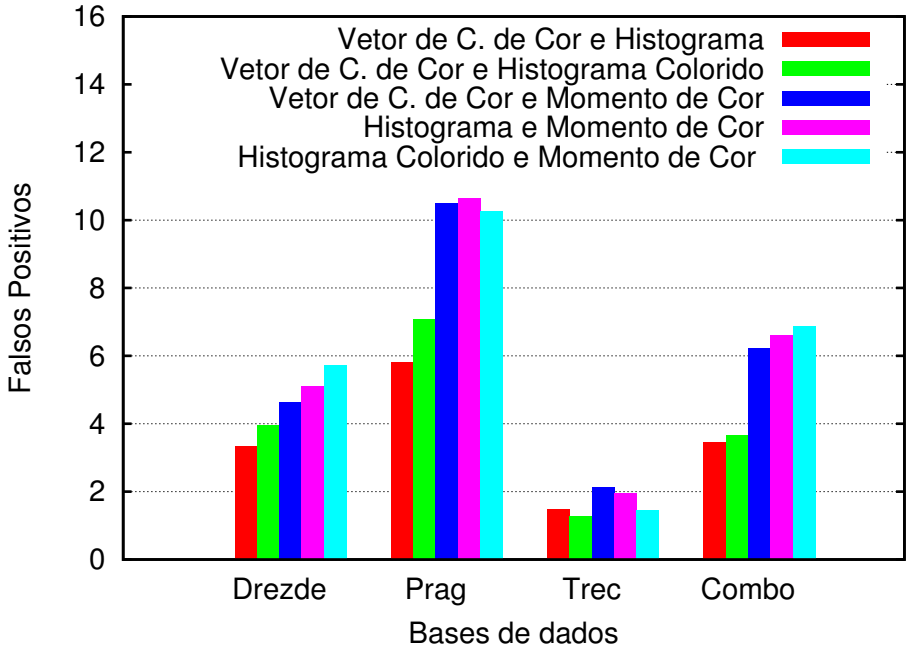


Figura 39: Comparação das taxas de falsos positivos obtidas com as combinações das características das imagens reduzidas.

## 4.7 Comparação entre os resultados obtidos sobre imagens originais e reduzidas

Como visto na seção anterior (seção 4.6.6), as características vetor de coerência de cor e histograma formam a melhor combinação para a representação tanto das imagens originais quanto das imagens reduzidas para a resolução de 100x100 *pixels*. A figura 40 compara as taxas de detecção de *spam* e de falsos positivos obtidos com o uso da combinação destas duas características para a representação das imagens originais e reduzidas.

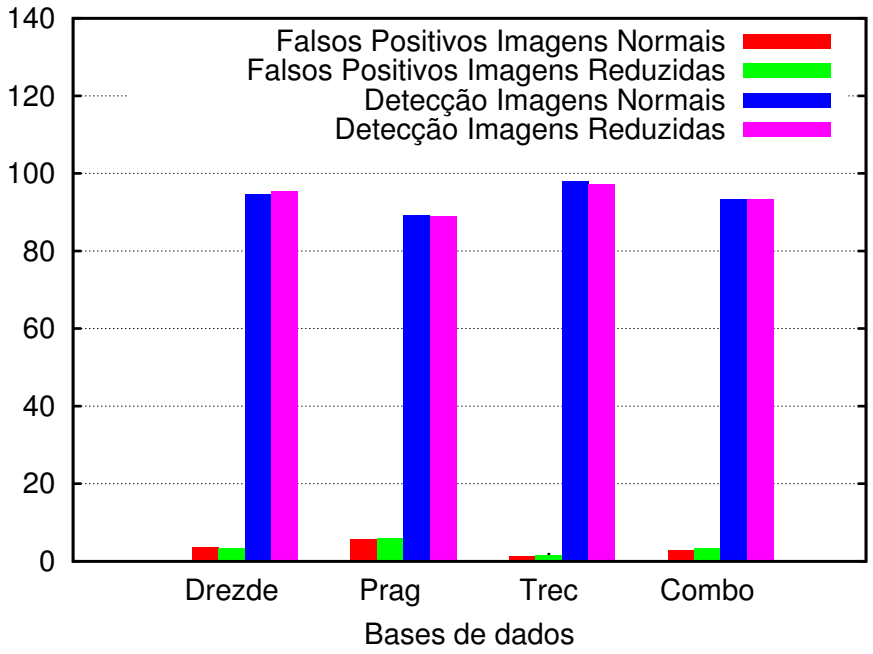


Figura 40: Comparação dos resultados com imagens originais e reduzidas: características vetor de coerência de cor e histograma.

A figura 40 indica que a redução na resolução das imagens não afeta a qualidade da representação produzida pela combinação das características vetor de coerência de cor e histograma. Assim, para esta combinação de características, a redução na resolução das imagens torna-se vantajosa, pois o custo computacional para extração do vetor de coerência de cor de imagens reduzidas é menor (seção 4.10).

As figuras 41, 42, 43 e 44 comparam as taxas de detecção de *spam* e de falsos positivos obtidos com o uso das demais combinações das características para a representação das imagens originais e reduzidas.

Das figuras 41, 42, 43 e 44, é possível perceber-se que a redução na resolução das imagens não afeta a qualidade da representação produzida pela combinação das demais características — vetor de coerência de cor com histograma colorido, vetor de coerência

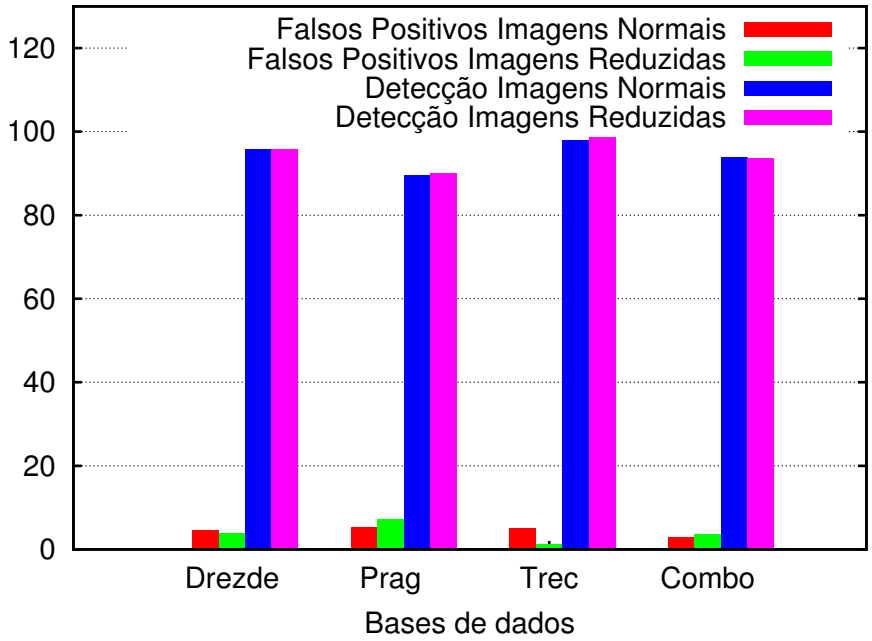


Figura 41: Comparação dos resultados com imagens originais e reduzidas: características vetor de coerência de cor e histograma colorido.

de cor com momento de cor, histograma com momento de cor e histograma colorido com momento de cor. Para as combinações que usam a característica vetor de coerência de cor, porém, a redução na resolução das imagens é vantajosa, pois reduz o custo computacional para extração desta característica.

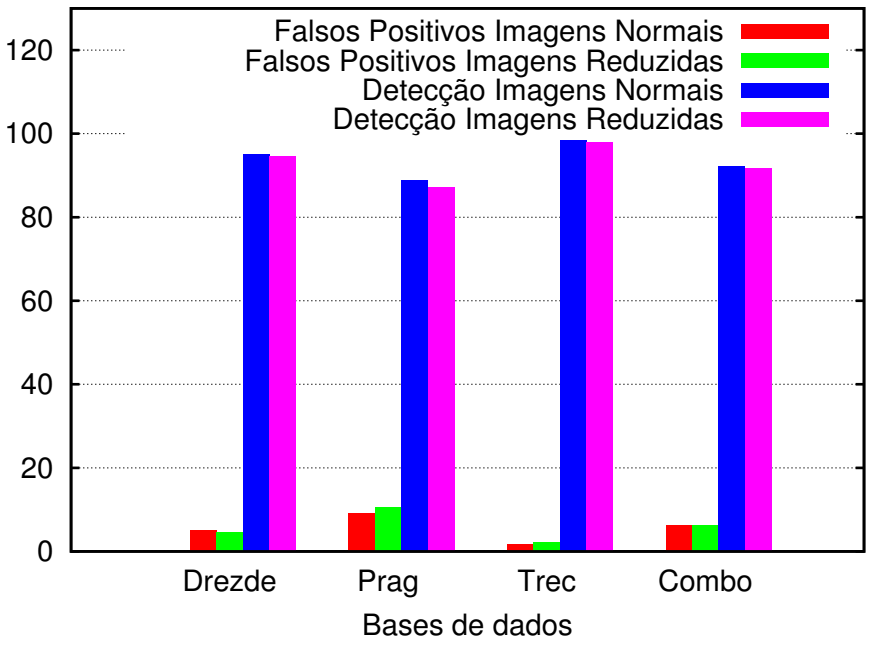


Figura 42: Comparação dos resultados com imagens originais e reduzidas: características vetor de coerência de cor e momento de cor.

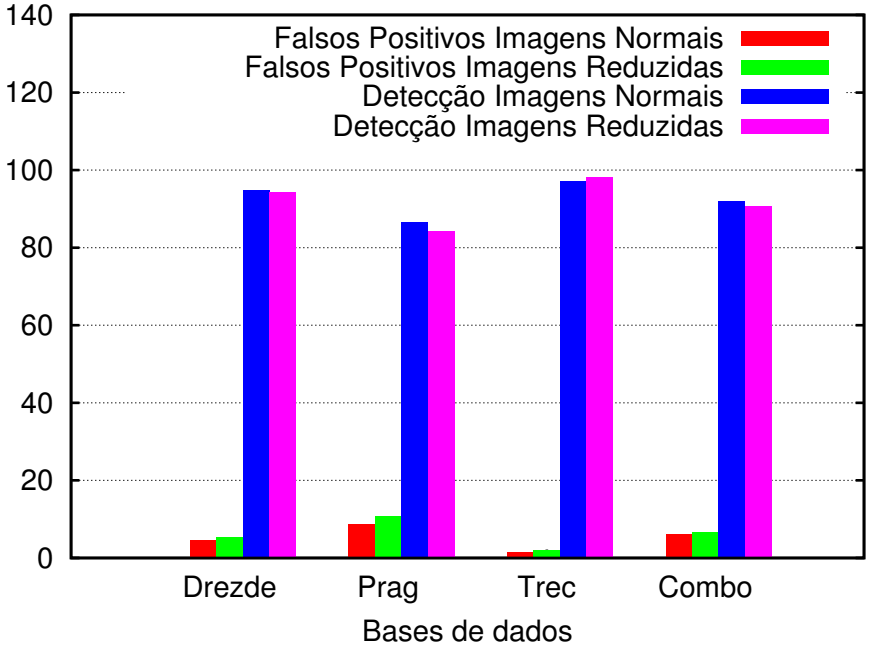


Figura 43: Comparação dos resultados com imagens originais e reduzidas: características histograma e momento de cor.

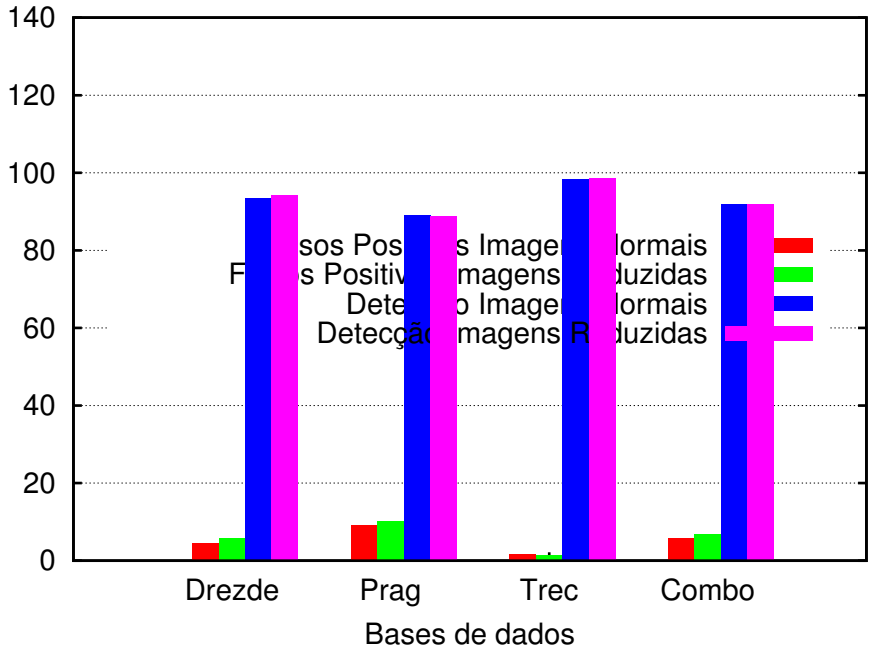


Figura 44: Comparação dos resultados com imagens originais e reduzidas: características histograma colorido e momento de cor.

## 4.8 Comparação entre os resultados obtidos com a melhor característica individual e a melhor combinação de características

Como visto na seção 4.5.6, o histograma colorido é a característica individual que mais bem representa as imagens originais e reduzidas. Por sua vez, como descrito na seção 4.6.6, o vetor de coerência de cor e histograma é a combinação de características que mais bem representa as imagens originais e reduzidas. Assim, as figuras 45 e 46 comparam, respectivamente, as taxas de detecção de *spam* e de falsos positivos obtidas com o uso da melhor característica individual e da melhor combinação de características.

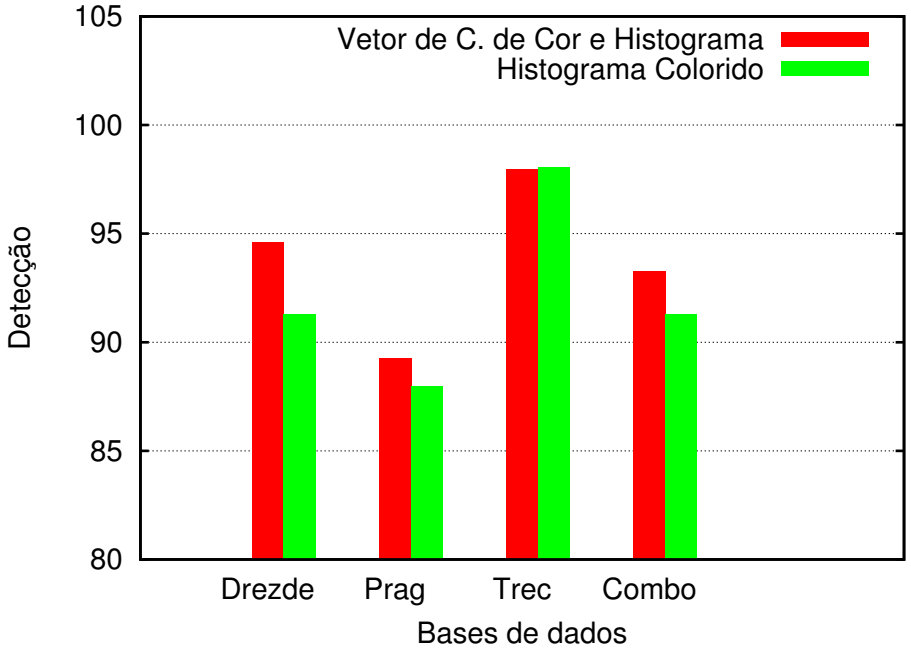


Figura 45: Comparação da taxa de detecção de *spam* obtida com o histograma colorido contra a obtida com a combinação vetor de coerência de cor e histograma.

As figuras 45 e 46 indicam que o uso da combinação das características vetor de coerência de cor e histograma produz melhores resultados que o uso individual da característica histograma colorido. Assim, dentre todas as características individuais e combinadas estudadas, a combinação vetor de coerência de cor e histograma é a que fornece a melhor qualidade de informação representacional para a classificação das imagens nas duas classes — *ham* e *spam*.

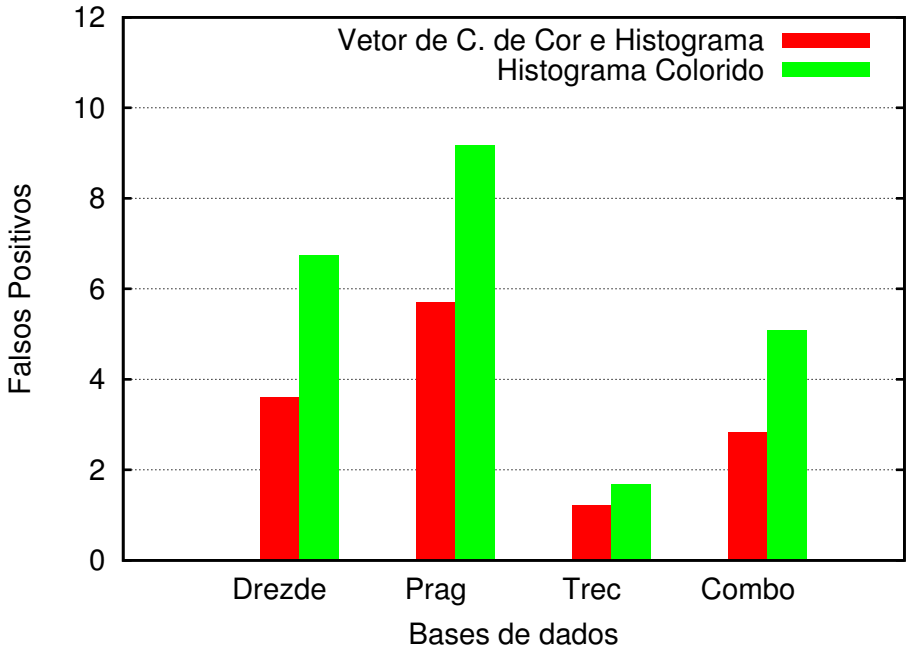


Figura 46: Comparação da taxa de falsos positivos obtida com o histograma colorido contra a obtida com a combinação vetor de coerência de cor e histograma.

## 4.9 Análise da redução na resolução das imagens

Em todos os estudos realizados, foram também analisados os resultados obtidos com imagens reduzidas à resolução de 100x100 *pixels*. Mehta et al. (2008) afirma que a redução das imagens para esta resolução é benéfica pois, além de eliminar ruídos das imagens, diminui o custo computacional para a extração de algumas de suas características.

De fato, conforme descrito na próxima seção (seção 4.10), a redução na resolução das imagens reduz significativamente o custo computacional para extração de suas características. Com o uso de algumas características (ou combinações de características), porém, conforme apresentado nos estudos de 1 a 10, a Rede Neural Artificial alcançou, com imagens reduzidas, resultados inferiores aos alcançados com imagens originais. Os resultados alcançados pela rede nestes estudos sugerem que, além de eliminar ruídos, como mencionado por Mehta et al., a redução na resolução reduz, igualmente, a qualidade da informação representacional provida por algumas características.

Para ilustrar a redução na qualidade da informação representacional provida pela característica histograma, pode-se tomar, como exemplo, a imagem *spam* apresentada na figura 47. Esta imagem contém ruídos. Apesar disto, foi classificada corretamente pela Rede Neural Artificial.

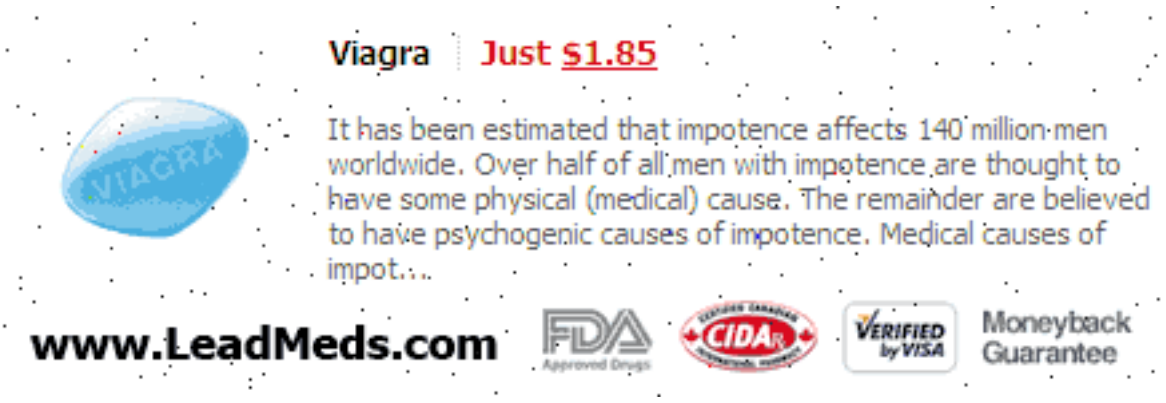


Figura 47: Imagem *spam* analisada.

A figura 48 apresenta a imagem da figura 47 reduzida à resolução de 100x100 *pixels*. É possível observar que a redução na resolução reduziu os ruídos da imagem original, mas não os eliminou completamente. Apesar da redução dos ruídos, foi classificada incorretamente pela Rede Neural Artificial.

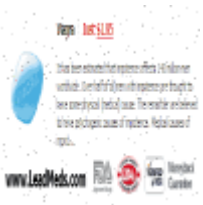

Figura 48: Imagem *spam* analisada, reduzida para a resolução de 100x100 pixels.

As figuras 49 e 50 apresentam os histogramas da imagem original (figura 47) e da imagem reduzida (figura 48), respectivamente.

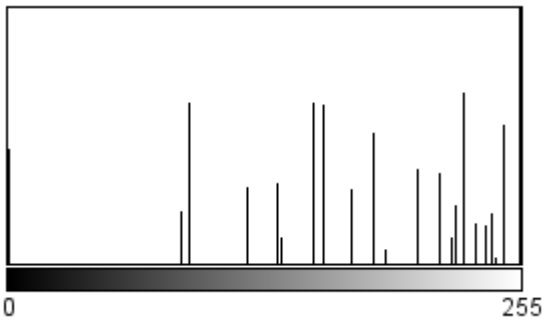


Figura 49: Histograma em tons de cinza da imagem *spam* original.

É possível observar que o histograma da imagem reduzida é significantemente diferente do da imagem original, apresentando um comportamento semelhante ao de uma equalização. Vários picos foram eliminados e as cores da imagem estão mais bem distribuídas. As diferenças nos histogramas sugerem que a redução na resolução das imagens pode redu-

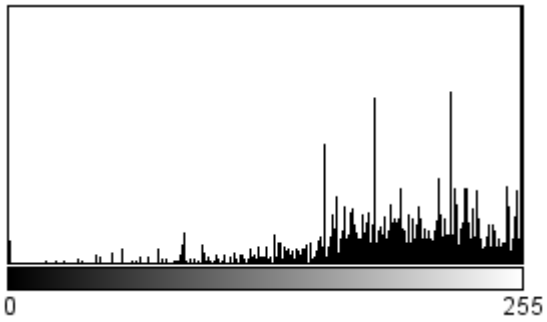


Figura 50: Histograma em tons de cinza da imagem *spam* reduzida para a resolução de 100x100 *pixels*.

zir, igualmente, a qualidade da informação representacional fornecida pela característica, afetando a capacidade de classificação da Rede Neural Artificial.

## 4.10 Análise de desempenho do sistema anti-spam

O custo computacional do sistema anti-spam pode ser medido através dos tempos computacionais gastos no carregamento das imagens originais e reduzidas, na extração de suas características e em suas classificações.

O tempo computacional gasto pela biblioteca *ImageJ* para carregar em memória principal cada imagem do disco rígido é alto. No carregamento das imagens, a biblioteca também extrai delas diversas informações, tais como, as características histograma, cor média, dentre outras. Nos estudos realizados, verificou-se que a biblioteca *ImageJ* consegue carregar aproximadamente 36 imagens por segundo. Um programa em linguagem C++, utilizando a biblioteca *OpenCV*[2], é capaz de carregar aproximadamente 118 imagens por segundo. O tempo computacional gasto pela *ImageJ* no carregamento das imagens é, portanto, alto.

É importante ressaltar que, nos estudos realizados, as imagens são lidas diretamente do disco rígido do computador. No entanto, em um sistema anti-spam, não necessariamente as imagens serão lidas do disco.

Como mencionado anteriormente, a biblioteca *ImageJ* já extrai, durante o carregamento de cada imagem, algumas de suas características, como, por exemplo, seu histograma. No entanto, tanto estas características quanto as demais foram extraídas individualmente de cada imagem, para mensurar-se o tempo computacional gasto na extração.

[2]Open Source Computer Vision (http://opencv.org/)

O tempo computacional de classificação é mensurado usando-se a melhor arquitetura de Rede Neural Artificial em cada estudo. As melhores arquiteturas estão descritas junto com os resultados dos estudos, nas seções 4.5.1 a 4.6.5.

As tabelas 23 e 24 mostram, respectivamente, os tempos obtidos, nos estudos, com o uso individual e combinado das características. Os tempos são medidos pelo número de imagens processadas por segundo. Como os tempos computacionais para carregamento e classificação das imagens originais são praticamente idênticos aos das imagens reduzidas, as tabelas incluem apenas uma coluna com informação sobre os tempos para carregamento e classificação das imagens.

Tabela 23: Tempos computacionais obtidos com o uso individual das características — todos os tempos são medidos pelo número de imagens processadas por segundo.

| | **Carga** | **Extração** | **Extração Red.** | **Classificação** |
|---|---|---|---|---|
| *Histograma* | 36,87 | 186,91 | 955,26 | 1.997,06 |
| *Histograma Colorido* | 36,75 | 220,54 | 980,59 | 1.285,93 |
| *RGB Médio* | 36,93 | 248,22 | 1.566,71 | 2.5641,27 |
| *Momento de Cor* | 37,24 | 121,26 | 1.091,42 | 2.1862,24 |
| *V. de C. de Cor* | 36,84 | 22,77 | 286,34 | 3.610,07 |

Tabela 24: Tempos computacionais obtidos com o uso combinado das características — todos os tempos são medidos pelo número de imagens processadas por segundo.

| | **Carga** | **Extração** | **Extração Red.** | **Classificação** |
|---|---|---|---|---|
| *V. de C. de Cor e Histograma* | 36,14 | 20,52 | 251,92 | 1501,467 |
| *V. de C. de Cor e Histograma C.* | 36,99 | 20,51 | 245,52 | 892,26 |
| *V. de C. de Cor e Momento de C.* | 36,83 | 19,14 | 245,02 | 4.360,51 |
| *Histograma e Momento de C.* | 36,78 | 75,30 | 621,30 | 1.814,29 |
| *Histograma C. e Momento de C.* | 36,64 | 77,56 | 631,63 | 1.211,52 |

Os tempos computacionais para carregamento das imagens deveriam ser idênticos em todas as linhas das tabelas 23 e 24. As pequenas diferenças ocorrem devido ao fato dos estudos terem sido realizados sobre um sistema operacional multi-tarefas (no caso, Microsoft Windows 7), que pode escalonar processos de forma preemptiva.

Em ambas as tabelas, os tempos computacionais mais baixos são obtidos no processo de classificação das imagens. Os mais altos, com exceção para a característica do vetor de coerência de cor, são obtidos no processo de carregamento das imagens.

Pelos valores apresentados em ambas as tabelas, é possível notar que o tempo computacional para a extração das características é gasto majoritariamente com a varredura (*scanning*) dos *pixels* das imagens. Para a extração das características momento de cor e vetor de coerência de cor, os *pixels* das imagens são varridos, respectivamente, por

duas e por diversas vezes. Por isto, os tempos computacionais para extração destas duas características são superiores aos de extração das demais características.

O algoritmo utilizado, nos estudos, para extração do vetor de coerência de cor é recursivo. Por isto, varre os *pixels* das imagens por diversas vezes. O algoritmo foi escolhido devido unicamente a sua simplicidade. São descritos na literatura, porém, algoritmos que realizam apenas uma única varredura nos *pixels* (PASS; ZABIH; MILLER, 1996).

Pelos valores apresentados em ambas as tabelas, é possível notar, igualmente, que a redução da resolução das imagens reduz o tempo computacional para extração de suas características. Este resultado, logicamente esperado, é consequência direta da menor quantidade de *pixels* contida nas imagens reduzidas.

# 5 Conclusão

Nos últimos anos, o correio eletrônico ou *e-mail* tornou-se um dos meios de comunicação mais populares. É utilizado pelo público em geral, no ambiente corporativo e no meio acadêmico, local onde foi criado. Com sua grande popularização, surgiu também seu uso indevido, causado, em grande parte, pelo fato do protocolo do serviço de *e-mails*— SMTP — ter sido projetado para uso em redes com sistemas computacionais confiáveis.

Na década de 90, usuários do serviço passaram a receber *e-mails* indesejados, tais como mensagens com propaganda de serviços ou mensagens maliciosas, que tentam enganar ou roubar dados dos usuários. Estas mensagens foram apelidadas de *spam*.

Inicialmente, as mensagens *spam* causaram incômodos pequenos e ocasionais. Com o tempo, porém, o volume de mensagens *spam* superou o de mensagens legítimas, também conhecidas como mensagens *ham*. Mensagens *spam* causam, atualmente, diversos problemas, tais como perda de tempo para seu descarte, sobrecarga de servidores e outros dispositivos de rede, bem como transtornos causados por *spams* fraudulentos.

Os transtornos e prejuízos causados por mensagens *spam* incentivou o desenvolvimento de sistemas anti-spam (SAS). Estes sistemas de software têm, por objetivo, identificar e filtrar mensagens *spam*, evitando que cheguem aos usuários. Os SAS tanto podem executar nos servidores de *e-mail*, removendo mensagens antes delas alcançarem as caixas de correio (*mailboxes*) dos usuários, quanto podem executar nos clientes de *e-mail* dos usuários.

Um dos objetivos de pesquisa na área, é reduzir a ocorrência de falsos positivos nas classificações de *e-mails* realizadas por sistemas anti-spam. Um falso positivo é a classificação incorreta, como *spam*, de um *e-mail* legítimo do usuário. Falsos positivos causam transtornos aos usuários, uma vez que estes, por erro do SAS, podem desconsiderar mensagens legítimas importantes. Embora causem transtornos menos graves aos usuários, falsos negativos também os incomodam. Um falso negativo é a classificação incorreta, como *e-mail* legítimo do usuário, de um *spam*.

Sistemas anti-spam alcançam, hoje em dia, razoáveis taxas de acerto na classificação de *e-mails* contendo mensagens de texto. Esta eficiência na classificação não passou despercebida aos *spammers*[1]. Para iludirem os SAS, *spammers* passaram a enviar mensagens com imagens *spam* em imagens anexas ao *e-mail* ou inclusas em seu próprio corpo de mensagem.

Uma imagem é usualmente representada por um vetor de números inteiros. Os valores contidos nos elementos do vetor especificam as cores dos pontos (*pixels*) da imagem. Assim, a quantidade de informação necessária para representação de uma imagem é significativamente maior que a necessária para representar uma mensagem em texto. O custo computacional dos sistemas anti-spam para classificação de imagens é, portanto, maior. São, igualmente, maiores os prejuízos causados por imagens *spam*, devido ao maior uso de banda de rede e ao maior tempo de processamento exigido dos servidores de *e-mail*.

A disseminação de imagens *spam* na Internet induziu a pesquisa de sistemas capazes de detectá-las e filtrá-las. Contudo, muitos dos sistemas propostos apresentam sérios inconvenientes, tais como custo computacional elevado e altas taxas de falsos positivos.

Métodos OCR (*optical character recognition*) são empregados para extração de textos contidos em imagens. Sistemas anti-spam que empregam OCR partem do princípio de que uma imagem *spam* vai conter textos *spam*. Assim, após a extração dos textos, estes sistemas ou classificam os textos ou encaminham-nos para SAS textuais convencionais. A inclusão de métodos OCR em sistemas anti-spam costuma ser ineficaz, por dois motivos principais. Primeiro, porque os *spammers* facilmente os confundem, ao empregar ruídos e outros artifícios nos textos das imagens. Segundo, devido ao alto custo computacional destes métodos, inviabilizando o uso em servidores com alto volume de mensagens.

Sistemas anti-spam podem utilizar diversos outros métodos para extração de características não-textuais das imagens. Representam as imagens por intermédio de suas características e classificam-nas como *ham* ou *spam*. Métodos para extração de características têm demonstrado ser bem mais eficientes que métodos OCR.

A maioria das pesquisas com sistemas para detecção de imagens *spam* reportadas na literatura utiliza bases privadas de imagens. Poucas são as pesquisas que utilizam bases públicas. Isto é devido ao fato de existirem poucas bases públicas contendo imagens *spam*. Além disto, devido a questões de privacidade, estas bases ou contêm muito poucas imagens *ham* ou não as contêm absolutamente. A existência de poucas bases públicas contendo tanto imagens *ham* quanto *spam* dificulta a comparação dos resultados das

[1]Entidades que produzem ou enviam *spam*.

pesquisas reportadas na literatura, não permitindo uma avaliação precisa e abrangente da capacidade representacional de cada característica de imagem empregada nestas pesquisas.

Este trabalho de pesquisa de dissertação propõe um sistema anti-spam de imagens. O sistema é composto por dois estágios. No primeiro, alguns métodos são empregados para extração de características das imagens. No segundo estágio, as imagens, representadas por suas características, são classificadas, através de Redes Neurais Artificiais, como imagens *ham* ou *spam*.

Para avaliação da capacidade representacional das características, foram utilizadas apenas bases públicas de imagens *ham* e *spam*, já utilizadas em algumas poucas pesquisas reportadas na literatura. Com isto, além de possibilitar a comparação dos resultados desta pesquisa com os destas poucas pesquisas reportadas na literatura, esta pesquisa de dissertação também permite que seus resultados sejam comparados com os de pesquisas futuras que empreguem as mesmas bases utilizadas e disponibilizadas por esta pesquisa.

Cinco métodos de extração de características — RGB médio, histograma em tons de cinza, histograma colorido, momento de cor e vetor de coerência de cor — foram utilizados e avaliados. O custo computacional e a capacidade representacional de cada uma destas características são avaliados através de cinco métricas — tempo gasto para extração da característica, tempo gasto para treinamento da Rede Neural Artificial, tempo gasto para classificação das imagens, taxa de detecção de *spam* e taxa de falsos positivos.

Os primeiros cinco estudos avaliam as características individualmente. Os cinco últimos, avaliam-nas combinadas. Nos dez estudos as características são extraídas tanto das imagens originais contidas nas bases quanto destas imagens reduzidas à resolução de 100x100 *pixels*. A redução da resolução foi estudada porque a literatura reporta que este processo reduz ruídos das imagens.

Na avaliação individual das características, a característica histograma colorido foi a que apresentou os melhores resultados. Com o uso desta característica, a Rede Neural Artificial produziu as melhores taxas de detecção de *spam* e de falsos positivos. As características histograma (em tons de cinza) e vetor de coerência de cor também apresentam bons resultados.

Na avaliação combinada das características, a combinação das características vetor de coerência de cor e histograma foi a que apresentou os melhores resultados. Com o uso desta combinação, a Rede Neural Artificial produziu as melhores taxas de detecção de *spam* e de falsos positivos. Com o uso combinado de características, a Rede Neu-

ral Artificial produz resultados superiores aos produzidos com o uso individual de cada característica. Estes resultados sugerem que cada característica fornece informação representacional própria, exclusiva, das imagens, informação esta não fornecida por qualquer uma das demais características.

Os resultados obtidos nos estudos são promissores. São similares e, em alguns casos, superiores aos de estudos, reportados na literatura, que utilizam ou métodos de extração de características significativamente mais complexos ou proporções entre os tamanhos dos conjuntos de treinamento e teste mais favoráveis que as utilizadas neste trabalho.

Por fim, é relevante destacar que o sistema anti-spam de imagens, desenvolvido em linguagem Java, para este trabalho de pesquisa, possui arquitetura modular. Esta arquitetura permite facilmente não só a adição de novos módulos que implementem, por exemplo, outros métodos de extração de características, como também a conversão do sistema em um produto tecnológico com licença livre.

## 5.1 Trabalhos Futuros

Algumas sugestões podem ser dadas para dar prosseguimento a pesquisa realizada neste trabalho. Primeira, o estudo e avaliação de novas características para representação das imagens, bem como dos métodos para extraí-las. Segunda, o estudo e avaliação do uso combinado de três ou mais características para representação das imagens.

Terceira, o estudo e avaliação de outras bibliotecas para processamento de imagens, como, por exemplo, a OpenCV. Quarta, o estudo e avaliação do uso de GPUs (*Graphic Processing Units*) para extração de características das imagens. O uso de GPUs pode também possibilitar o uso de características mais complexas, que seriam inviáveis em sistemas anti-spam, devido a seus altos custos computacionais.

Quinta, análise do comportamento das técnicas estudadas com o acréscimo de ruídos nas imagens. Sexta, utilização de diferentes espaços de cores com as técnicas de extração. Sétima, o estudo e avaliação de outros modelos para classificação das imagens, como, por exemplo, máquinas de vetor de suporte. Oitavo, o estudo e avaliação da integração do sistema anti-spam de imagens desenvolvido a um sistema anti-spam de mensagens de texto.

Nono, o estudo e avaliação do desempenho do sistema anti-spam desenvolvido em um ambiente real de produção. Por fim, a conversão do sistema desenvolvido em um produto

tecnológico com licença livre.

# Referências

AL-DUWAIRI, B.; KHATER, I.; AL-JARRAH, O. Texture analysis-based image spam filtering. In: IEEE. *Internet Technology and Secured Transactions (ICITST), 2011 International Conference for*. [S.l.], 2011. p. 288–293.

ARADHYE, H. B.; MYERS, G. K.; HERSON, J. A. Image analysis for efficient categorization of image-based spam e-mail. In: *Document Analysis and Recognition, 2005. Proceedings. Eighth International Conference on*. [S.l.: s.n.], 2005. p. 914 – 918 Vol. 2. ISSN 1520-5263.

ATTAR, A.; RAD, R. M.; ATANI, R. E. A survey of image spamming and filtering techniques. *Artificial Intelligence Review*, Springer Netherlands, v. 40, n. 1, p. 71–105, 2013. ISSN 0269-2821.

BIGGIO, B. et al. Improving image spam filtering using image text features. In: *Proceedings of the Fifth Conference on Email and Anti-Spam (CEAS)*. [S.l.: s.n.], 2008.

BURGER, W.; BURGE, M. J. *Digital Image Processing, An Algorithm Introduction Using Java*. [S.l.]: Springer, 2008. (Texts in Computer Science). ISBN 9781846283796.

BYUN, B. et al. A discriminative classifier learning approach to image modeling and spam image identification. In: CITESEER. *CEAS*. [S.l.], 2007.

CAMPBELL, K. *A NET.CONSPIRACY SO IMMENSE...* 1994. `http://w2.eff.org/legal/cases/Canter_Siegel/`. "[Online; acessado 12/04/2013]".

CHENG, H. et al. A novel spam image filtering framework with multi-label classification. In: IEEE. *Communications, Circuits and Systems (ICCCAS), 2010 International Conference on*. [S.l.], 2010. p. 282–285.

CHENG, H. et al. Spam image discrimination using support vector machine based on higher-order local autocorrelation feature extraction. In: IEEE. *Cybernetics and Intelligent Systems, 2008 IEEE Conference on*. [S.l.], 2008. p. 1017–1021.

CORMACK, G. V.; LYNAM, T. R. *TREC 2005 Spam Public Corpora*. 2005. `http://plg.uwaterloo.ca/~gvcormac/treccorpus/`. "[Online; acessado 12/04/2013]".

DHANARAJ, S.; KARTHIKEYANI, V. A study on e-mail image spam filtering techniques. In: IEEE. *Pattern Recognition, Informatics and Medical Engineering (PRIME), 2013 International Conference on*. [S.l.], 2013. p. 49–55.

DREDZE, M.; GEVARYAHU, R.; ELIAS-BACHRACH, A. Learning fast classifiers for image spam. In: *Conference on Email and Anti-Spam*. [S.l.: s.n.], 2007.

FUMERA, G.; PILLAI, I.; ROLI, F. Spam filtering based on the analysis of text information embedded into images. *J. Mach. Learn. Res.*, JMLR.org, v. 7, p. 2699–2720, dez. 2006. ISSN 1532-4435. Disponível em: <http://dl.acm.org/citation.cfm?id=1248547.1248645>.

HAYATI, P.; POTDAR, V. Evaluation of spam detection and prevention frameworks for email and image spam: a state of art. In: *Proceedings of the 10th International Conference on Information Integration and Web-based Applications & Services*. New York, NY, USA: ACM, 2008. (iiWAS '08), p. 520–527. ISBN 978-1-60558-349-5. Disponível em: <http://doi.acm.org/10.1145/1497308.1497402>.

HAZZA, Z. M.; AZIZ, N. A. Detecting computer generated images for image spam filtering. In: *Proceedings of the 2012 International Conference on Advanced Computer Science Applications and Technologies*. Washington, DC, USA: IEEE Computer Society, 2012. (ACSAT '12), p. 313–317. ISBN 978-0-7695-4959-0. Disponível em: <http://dx.doi.org/10.1109/ACSAT.2012.38>.

LIU, T.-J.; TSAO, W.-L.; LEE, C.-L. A high performance image-spam filtering system. In: IEEE. *Distributed Computing and Applications to Business Engineering and Science (DCABES), 2010 Ninth International Symposium on*. [S.l.], 2010. p. 445–449.

MEHTA, B. et al. Detecting image spam using visual features and near duplicate detection. In: ACM. *Proceedings of the 17th international conference on World Wide Web*. New York, NY, USA: ACM, 2008. (WWW '08), p. 497–506. ISBN 978-1-60558-085-2. Disponível em: <http://doi.acm.org/10.1145/1367497.1367565>.

PASS, G.; ZABIH, R.; MILLER, J. Comparing images using color coherence vectors. In: . New York, NY, USA: ACM, 1996. (MULTIMEDIA '96), p. 65–73. ISBN 0-89791-871-1.

QU, Z.; ZHANG, Y. Filtering image spam using image semantics and near-duplicate detection. In: IEEE. *Intelligent Computation Technology and Automation, 2009. ICICTA '09. Second International Conference on*. [S.l.], 2009. v. 1, p. 600–603.

RUMELHART, D. E.; HINTON, G. E.; WILLIAMS, R. J. *Learning internal representations by error propagation*. [S.l.], 1985. 318-362 p.

RUSSEL, S. J.; NORVIG, P. *Artificial Intelligence: A Modern Approach*. First edition. [S.l.]: Prentice Hall, 1995. ISBN 978-0131038059.

SORANAMAGESWARI, M.; MEENA, C. Statistical feature extraction for classification of image spam using artificial neural networks. In: IEEE. *Machine Learning and Computing (ICMLC), 2010 Second International Conference on*. [S.l.], 2010. p. 101–105.

SORANAMAGESWARI, M.; MEENA, C. A novel approach towards image spam classification. *International Journal of Computer Theory and Engineering*, v. 3, n. 1, p. 84–88, 2011.

STERN, H. A survey of modern spam tools. In: CITESEER. *CEAS*. [S.l.], 2008.

STRICKER, M.; ORENGO, M. Similarity of color images. In: INTERNATIONAL SOCIETY FOR OPTICS AND PHOTONICS. *IS&T/SPIE's Symposium on Electronic Imaging: Science & Technology*. [S.l.], 1995. p. 381–392.

SWAIN, M. J.; FRANKEL, C.; ATHITSOS, V. Webseer: An image search engine for the world wide web. University of Chicago, Chicago, IL, USA, 1996.

WANG, C. et al. Image spam classification based on low-level image features. In: IEEE. *Communications, Circuits and Systems (ICCCAS), 2010 International Conference on.* [S.l.], 2010. p. 290–293.

WU, C.-t. et al. Using visual features for anti-spam filtering. In: IEEE. *Image Processing, 2005. ICIP 2005. IEEE International Conference on.* [S.l.], 2005. v. 3, p. 509–512.

ZHEN, X.; HONG-GUO, W.; ZENG-ZHEN, S. Evaluation of image spam classification system based on ahp. In: IEEE. *Computational Intelligence and Software Engineering, 2009. CiSE 2009. International Conference on.* [S.l.], 2009. p. 1–4.