

FACULDADE DE ENGENHARIA DA UNIVERSIDADE DO PORTO



Olhó-passarinho: uma extensão do TweeProfiles para fotografias

Ivo Filipe Valente Mota

PARA APRECIAÇÃO POR JÚRI

Mestrado Integrado em Engenharia Eletrotécnica e de Computadores

Orientador: Luís Filipe Pinto de Almeida Teixeira (PhD)

Co-orientador: Carlos Manuel Milheiro de Oliveira Pinto Soares (PhD)

7 de Julho de 2014

© Ivo Mota, 2014

Resumo

O Twitter é uma das redes sociais atuais que mais informação gera todos os dias. Face à sua dimensão, foi desenvolvido o TweeProfiles, uma ferramenta que recorre a esta rede social, mais concretamente às mensagens partilhadas neste serviço. Esta ferramenta utiliza técnicas de *Data Mining* para identificar padrões, apresentados através de *clusters* de Tweets, em que são analisados, o conteúdo na forma de texto, as ligações sociais, e as dimensões espaço-temporais das mensagens.

Face ao aumento do número de utilizadores que recorrem a smartphones para acederem ao Twitter, o número de fotografias partilhadas neste serviço tem crescido significativamente nos últimos anos. Esta dissertação teve como objetivo principal o desenvolvimento de uma extensão da ferramenta TweeProfiles, através de técnicas de visão por computador e *Data Mining*, que permita a identificação de padrões espaço-temporais através da informação das imagens partilhadas no serviço de *microblogging* Twitter. Para a sua concretização foi desenvolvido um módulo que utiliza o conceito de vocabulário visual para a representação das imagens de uma forma mais compacta e eficiente.

Os resultados obtidos podem ser visualizados através de uma aplicação web que permite a navegação e visualização pelas imagens e dimensões espaço-temporais dos *clusters*.

Abstract

Twitter is one of social networks that generates more information every day. Due to it is dimension, was developed a tool called TweeProfiles, which uses this social network, more specifically the messages shared in this service. This tool uses data mining techniques to identify patterns presented through clusters of Tweets, each of them are analyzed in their respective: content as text, social connections, and dimensions of spatial and temporal messages.

Given the increasing number of users who use smartphones to access Twitter, the number of shared photos in this service has grown significantly in recent years. This thesis had the main objective of developing an extension of TweeProfiles tool. Through techniques of computer vision and data mining, this tool allow the identification of spatiotemporal patterns using all information in the shared images in the microblogging service Twitter. For its implementation, a module was developed using the concept of visual vocabulary for representing images in a more compact and efficient way.

The results can be visualized through a web application that allows browsing and viewing the images and spatial and temporal dimensions of clusters.

Agradecimentos

Em primeiro lugar quero deixar os meus agradecimentos aos meus orientadores, Professor Luís Filipe Teixeira e Professor Carlos Soares, que foram o pilar para desenvolvimento deste projeto de dissertação, com

Ivo Mota

This work is partially funded by National Funds through the FCT – Fundação para a Ciência e a Tecnologia (Portuguese Foundation for Science and Technology) within projects "REACTION (UTAustin/EST-MAI/0006/2009)" and "POPSTAR (PTDC/CPJ-CPO/116888/2010)" as well as Project "NORTE-07-0124-FEDER-000059", which is funded by the North Portugal Regional Operational Programme (ON.2 – O Novo Norte), under the National Strategic Reference Framework (NSRF), through the European Regional Development Fund (ERDF), and the Portuguese funding agency, Fundação para a Ciência e a Tecnologia (FCT).

*"Logic will get you from A to Z.
Imagination will get you everywhere."*

Albert Einstein

Conteúdo

Agradecimentos	v
1 Introdução	1
1.1 Contexto	1
1.2 Motivação	2
1.3 Objetivos	2
1.4 Estrutura do documento	2
2 Conceitos e Trabalhos Relacionados	5
2.1 Clustering	5
2.1.1 Clustering por Partição	6
2.1.2 Clustering Hierárquico	7
2.1.3 Clustering Baseado em Densidade	9
2.1.4 Clustering Baseado em Grelhas	10
2.1.5 Funções de Distância	10
2.2 TweeProfiles	13
2.2.1 Descrição e objetivos	13
2.2.2 Resultados ilustrativos	14
2.2.3 Prós e contras	16
2.3 Representação de Informação Visual	18
2.3.1 Representação Matricial	18
2.3.2 Histogramas	19
2.3.3 Descritores de Cor	19
2.3.4 Descritores de Textura	22
2.3.5 Descritores de Forma	23
2.3.6 Descritores Locais	24
2.3.7 Descritores Baseado em Vocabulário Visual	29
3 Módulo da Informação Visual	31
3.1 Recolha dos Dados	31
3.1.1 Descrição dos Dados	31
3.1.2 Filtragem dos Dados	32
3.1.3 Conjunto de Dados Final	33
3.2 Extração, Processamento e Armazenamento da Informação Visual	34
3.2.1 Extração dos Pontos de Interesse e Descritores Locais	34
3.2.2 Criação do Vocabulário Visual	36
3.2.3 Armazenamento da Informação Visual	36
3.3 Matriz de Distâncias entre Imagens	37

4	Olhó-passarinho	39
4.1	Arquitetura do Sistema	39
4.2	Informação Espaço-Temporal	40
4.3	Clustering da Informação Visual, Espacial e Temporal	41
4.4	Visualização	42
4.5	Resultados Ilustrativos	43
5	Conclusões e Trabalho Futuro	47
5.1	Resumo do Trabalho Realizado	47
5.2	Trabalho Futuro	48
A	Exemplo objeto JSON de um tweet	49
	Referências	55

Listas de Figuras

2.1	Dendrograma ilustrativo da divisão entre nós do conjunto completo ao objeto individual no fundo [1]	8
2.2	Matriz confusão de dois objetos com atributos binários	12
2.3	Exemplo ilustrativo da distribuição espacial dos <i>clusters</i> calculada apenas com a consideração da dimensão espacial. <i>Retirada de</i> [10]	14
2.4	Exemplo ilustrativo da distribuição temporal de um <i>cluster</i> . <i>Retirada de</i> [10]	15
2.5	Exemplos de resultados de <i>clusters</i> com pesos diferentes para as diferentes dimensões. <i>Retiradas de</i> [10]	15
2.6	<i>Clusters</i> em Portugal: Conteúdo 25% + Espacial 25% + Temporal 25% + Social 25%. <i>Retirada de</i> [10]	16
2.7	Visualização de informação mais detalhada de um <i>cluster</i> incluindo o seu grafo social, e da informação relativa a um determinado tweet. <i>Retirada de</i> [10]	16
2.8	Tweets pertencentes a um <i>cluster</i> para a seguinte distribuição de pesos: Conteúdo 25% + Espacial 25% + Temporal 25% + Social 25%. <i>Retirada de</i> [10]	17
2.9	Imagen em tons cinza e respetivo histograma	19
2.10	Sub-imagem e bloco de imagem. <i>Retirada de</i> [2]	23
2.11	Exemplo de formas de objetos que podem ser descritas eficazmente pelo descritor baseado em região. <i>Retirada de</i> [3]	24
2.12	Construção das Diferenças Gaussianas e formação de oitavas <i>Retirada de</i> [4]	26
2.13	Ilustração do processo de deteção de máximos e mínimos das imagens de Diferença Gaussiana. O pixel candidato está marcado com X e os vizinhos com um circulo. <i>Retirada de</i> [4]	27
2.14	Imagen	28
2.15	(a) Filtros Gaussianos de segunda ordem nas direções yy e xy; (b) aproximação por filtros de caixa; (c) filtros de Haar; As regiões a cinzento têm valor igual a zero. <i>Retirada de</i> [5].	29
3.1	Arquitetura do módulo de extração, processamento e armazenamento da informação visual. <i>Adaptada de</i> [6]	35
3.2	Exemplo de projeção de centroides após tarefa de <i>clustering</i> num espaço a duas dimensões.	37
3.3	Estrutura de uma matriz de distâncias. <i>Retirada de</i> [7]	38
4.1	Arquitetura do sistema completo	40
4.2	Distribuição de tweets na dimensão espacial	42
4.3	Exemplo ilustrativo da ferramenta Timeline. <i>Retirada de</i> [8]	42
4.4	Exemplo ilustrativo da visualização da distribuição espacial dos <i>clusters</i> para o caso em que a distribuição dos pesos entre as três dimensões é igual	43

4.5 Exemplo ilustrativo da visualização da distribuição temporal dos <i>clusters</i> para o caso em que a distribuição dos pesos entre as três dimensões é igual	44
4.6 Exemplo ilustrativo da visualização completa da aplicação web com a visualização pelas várias dimensões	44
4.7 Exemplo ilustrativo da visualização completa da aplicação web com a visualização mais detalhada de uma das imagens	45
4.8 Exemplo ilustrativo da visualização de um conjunto de imagens pertencentes a um <i>Cluster</i> e em que foi atribuído o peso total apenas à dimensão do conteúdo visual	46

Lista de Tabelas

3.1	Descrição em números do total de tweets com indicação, nos que contém URL para imagem, do número de tweets por serviço de partilha de imagem	32
3.2	Esquema da base de dados SQLite	38
4.1	Tabela com as possíveis distribuições de pesos entre as várias dimensões	41

Abreviaturas e Símbolos

BoW	<i>Bag Of Words</i>
DBSCAN	<i>Density-Based Spatial Clustering of Applications with Noise</i>
DoG	<i>Diference of Gaussian</i>
GLOH	<i>Gradient Location and Orientation Histogram</i>
HOG	<i>Histogram of Oriented Gradients</i>
MS	<i>Maximally Stable</i>
SA	<i>Shape Adapted</i>
SIFT	<i>Scale-Invariant Feature Transform</i>
SURF	<i>Speeded Up Robust Features</i>

Capítulo 1

Introdução

Neste capítulo é feita uma introdução ao projeto desenvolvido no âmbito da disciplina de dissertação com a apresentação do seu contexto, a motivação para o seu desenvolvimento, os objetivos a alcançar e a descrição da estrutura deste documento.

1.1 Contexto

As redes sociais são uma excelente fonte de informação sempre em atualização, que fornecem aos investigadores uma vasta quantidade e variedade de dados. Estes dados apresentam-se de diferentes formas como textos, imagens e vídeos.

O Twitter faz parte do grupo de redes sociais existentes que mais informação gera todos os dias, sendo caracterizado como um serviço de microblogging, que permite aos utilizadores partilharem mensagens, designadas por tweets, até um máximo de 140 caracteres. Essas mensagens podem conter, para além de texto, imagens ou links para imagens de outros serviços, como por exemplo, o Instagram ou o Twitpic.

Ao contrário do que acontece com outras redes sociais como o Facebook e LinkedIn que utilizam uma rede de comunicação bi-direcional, o Twitter utiliza uma infraestrutura assimétrica onde existem os "*friends*" e os "*followers*". Supondo que é um utilizador do Twitter, os "*friends*" corresponde às contas das pessoas que o utilizador segue e os "*followers*" corresponde às contas das pessoas que o seguem [9].

O TweeProfiles [10], é uma ferramenta que recorre à rede social Twitter com o objetivo de identificar padrões em mensagens escritas em português partilhadas nesta rede social. Esta ferramenta utiliza técnicas de *Data Mining*, mais precisamente de *Text Mining*. A principal característica do TweeProfiles é o facto de utilizar a tarefa de *clustering* para identificar padrões em mensagens partilhadas no Twitter, através do conteúdo das mensagens (o texto) e das dimensões espaço-temporais das mesmas.

1.2 Motivação

Devido ao grande número de utilizadores e de informação partilhada a todo o instante no Twitter, este torna-se um excelente serviço de recolha de dados, proporcionando aos investigadores e empresas uma quantidade e variedade de dados necessários para o desenvolvimento de ferramentas de análise de dados e extração de conhecimento.

As mensagens partilhadas no Twitter sobre a forma de texto têm sido uma das grandes fontes de dados utilizadas por muitas ferramentas como o TweetProfiles [10], mas apesar de se tratar de uma rede social em que a maioria da informação disponível se encontra forma de texto, o Twitter também permite a partilha de imagens a partir do seu próprio serviço, ou através de outros serviços como Twitpic ou Instagram. Estas imagens também podem ser utilizadas para a análise e extração de conhecimento, pois o seu conteúdo pode mesmo em muitos casos complementar o texto ou até mesmo, o substituir.

A análise de informação visual é assim um acréscimo importante para o desenvolvimento de ferramentas de extração de conhecimento através de redes sociais.

1.3 Objetivos

Esta dissertação tem como principal objetivo a criação de uma extensão para o TweeProfiles através de técnicas de visão por computador e *Data Mining*, que permita a identificação de padrões em imagens partilhadas no serviço de microblogging Twitter, através da representação de *clusters*.

Será assim necessário realizar a recolha dos dados alojados numa base de dados Mongodb criada através da plataforma Socialbus (anteriormente designada por TwitterEcho [11]), que consiste num projeto open source, responsável por extrair e armazenar tweets de uma determinada comunidade de utilizadores, tendo sido desenvolvido com o intuito de ajudar os investigadores a terem facilidade de acesso a uma base de dados de redes sociais, na sua maioria, com texto na língua portuguesa. Após recolhidos os dados será necessário o desenvolvimento de um módulo responsável pela recolha das imagens através do *URL* existente nos tweets, do processamento da informação visual de modo a torna-la mais compacta e eficiente, e do armazenamento dessa informação. Por fim, a informação visual deverá ser integrada na ferramenta TweeProfiles, com objetivo de realizar o processo de *Data Mining*, mais especificamente a tarefa de *clustering* e desenvolver a aplicação para visualizar os *clusters* nas diferentes dimensões.

1.4 Estrutura do documento

O documento está organizado da seguinte forma: o Capítulo 2 descreve conceitos e trabalhos relacionados, onde é apresentado uma pesquisa sobre os vários domínios científicos relacionados com as necessidades para o desenvolvimento do projeto de dissertação. No Capítulo 3 é apresentado o modelo desenvolvido para a extração, processamento e armazenamento da informação visual. Já no Capítulo 4 é apresentado a ferramenta Olhó-passarinho com a integração do módulo

desenvolvido para a informação visual com a ferramenta TweeProfiles. Para finalizar é apresentado o Capítulo 5 com as conclusões retiradas do desenvolvimento deste projeto de dissertação e o trabalho futuro possível para dar continuidade ao trabalho desenvolvido.

Capítulo 2

Conceitos e Trabalhos Relacionados

Neste capítulo é apresentado o estudo realizado, tendo em vista a aquisição de competências e conhecimentos necessários para o desenvolvimento do projeto, que se focam essencialmente em análise de métodos de *data mining* e em técnicas aplicadas em visão por computador. Em primeiro lugar será exposto conteúdo relativamente a métodos de *clustering* como uma tarefa de *data mining*. Em seguida serão apresentadas formas de representação de imagens. Por fim, são referenciados alguns trabalhos relacionados com o projeto a desenvolver nesta dissertação.

2.1 Clustering

Extração de conhecimento em base de dados ou *Data Mining* é um processo de exploração de grandes quantidades de dados que procura encontrar padrões "interessantes" [7]. Trata-te assim de uma fusão de estatística aplicada, sistemas de lógica, inteligência artificial, *Machine Learning* e gestão de base de dados [12]. Este processo é caracterizado por várias tarefas possíveis de ser aplicadas, dependendo do problema abordado, tais como [13]:

- Deteção de anomalias (outliers/ alterações/ desvios) - Identifica regtos de dados incomuns, podendo ser erros nos dados ou objetos interessantes que apresentam comportamento diferente dos restantes;
- Regras de associação - Procura relações entre variáveis que ocorram frequentemente;
- Classificação - É a tarefa de generalizar uma estrutura conhecida e aplicar a novos dados, sendo essencialmente utilizada em tarefas de previsão;
- Regressão - Tenta encontrar uma função que modela os dados com o mínimo de erro;
- Resumo - Trata-se da representação mais compacta do conjunto de dados, que pode incluir visualização e descrição através de um relatório,
- *Clustering* - Tarefa de descobrir grupos em que os dados apresentam de alguma forma semelhanças, sem o uso de estruturas previamente conhecidas

Este projeto terá como uma das principais tarefas a realização *clustering* sobre dados recolhidos e tratados de fotografias partilhadas no Twitter. Pode-se definir *clustering* como "um processo de agrupamento de um conjunto de objetos de dados em vários grupos ou *clusters*, de modo que os objetos dentro de um *cluster* apresentem alta similaridade, mas que sejam muito diferentes de objetos de outros *clusters*. Diferenças e semelhanças são avaliados com base nos valores de atributos que descrevem os objetos e muitas vezes envolvem medidas de distância" [7].

A realização de *clustering* é assim uma escolha lógica para a extração de padrões em dados não supervisionados e para o agrupamento de *tweets* pela sua semelhança em conteúdo, neste caso as imagens, e com a integração de outras dimensões, como o tempo e espaço.

O *clustering* faz parte de um conjunto de técnicas aplicadas na aprendizagem não supervisionada. Enquanto que na aprendizagem supervisionada existe um conjunto de dados previamente analisados e rotulados que são usados para treinar um modelo capaz de encontrar relação entre os atributos desses dados com novos conjuntos de dados, na aprendizagem não supervisionada, não são utilizados conjuntos de dados previamente analisados e rotulados. Assim o processo de descoberta de padrões nos dados apenas tem em conta os dados presentes, tentando organizar as instâncias em grupos semelhantes [14].

Nesta secção serão apresentadas as principais características e técnicas para aplicação da tarefa de *clustering*, tais como, *clustering* por partição, *clustering* hierárquico, *clustering* baseado em densidade, *clustering* baseado em grelhas, funções de distância para o cálculo da similaridade entre objetos e por fim, a avaliação de *clusters*.

2.1.1 Clustering por Partição

A utilização de métodos baseados em partições é a forma mais simples e elementar de realizar análise por *clustering*, em que um conjunto de objetos é distribuído em vários grupos ou *clusters* mais pequenos. É assumido que o número de *clusters* é conhecido antes da realização da tarefa, sendo esse valor tomado como o ponto de partida para aplicação de métodos baseados em partição [7].

O algoritmo *k-means* é o melhor algoritmo de *clustering* por partição e o mais utilizado devido à sua simplicidade e eficiência [14]. É apresentado como sendo um algoritmo de *clustering* por partição, pois este divide o conjunto de dados em partições mais pequenas, formando assim os *clusters*.

Inicialmente é necessário que o utilizador indique o valor de *k* e o algoritmo irá iterativamente dividir o conjunto de objetos em *k-clusters* diferentes, baseado em funções de distância [14] que são apresentadas na secção 2.1.5.

Cada *cluster* apresenta um centroide que é o representante do grupo, sendo o valor médio de todos os objetos (instâncias) pertencentes ao *cluster*. Este centroide é recalculado de forma iterativa até que seja atingido o critério de paragem. A convergência ou critério de paragem pode ser um dos seguintes enumerados:

1. Não ocorre (ou ocorre um valor mínimo) de alterações dos objetos para diferentes *clusters*.

2. Não ocorre (ou ocorre um valor mínimo) de alterações dos centroides.
3. Diminuição mínima da **soma do erro quadrático** (SEQ),

$$SEQ = \sum_{j=1}^k \sum_{x \in C_j} dist(x, m_j)^2, \quad (2.1)$$

onde k é o número de *clusters* pretendidos, C_j é i-ésimo *cluster*, m_j é o centroide do *cluster* C_j e $dist(x, m_j)$ é a distância entre uma instância x e o centroide m_j .

Assim, "o algoritmo *k-means* pode ser usado em qualquer aplicação com um conjunto de dados onde a média pode ser definida e calculada" [14].

No **espaço euclidiano**, centroide de um *cluster* é calculado da seguinte forma:

$$m_j = \frac{1}{|C_j|} \sum_{x_i \in C_j} x_i, \quad (2.2)$$

onde $|C_j|$ é o número de pontos (instâncias) no *cluster* C_j . A distância entre um ponto x_i a um centroide m_j é calculado da seguinte forma:

$$dist(x_i, m_j) = \|x_i - m_j\| = \sqrt{(x_{i1} - m_{j1})^2 + (x_{i2} - m_{j2})^2 + \dots + (x_{ir} - m_{jr})^2}. \quad (2.3)$$

O pseudo-código deste algoritmo é apresentado no algoritmo 1.

Algoritmo 1 K-Means

```

1: procedure K-MEANS( $k$ : clusters,  $D$ : conjunto de dados)
2:   escolher  $k$  objetos de  $D$  como centroides dos clusters iniciais;
3:   repeat
4:     (re) atribuir cada objeto a ao cluster ao qual o objeto é o mais similar;
5:     actualizar o centroide do cluster;
6:   until clusters sem alterações;
7:   return conjunto de  $k$  clusters;
8: end procedure
```

2.1.2 Clustering Hierárquico

O *clustering* hierárquico é outra abordagem importante na tarefa de *clustering*. Os *clusters* são criados sobre a forma de uma sequência em árvore (dendrograma). Os objetos (instâncias) encontram-se no fundo do diagrama, enquanto que o conjunto de todos os objetos encontra-se no topo do diagrama. Cada nó que se encontra no interior do diagrama possui nós filhos, sendo que cada nó representa um *cluster*. Assim designam-se por *clusters* irmãos, aqueles que derivam de um mesmo *cluster*, isto é, do nó parente [14]. A Figura 2.1 exemplifica a representação ilustrativa de um dendrograma onde o topo é representado o conjunto de todas as letras e o fim de cada letra individual.

Existem dois tipos principais de métodos de *clustering* hierárquico, sendo eles [14] :

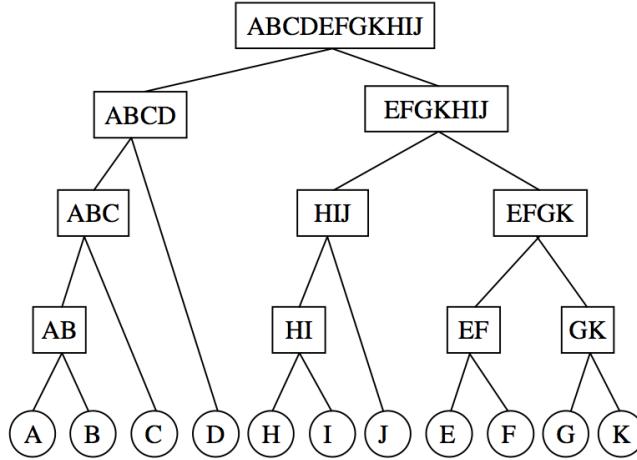


Figura 2.1: Dendrograma ilustrativo da divisão entre nós do conjunto completo ao objeto individual no fundo [1]

Clustering por aglomeração: O dendrograma é construído do nível mais baixo até ao mais alto, juntando sucessivamente e iterativamente os *clusters* com maiores semelhanças até existir um único *cluster* com todo o conjunto dos dados.

Clustering por divisão: O dendrograma é construído do nível mais alto até ao nível mais baixo, onde o processo tem início com um único *cluster* que possui todos os objetos, sendo dividido sucessivamente em *clusters* mais pequenos, até que estes sejam constituídos apenas por um único objeto.

Ao contrário do algoritmo *k-means*, que apenas calcula a distância entre os centroides de cada grupo ou *cluster*, no clustering hierárquico podem ser usados os vários métodos apresentados em seguida para determinar a distância entre dois *clusters* [14]:

Método Single-Link: Neste método, a distância entre dois *clusters* é determinada pela distância entre os dois objetos mais próximos (vizinhos mais próximo) pertencentes a *clusters* diferentes.

Método Complete-Link: Neste método, a distância entre dois *clusters* é determinada pela maior distância entre dois objetos (vizinhos mais distante).

Método Average-Link: Este método tenta manter um compromisso entre a sensibilidade a *outliers* do método *Complete-Link* e a sensibilidade do método *Single-Link* ao ruído existente nos dados. Para isso, é determinada a distância entre dois *clusters* através da distância média entre todos os pares de objetos nos dois *clusters*.

Método Ward: Este método, tenta minimizar a variância entre dois *clusters* unidos.

Assim conclui-se que o clustering hierárquico apresenta-se para determinados domínios, como bastante intuitivo para humanos, mas a interpretação dos resultados pode ser por vezes subjetiva.

Outra característica interessante é o facto de, ao contrário do *clustering* por partição, no *clustering* hierárquico não ser necessário especificar logo à partida o número de *clusters*.

2.1.3 Clustering Baseado em Densidade

Os métodos de *clustering* por partição ou hierárquico estão preparados para encontrar *clusters* que apresentam formas geométricas circulares, sendo ineficiente quando as formas destes grupos são por exemplo elípticas. Assim, para descobrir *clusters* com formas arbitrárias, podem ser usados métodos baseados na densidade dos objetos [7]. Um dos algoritmos mais conhecidos que utiliza este tipo de método é o DBSCAN (*Density-Based Spatial Clustering of Applications with Noise*) que é capaz de encontrar *clusters* através da análise da densidade e proximidade dos objetos pertencentes a um conjunto de dados. Para esta análise é necessário previamente atribuir um valor que definirá o raio da vizinhança considerada para cada objeto. Esse parâmetro é designado por ε e terá de ser necessariamente maior que 0. Assim, a ε -vizinhança de um objeto x é o espaço dentro de um raio com valor ε , centrado em x [7]. Já para determinar a densidade de uma vizinhança, é utilizado o parâmetro *MinPts* também previamente definido, que especifica o número mínimo de objetos vizinhos que um objeto necessita ter em seu redor, para ser considerado como objeto central. Os passos necessários para a implementação deste método são apresentados no Algoritmo 2

Algoritmo 2 DBSCAN

```

1: procedure DBSCAN(MinPts: limiar da vizinhança, D: conjunto de dados,  $\varepsilon$  : raio )
2:   Marcar todos os objetos como não selecionados;
3:   repeat
4:     escolher aleatoriamente um objeto  $p$  não selecionado;
5:     actualizar objeto  $p$  como selecionado;
6:     if o  $\varepsilon$ -vizinhança de  $p$  tem pelo menos MinPts objetos then
7:       criar um novo cluster  $C$  e adicionar objeto  $p$  ao cluster  $C$ ;
8:       Seja  $N$  o conjunto de objetos na  $\varepsilon$ -vizinhança de  $p$ ;
9:       for cada ponto  $p'$  em  $N$  do
10:        if  $p'$  não selecionado then
11:          Marcar  $p'$  como selecionado
12:          if a  $\varepsilon$ -vizinhança de  $p'$  tem pelo menos MinPts then
13:            adicionar ponto a  $N$ 
14:          end if
15:        end if
16:        if  $p'$  não pertence a nenhum cluster then
17:          adicionar  $p'$  a  $C$ 
18:        end if
19:      end for
20:      output  $C$ ;
21:      else marcar  $p$  como ruido
22:      end if
23:    until todos os objetos selecionados;
24: end procedure

```

Uma das grandes vantagens do *clustering* baseado em densidade é que neste não é necessário uma prévia definição do número de *clusters*, sendo apresentados os que encontrar consoante os dados que possui e os parâmetros definidos.

2.1.4 Clustering Baseado em Grelhas

Os métodos de *clustering* discutidos até agora, apresentam algoritmos que se adaptam a distribuição dos dados no espaço. Em alternativa, o *clustering* baseado em grelhas é orientado ao espaço, na medida em que divide o espaço em células independentemente da distribuição dos objetos de entrada. Este quantifica o espaço num número finito de células, que formam uma estrutura de grelhas sobre a qual são executadas as operações de *clustering*. Este método apresenta como principal vantagem o tempo baixo de processamento, que normalmente é independente da quantidade de dados, no entanto, este depende do número de células em cada uma das dimensões no espaço quantizado [7].

O algoritmo STING (*Statistical Information Grid*) [15] é um dos algoritmos utilizados para *clustering* baseado em grelhas. Este divide o espaço em células retangulares, correspondente a diferentes resoluções que forma uma estrutura hierárquica, sendo a base o nível 1, os filhos o nível 2, e assim sucessivamente. Cada célula pertencente a um nível superior é dividida para formar células de menor dimensão no nível inferior seguinte. Assim, sabe-se que o nível mais baixo apresenta uma maior resolução. Isto permite que os *clusters* sejam encontrados recorrendo a uma pesquisa de cima para baixo (*clustering* por divisão, como explicado na secção 2.1.2), passando por cada nível até atingir o mais baixo, retornando no fim as células mais relevantes para a consulta especificada. A informação estatística de cada célula é calculada e armazenada para o processamento de consultas futuras. Também é necessário ter em atenção que apenas é considerado para este algoritmo um espaço bidimensional. Um outro algoritmo com características semelhantes é o CLIQUE [16], que "identifica *clusters* densos em sub-espacos de máxima dimensão", isto é, são detetados todos os *clusters* em todos sub-espacos existentes e em que um ponto pode pertencer a vários *clusters* em sub-espacos diferentes.

2.1.5 Funções de Distância

As funções de distância ou similaridade têm um papel fulcral em todos os algoritmos de *Clustering*. Existem inúmeras funções de distância usadas para diferentes tipos de atributos (ou variáveis) [14]. Em seguida serão apresentadas diferentes funções distância para diferentes atributos: numéricos, binários e nominal. Também serão apresentadas funções distância utilizadas para as dimensões temporal e de conteúdo.

2.1.5.1 Atributos Numéricos

As funções de distância mais utilizadas para variáveis numéricas são a **Distância Euclidiana** e **Distância Manhattan**. É utilizado $dist(x_i, x_j)$ para representar a distância entre duas instâncias

de r dimensões. Ambas as funções referidas anteriormente são casos especiais da função mais geral chamada **Distância Minkowski** [14]:

$$dist(x_i, x_j) = (|x_{i1} - x_{j1}|^h + |x_{i2} - x_{j2}|^h + \dots + |x_{ir} - x_{jr}|^h)^{\frac{1}{h}}, \quad (2.4)$$

onde h é um inteiro positivo.

Se $h=2$, temos a **Distância Euclidiana**,

$$dist(x_i, x_j) = \sqrt{(x_{i1} - x_{j1})^2 + (x_{i2} - x_{j2})^2 + \dots + (x_{ir} - x_{jr})^2}. \quad (2.5)$$

Se $h=1$, temos a **Distância City-block (Manhattan)**,

$$dist(x_i, x_j) = |x_{i1} - x_{j1}| + |x_{i2} - x_{j2}| + \dots + |x_{ir} - x_{jr}|. \quad (2.6)$$

Também, não menos importantes, são outras funções distância apresentadas em seguida:

Distância Euclidiana Ponderada : A ponderação é atribuída através de pesos dados pela importância que cada atributo representa relativamente a outros atributos.

$$dist(x_i, x_j) = \sqrt{w_1(x_{i1} - x_{j1})^2 + w_2(x_{i2} - x_{j2})^2 + \dots + w_r(x_{ir} - x_{jr})^2}. \quad (2.7)$$

Distância Euclidiana Quadrática : Trata-se de uma alteração da função **Distância Euclidiana**, elevando a mesma ao quadrado, o que faz com que seja progressivamente atribuído peso maior a pontos dos dados que estejam mais afastados.

$$dist(x_i, x_j) = (x_{i1} - x_{j1})^2 + (x_{i2} - x_{j2})^2 + \dots + (x_{ir} - x_{jr})^2. \quad (2.8)$$

Distância Chebyshev : Utilizada para casos em que há necessidade de definir dois pontos dos dados como diferentes, caso sejam diferentes em qualquer dimensão.

$$dist(x_i, x_j) = \max(|x_{i1} - x_{j1}| + |x_{i2} - x_{j2}| + \dots + |x_{ir} - x_{jr}|). \quad (2.9)$$

2.1.5.2 Atributos Binários e Nominais

As funções apresentadas anteriormente apenas podem ser utilizadas com atributos do tipo numérico, assim serão necessárias funções de distância específicas para atributos do tipo binário e nominal.

Uma variável binária é aquela que apenas pode assumir dois estados ou valores, sendo normalmente representado pelo valor 0 e 1. Mas estes estados não apresentam um ordem definida. Por exemplo, no caso de uma lâmpada, esta pode assumir apenas dois estados, ligado ou desligado, ou o género de uma pessoa, masculino ou feminino. Estes exemplos apresentam dois valores diferentes mas que não possuem qualquer ordem. As funções distância existentes para atributos binários

são baseadas na proporção, sendo que a melhor maneira de representar é através de uma matriz confusão [14].

		Objeto j		
		x = 1	x = 0	
Objeto i	x = 1	a	b	$a+b+c+d = \text{número de variáveis}$
	x = 0	c	d	

Figura 2.2: Matriz confusão de dois objetos com atributos binários

Os atributos binários ainda podem ser divididos em dois tipos de atributos diferentes, os simétricos e os assimétricos, sendo em seguida apresentado as funções distância para ambos os casos [14].

Atributos simétricos: Um atributo é simétrico quando ambos os estados (0 ou 1) têm a mesma importância e o mesmo peso, tal como ocorre no exemplo dado anterior com o atributo gênero (masculino e feminino). Para este caso, a função distância mais utilizada é designada por *simple matching distance*, que corresponde à proporção de incompatibilidade ou desacordo(equação 2.10).

$$dist(x_i, x_j) = \frac{b + c}{a + b + c + d} \quad (2.10)$$

Atributos assimétricos: Um atributo é assimétrico se um dos estados apresenta maior importância ou valor do que o outro. Normalmente o estado mais valioso é o que ocorre com menor frequência. No nosso caso iremos considerar o estado 1 como o mais valioso. Assim, a função distância mais frequentemente utilizada para atributos assimétricos é a *Jaccard distance*:

$$dist(x_i, x_j) = \frac{b + c}{a + b + c} \quad (2.11)$$

No caso de atributos nominais com mais de dois estados ou valores, a função distância mais utilizada, é baseada na *simple matching distance*. Dados dois objetos i e j , r corresponde ao número total de atributos e q ao número de valores que são mutuamente correspondidos entre os objetos i e j :

$$dist(x_i, x_j) = \frac{r + q}{r} \quad (2.12)$$

2.1.5.3 Dimensão Temporal

O tempo é representado apenas por uma dimensão, sendo que para calcular a distância, por exemplo, entre dois tweets t_i e t_j , apenas é necessário calcular a diferença dos tempos entre os mesmos. Supondo que os valores dos tempos são respetivamente Δ_i e Δ_j , o intervalo de tempo pode ser definido pela seguinte equação:

$$dist^T(t_i, t_j) = |\Delta_i - \Delta_j| \quad (2.13)$$

2.1.5.4 Dimensão Espacial

Ao contrário da dimensão temporal, a dimensão espacial apresenta mais do que uma dimensão, latitude e longitude. Estas apresentam-se sobre a forma numérica, sendo possível o cálculo da distância entre dois objetos através de funções de distância para atributos numéricos como referido anteriormente (2.1.5.1). Assim, para o cálculo entre pontos distribuídos num espaço poder-se-á recorrer à função Minkowski (equação 2.4), à função Euclidiana (equação 2.5), à função Manhattan (equação 2.6) ou mesmo à função de Chebychev (equação 2.9), sendo que no caso mais específico de uma distribuição espacial geográfica, em que os pontos possuem latitude e longitude, é considerada mais apropriada a utilização da função distância Haversine pois esta toma em consideração a forma esférica da Terra [17]. Assim, obtendo um par de objetos x_i e x_j distanciados geograficamente, são consideradas a latitude ϕ_{x_i} e ϕ_{x_j} e a longitude λ_{x_i} e λ_{x_j} para determinar a distância entre os objetos através da equação 2.14

$$dist^{Sp}(x_i, x_j) = 2R \sin^{-1} \left(\left[\sin^2\left(\frac{\phi_{x_i} - \phi_{x_j}}{2}\right) + \cos \phi_{x_i} \cos \phi_{x_j} \sin^2\left(\frac{\lambda_{x_i} - \lambda_{x_j}}{2}\right) \right]^{0.5} \right) \quad (2.14)$$

onde R representa o raio da Terra e que determina as unidades do resultado retornado pela função, sendo comum a utilização das unidades no sistema internacional (SI), o metro, podendo também ser representado em quilómetros devido ao fator de escala.

2.2 TweeProfiles

Esta dissertação pretende dar continuidade e um trabalho designado por TweeProfiles [10]. O TweetProfiles é uma ferramenta de análise de dados recolhidos no Twitter e visualização *clusters* em várias dimensões, tais como, espacial, temporal, de conteúdo (o texto dos tweets) e social. Esta secção faz um breve introdução e descrição sobre esta ferramenta.

2.2.1 Descrição e objetivos

O TweeProfiles aborda o problema de identificar perfis de tweets envolvendo múltiplos tipos de informação: espacial, temporal, social e de conteúdo. A informação espacial, trata-se da informação de localização das mensagens de texto partilhadas no Twitter, a temporal é relativa à data

de publicação do tweet, a social às ligações entre os utilizadores, e por fim o conteúdo que neste caso é relativo ao texto contido em cada tweet.

Os objetivos do TweeProfiles foi o desenvolvimento de uma metodologia de Data Mining que identificasse perfis de tweets que combinem de forma flexíveis as várias dimensões consideradas, a criação de uma ferramenta de visualização para representar os resultados obtidos e a sua aplicação a um caso de estudo na twittosfera portuguesa.

A ferramenta de visualização está desenhada para uma utilização dinâmica e intuitiva, direcionada para a representação dos perfis de uma forma comprehensível e interativa. Esta apresenta vários widgets capazes de representar os padrões obtidos. O caso de estudo que o TweeProfiles aborda dados georeferenciados do Socialbus (antes designado como TwitterEcho). No entanto, esta ferramenta é adequada para tratar quaisquer mensagens georeferenciadas provenientes do Twitter.

2.2.2 Resultados ilustrativos

Através do TweeProfiles podemos visualizar *clusters* nas suas diversas dimensões. Esta ferramenta apresenta três secções distintas para visualização e inclui ainda controlos para navegação e seleção de determinados parâmetros. Uma das secções é constituída por um mapa onde é possível visualizar geograficamente os *clusters*, como podemos ver na Figura 2.3, em que são representados por círculo. Outra das secções é um gráfico que apresenta a distribuição temporal dos *clusters*, onde é possível visualizar a data de início de fim de um determinado *cluster*, como representado na Figura 2.4. Por fim, é nos apresentada uma secção onde é possível visualizar informação mais



Figura 2.3: Exemplo ilustrativo da distribuição espacial dos *clusters* calculada apenas com a consideração da dimensão espacial. Retirada de [10]

específica sobre os *clusters*, onde é incluído a informação relativa às dimensões de conteúdo e social com a representação das palavras mais utilizadas e as ligações entre utilizadores respetivamente nesse *clusters*. No caso dos controlos, estes permitem selecionar um dos três intervalos

de tempo existentes e os pesos para cada uma das dimensões, que podem assumir valores entre 0% e 100% com incrementos de 25%. Cada dimensão pode assumir como peso assim um dos seguintes valores percentuais: {0, 25, 50, 75, 100}, sendo que a soma dos pesos das diferentes dimensões deve ser igual a 1.

Em seguida são apresentados alguns resultados ilustrativos obtidos através da ferramenta TweeProfiles, e que demonstram o seu funcionamento global. Este resultados serão baseados apenas nos apresentados por Tiago Cunha [10], sendo ilustrados resultados de *clusters* em Portugal para diferentes combinações possíveis.



Figura 2.4: Exemplo ilustrativo da distribuição temporal de um *cluster*. Retirada de [10]

Na Figura 2.5 são apresentadas quatro combinações entre a dimensão do conteúdo e as restantes, exceto no primeiro caso 2.5a onde podemos ver um *cluster* em Portugal apenas tendo em consideração a dimensão do conteúdo. Já no caso da Figura 2.5b o peso é distribuído da mesma forma entre o conteúdo e a dimensão espacial, resultando em dois *clusters*. Na Figura 2.5c, tal como no caso anterior, é distribuído o peso de igual forma entre o conteúdo, e neste caso a dimensão temporal. O resultado obtidos no caso de Portugal, assemelha-se ao do primeiro caso da Figura 2.5a. Por fim, temos o caso da Figura 2.5d onde o peso é de 50% para o conteúdo e 50% para a dimensão social, tendo resultado para o caso de Portugal, dois *cluster*.

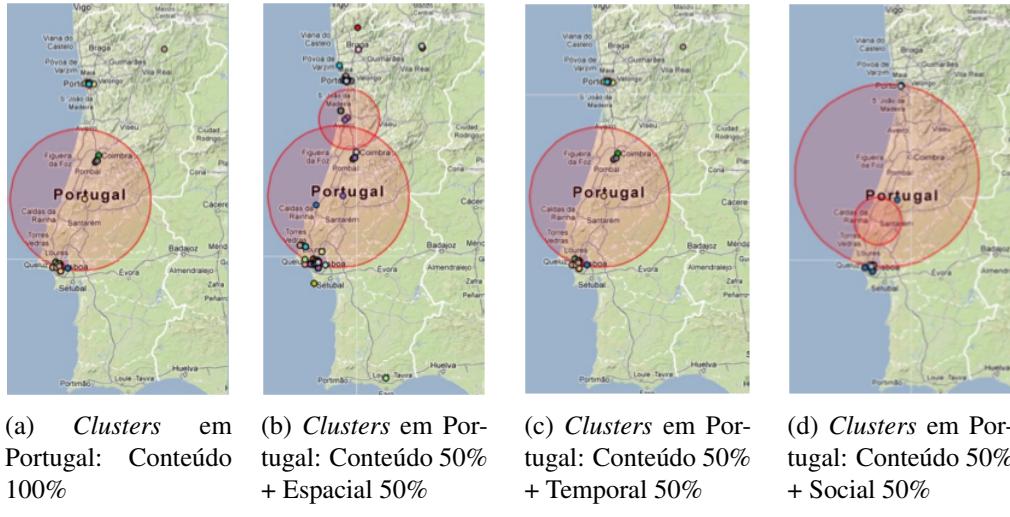


Figura 2.5: Exemplos de resultados de *clusters* com pesos diferentes para as diferentes dimensões. Retiradas de [10]

Outra das combinações apresentadas, foi a distribuição igual por todas as dimensões. Este caso é apresentado na Figura 2.6 onde todas as dimensões possuem o peso de 25% cada. Neste caso foram obtidos três *cluster* em Portugal.



Figura 2.6: *Clusters* em Portugal: Conteúdo 25% + Espacial 25% + Temporal 25% + Social 25%. Retirada de [10]

Por fim, na Figura 2.7 é possível visualizar a secção que apresenta ao utilizador informação sobre um *cluster* selecionado, o conteúdo desse *cluster*, como por exemplo, as palavras mais relevantes, e o grafo com as ligações sociais. É possível ainda visualizar com mais detalhe de um determinado tweet, como o nome do utilizador, informação temporal e espacial.

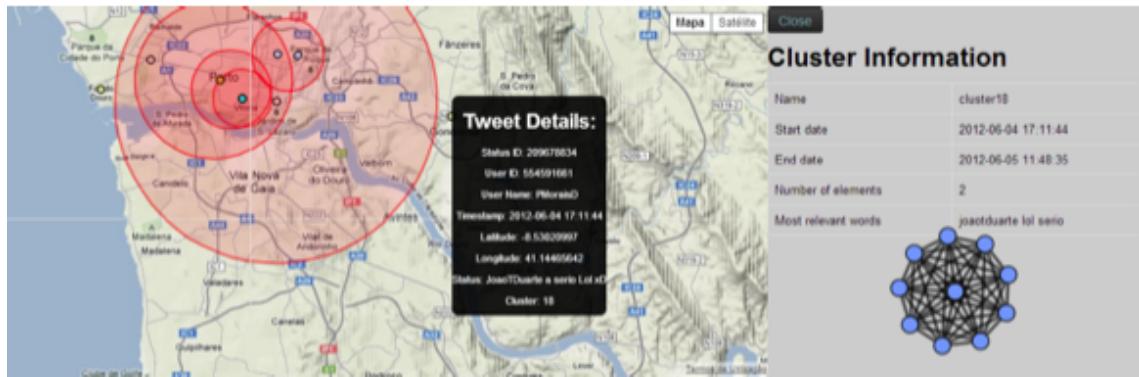


Figura 2.7: Visualização de informação mais detalhada de um *cluster* incluindo o seu grafo social, e da informação relativa a um determinado tweet. Retirada de [10]

2.2.3 Prós e contras

Como vimos na Secção 2.2.2, a ferramenta TweeProfiles permite uma grande variedade de combinações que produzem resultados que podem variar no número e posição geográfica dos *clusters* consoante a influência e peso e das dimensões em consideração.

Este projeto de dissertação não tem como objetivo melhorar O TweeProfiles, mas sim apresentar uma abordagem alternativa de visualização de análise de dados em redes sociais utilizando conteúdo diferente. Por este motivo, nesta secção somente abordamos os prós e contras relativamente à abordagem das dimensões utilizadas.

Uma das vantagens da utilização do texto como dimensão de conteúdo, é o facto de no Twitter este se apresentar como a maior fonte de informação, pois os tweets na sua maioria apresentam texto. Mesmo adicionando a dimensão espacial, que implica a existência de uma referência geográfica contida na informação dos tweets, o número de tweets disponíveis continua a ser relativamente razoável para uma análise dos dados, tal como podemos confirmar em [10]. Apesar disso, é possível verificar qu nos resultados apresentados existe muito conteúdo em forma de texto que apresenta informação muitas vezes com pouca relevância, como é o caso da partilha da sua localização através de outros serviços onde é utilizado um texto pré-definido pelo serviço utilizado, indicando apenas o local onde se encontra o utilizador, como podemos ver pela Figura 2.8.

Cluster	Tweet ID	User Name	Timestamp	Latitude	Longitude	Status
3	208529373	Madptron	01/06/2012 13:04	-9.20568466	38.66111817	I'm at Faculdade de Ciências e Tecnologia - Universidade Nova de Lisboa (Almada, Portugal) http://t.co/mDq2VH8y
3	209562126	Madptron	04/06/2012 09:27	-9.20568466	38.66111817	I'm at Faculdade de Ciências e Tecnologia - Universidade Nova de Lisboa (Almada, Portugal) http://t.co/noHVgTY
3	209982566	pedro_jose	05/06/2012 13:18	-9.20568466	38.66111817	I'm at Faculdade de Ciências e Tecnologia - Universidade Nova de Lisboa (Almada, Portugal) http://t.co/y8T8qQVI
3	210022881	diogoreis32	05/06/2012 15:58	-9.20568466	38.66111817	I'm at Faculdade de Ciências e Tecnologia - Universidade Nova de Lisboa (Almada, Portugal) http://t.co/SZQVE4fw

Figura 2.8: Tweets pertencentes a um *cluster* para a seguinte distribuição de pesos: Conteúdo 25% + Espacial 25% + Temporal 25% + Social 25%. Retirada de [10]

O TweeProfiles revela ainda que a dimensão espacial, quando tida em consideração, tem um influência visivelmente interessante nos resultados quando apresentados no mapa. Já a dimensão temporal, não revelou apresentar uma grande influência como vimos no caso apresentado na Figura 2.5c, mas esta permite um controlo e visualização de informação interessante através do gráfico de tempo como representado na Figura 2.4. No caso da dimensão social, verifica-se que apresenta alguma influência na divisão dos *clusters* mas a visualização da sua informação através do grafo, apresenta-se pobre e de difícil compreensão para o utilizador.

Assim, conclui-se que a utilização das dimensões espaço-temporal fazem todo o sentido no desenvolvimento desta dissertação, sendo necessária a sua inclusão para a extensão do TweeProfiles. Já no caso do conteúdo de texto, este não será tido em conta de forma a ser analisado se existe uma mais valia em realizar uma análise em tweets através exclusivamente do conteúdo visual partilhado pelos utilizadores com as possíveis combinações com as dimensões temporal e espacial. A dimensão social, também não será incluída por esta não adicionar ainda não apresentar a informação de forma clara para um utilizador, sendo necessário um estudo mais aprofundado sobre esta temática, que não faz parte dos objetivos desta dissertação.

2.3 Representação de Informação Visual

Na secção 2.1 foi apresentado o conceito e características da tarefa de *clustering* de uma forma geral. Nesta secção serão expostas formas de representar imagens como dados.

Como o objetivo desta dissertação passa por a realização da tarefa de *clustering*, utilizando como dados as imagens partilhadas no serviço Twitter, é necessário utilizar formas eficientes para descrever cada imagem de uma forma compacta e de forma a que seja descrito o conteúdo geral das imagens. É importante salientar que a tarefa de *clustering* em imagem é muitas vezes associada à técnica de segmentação de imagem [18], em que se pretende distinguir objetos individuais, não sendo este o objetivo deste projeto de dissertação. Pretende-se sim que para tarefa de *clustering* o objeto representativo de uma imagem descreva esta pelo seu conjunto, tal como referido anteriormente.

2.3.1 Representação Matricial

Uma imagem pode ser vista como um objeto (ou instância), sendo computacionalmente representada como uma matriz (um vetor bi-dimensional) de pixels. A matriz de pixels descreve assim a imagem como $N \times M$ m -bit pixels, onde N corresponde ao número de pontos ao longo do eixo horizontal, M o número de pontos ao longo do eixo vertical e m o número de bits por pixel que controla os níveis de brilho. Com m bits temos uma gama de valores para o brilho de 2^m , que varia entre 0 e $2^m - 1$. Assim se o valor de m for 8, os valores de brilho de cada pixel de uma imagem podem variar entre 0 e 255, que normalmente correspondem ao preto e branco respetivamente, sendo que os valores intermédios correspondem ao tons de cinza [19].

No caso de imagens a cores, o princípio é idêntico, no entanto ao invés de se usar apenas um plano, as imagens a cores são representadas por 3 componentes de intensidade, designado por modelo *RGB*, a que corresponde respetivamente às cores vermelho (Red), verde (Green) e azul (Blue). Para além deste esquema de cores, também existe outros como o *CMYK* composto pelas componentes de cor, azul turquesa, magenta, amarelo e preto. Usando qualquer esquema de cores, existem 2 métodos principais para representar a cor do pixel. No primeiro método é utilizado um valor inteiro para cada pixel, sendo esse valor como um índice para uma tabela, também conhecida como paleta da imagem, com a correspondência à intensidade de cada componente de cor. Este método tem como vantagem o facto de ser eficiente na utilização da memória, pois apenas é guardado um plano da imagem (os índices) e a paleta (tabela). Por outro lado, tem como desvantagem o facto de normalmente ser usado um conjunto reduzido de cores o que provoca uma redução da qualidade da imagem. Já o segundo método consiste na utilização de vários planos da imagem para armazenar a componente de cor de cada pixel. Este representa a imagem com mais precisão pois considera muito mais cores. O formato mais usual é 8 bits para cada uma das 3 componentes, no caso do *RGB*. Assim, são utilizados 24 bits para representar a cor de cada pixel, o que permite que uma imagem possa conter mais de 16 milhões de cores simultaneamente. Como

era de esperar, isto envolve um custo grande na utilização de memória, que mesmo apresentando-se como uma desvantagem, com a constante redução do custo das memórias esta passou ser uma boa alternativa à apresentada anteriormente [19].

Em suma, a representação matricial é uma forma fácil de representar uma imagem, mas ao contrário dos objetivos desta dissertação, não consegue fazê-lo de uma forma compacta, sendo necessária a existência de grandes quantidades de memória devido à sua dimensionalidade. Assim, esta não se apresenta como uma boa solução para o a extração da informação das imagens.

2.3.2 Histogramas

Outras das formas de representar a informação de uma imagem é através de um histograma. Um histograma de uma imagem apresenta a frequência de ocorrência de níveis individuais de brilho, representado através um gráfico que mostra o número de pixels da imagens com um determinado nível de brilho. No caso de pixels representados por 8-bit, o brilho vai variar de 0 (preto) até 255 (branco) [19]. Também pode ser apresentado informação de cor sobre uma imagem através de um histograma, sendo para isso necessário apresentar 3 histogramas diferenciados, um para cada componente de cor, no caso do esquema RGB. A figura 2.9 apresenta um exemplo de um histograma de uma imagem com tons cinza, onde são representados o número de pixeis para cada nível diferente de cinzento.

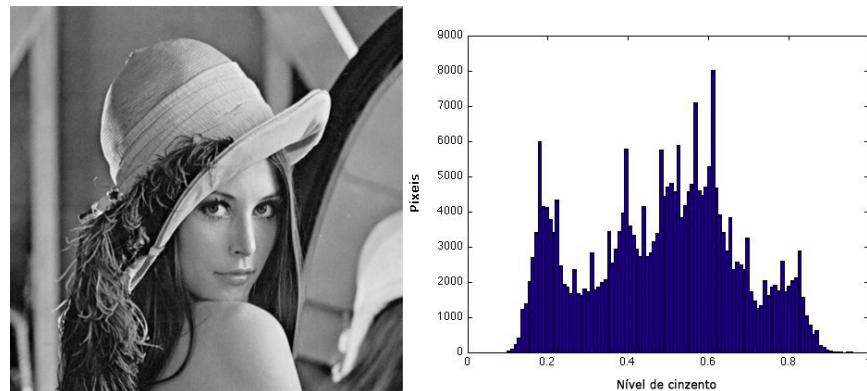


Figura 2.9: Imagem em tons cinza e respetivo histograma

2.3.3 Descritores de Cor

A cor apresenta-se como um importante atributo da imagem para o olho humano e processamento por computador. Nesta secção são apresentados vários descritores de cor, utilizados para extração de informação e reconhecimento de similaridade em imagens. Por exemplo, o histograma de cores, referido na secção 2.3.2, é um dos descritores de cor mais utilizados para caracterizar a distribuição da cor de uma imagem, mas apresenta uma baixa eficiência. Assim, em seguida é apresentado descritores de cor considerados pelo MPEG-7 [20, 21, 22, 23], que apresentam eficiência superior aos histogramas de cor.

2.3.3.1 Espaços de Cor

Nesta secção é apresentado os vários descritores de espaços de cor especificado no MPEG-7 [23]. Existe uma vasta seleção de espaços de cores, tais como, RGB, YCbCr, HSV, HMMS, Monocromático e Matriz linear de transformação com referência a RGB. Estes são usados por outro descritor de cor, mais especificamente, o descritor de dominância de cor que será falado posteriormente. É utilizado também, um sinalizador para indicar a referência a uma cor primária e de mapeamento de um valor de referência do branco padrão.

Em espaços de cor, as componentes de cor são definidas como entidades de valor contínuo, sendo que podem ser representadas por valores discretos através de uma quantização uniforme, em que é especificado um número de níveis de quantização para cada componente de cor no espaço de cor. A única exceção é o espaço de cor HMMD.

O espaço de cor RGB é um dos modelos referidos mais utilizados, que apresenta três componentes distintas, vermelho, verde e azul, tal como foi referido no secção 2.3.1. Neste modelo é utilizado a combinação das 3 cores primárias para representar as diferentes cores. O modelo YCbCr provém do padrão MPEG-1/2/4 [23] e é definido pela transformação linear do espaço de cor RGB como demonstrado na equação 2.15:

$$\begin{aligned} Y &= 0.299 \times R + 0.587 \times G + 0.114 \times B \\ Cb &= -0.169 \times R - 0.331 \times G + 0.500 \times B \\ Cr &= 0.500 \times R - 0.419 \times G - 0.081 \times B \end{aligned} \quad (2.15)$$

Para o espaço de cor Monocromático, é usado apenas a componente Y do modelo YCbCr.

O espaço de cor HSV apresenta uma especificação mais complexa, tendo sido desenvolvido para fornecer uma representação mais intuitiva e para se aproximar mais do sistema visual humano. A transformação do modelo RGB para o HSV não é linear, mas é reversível [20]. Uma das componentes é a matiz (*H - Hue*), que representa a componente de cor espectral dominante na sua forma mais pura, como o verde, amarelo, azul e vermelho. Ao ser adicionado branco à cor, esta sofre uma alteração, sendo que, adicionando mais branco, menos saturada se torna a cor. A saturação (*S - Saturation*) é precisamente outra das componentes deste modelo. Por fim, o valor (*V - Value*) corresponde ao brilho de cor.

O espaço de cor HMMD (*Hue-Max-Min-Diff*) [20, 23] é mais recente, que é caracterizado pela componente matiz, tal como o modelo HSV, pelo *max* e *min*, que são respetivamente o máximo e mínimo entre os valores R, G e B. Para descrever este modelo, também é utilizado a componente *Diff*, que corresponde à diferença entre o *max* e *min*. Para representar este espaço de cor, apenas é necessário três dos quatro componentes referidos anteriormente, como por exemplo, *Hue*, *Max*, *Min* ou *Hue*, *Diff*, *Sum*, onde *Sum* pode ser definida pela equação 2.16.

$$Sum = \frac{Max + Min}{2} \quad (2.16)$$

2.3.3.2 Cor Dominante

O descriptor de cor dominante fornece uma representação compacta das cores de uma imagem ou da região da imagem. Este apresenta a distribuição das cores mais representativas na imagem. Ao contrário do descriptor de cor por histograma, na especificação do descriptor de cor dominante, as cores mais representativas são calculadas a partir de cada imagem, em vez de ser fixado no espaço de cor, permitindo assim, uma representação das cores mais exata e compacta, presentes numa região de interesse.

O descriptor de cor dominante pode ser definido como,

$$F = c_i, p_i, v_i, s, (i = 1, 2, \dots, N)$$

onde N é o número de cores dominantes. Cada valor c_i da cor dominante é um vetor de valores das componentes do espaço de cor correspondente (por exemplo, um vetor de 3 dimensões no espaço de cor RGB). O valor p_i é a fração de pixels na imagem ou região da imagem (normalizado para um valor entre 0 e 1) que corresponde à cor c_i , sendo $\sum_i p_i = 1$. O opcional v_i descreve a variação dos valores de cor dos pixels em um *cluster* em torno da cor representativa correspondente. Por fim, a coerência espacial s é um único número representa a homogeneidade espacial global das cores predominantes na imagem [23].

2.3.3.3 Cor Escalável

O descriptor de cor escalável, pode ser interpretado como um esquema de codificação base que recorre à transformada de Haar, que é aplicada aos valores do histograma de cor no espaço de cor HSV (referido na secção 2.3.3.1). De uma forma mais específica, o descriptor de cor escalável, extraí, normaliza e mapeia de forma não linear os valores do histograma, numa representação inteira a 4-bit, dando assim mais relevância a valores mais pequenos. A transformada de Haar, é assim aplicada aos valores inteiros a 4-bit através das barras do histograma.

A extração do descriptor é realizada com computação de um histograma de cor com 256 níveis no espaço de cor de HSV com a componente matiz (H) quantificada a 16 níveis, e a saturação (S) e o valor (V) quantificado cada um para 4 níveis [21].

A aplicação típica do descriptor é na a busca de similaridade numa base de dados com conteúdo multimédia e pesquisa em enormes base de dados.

2.3.3.4 Estrutura de Cor

Este descriptor é uma generalização do histograma de cores, que apresenta algumas características espaciais da distribuição de cores numa imagem. Este tem a particularidade de, para além de apresentar o conteúdo da cor de forma semelhante a um histograma de cor, também apresentar informações sobre a estrutura de uma imagem, sendo esta a característica diferenciadora deste descriptor de cor. Em vez de considerar cada pixel separadamente, o descriptor recorre a uma estrutura

de 8x8 pexels que desliza sobre a imagem. Ao contrário do histograma de cor, este descritor consegue distinguir duas imagens em que uma determinada cor está presente em quantidades iguais, mas que apresenta uma estrutura num dos grupos de pixels 8x8 com uma cor diferente nas duas imagens. Os valores de cores são representadas no espaço de cor HMMD com cone duplo, sendo o espaço quantificado de maneira não uniforme em 32, 64, 128 ou 256 níveis. Cada valor de amplitude de um nível é representado por um código de 8 bits. Este descritor apresenta um bom desempenho na tarefa de recuperação de imagens baseado na similaridade [24].

2.3.4 Descritores de Textura

A textura das imagens é uma característica visual importante, que tem muitas aplicações na recuperação, navegação e indexação de imagens. Existem três descritores de textura, referenciados na norma MPEG-7 [2]. Em seguida é apresentada uma pequena descrição dos mesmos.

2.3.4.1 Descritor de Textura Homogénea

O descritor de textura homogénea (HTD - *Homogeneous Texture Descriptor*) descreve a distribuição estatística da textura de uma imagem. Existem neste descritor 62 interfaces de recurso, sendo 2 no domínio espacial e 60 no domínio das frequências. No domínio espacial, é extraído a média e o desvio padrão de uma imagem. No domínio das frequências, o espaço é dividido em 30 canais, sendo calculado o valor energético e o valor do desvio de energia da resposta do filtro Gabor em cada canal [2, 25]. Este desenho baseia-se no facto de a resposta do córtex visual possuir banda limitada e do facto do cérebro decompor o espectro em bandas na frequência espacial [2]. O descritor de textura homogénea é essencialmente utilizado em aplicações de recuperação de imagens por similaridade.

2.3.4.2 Descritor de Histograma de Borda

O descritor de histograma de borda (EHD - *Edge Histogram Descriptor*) apresenta-se sobre a forma de um histograma de 80 níveis, que representa a distribuição de borda local de uma imagem. Este descreve as bordas em cada sub-imagem. Estas sub-imagens são obtidas através da divisão da imagem numa grelha 4x4 como pode ser visto na Figura 2.10. Existem 5 tipos de classificação diferentes das bordas de cada sub-imagem, sendo elas: vertical, horizontal, 45-graus, 135-graus e não direcional [2].

Este descritor é utilizado na recuperação de imagens, como por exemplo, imagens naturais ou de esboço, devido à sua textura homogénea. É também suportado por este descritor, a pesquisa baseada em blocos de imagem.

2.3.4.3 Descritor de Navegação Percetual

O descritor de navegação perceptual (HTD - *Perceptual Browsing Descriptor*) foi projetado para navegação em base de dados, mas principalmente para quando essa navegação necessita de

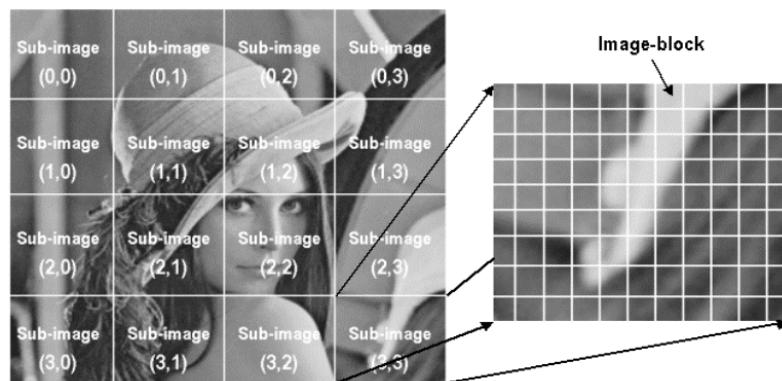


Figura 2.10: Sub-imagem e bloco de imagem. *Retirada de [2]*

recursos com sentido perceptual [2]. Este descritor é bastante compacto, que requer apenas 12 bits (máximo) para caracterizar regularidade (2 bits), direcionamento (3 bits x 2) e grosseirismo (2 bits x 2) da textura de uma imagem. A regularidade de uma textura pode apresentar valores numa escala entre 0 e 3, em que 0 indica uma textura irregular ou aleatória e 3 indica um padrão com direção e grosseirismo bem definidos. O direcionamento de uma textura é quantizada em 6 valores, variando de 0 a 150 em degraus de 30. Por fim, o grosseirismo de uma textura está relacionado com a escala e resolução de uma imagem. É quantizado em 4 níveis de 0 a 3, sendo 0 para um grão fino e 3 para uma textura grosseira. Estes valores têm uma relação com a divisão do espaço de frequência usado no cálculo do descritor de textura homogénea (HTD) [20]

2.3.5 Descritores de Forma

O descritores de forma são dos descritores mais poderosos no reconhecimento de objetos. Isto deve-se ao facto de os seres humanos serem exímios no reconhecimento de objetos característicos exclusivamente através da suas formas, provando que a forma muitas vezes possui informação semântica [3].

2.3.5.1 Descritor de Forma Baseado em Região

O descritor de forma baseado em região (RSD - *Region-based Shape Descriptor*) [3] apresenta a distribuição de um pixel dentro de uma região de um objeto em 2 dimensões. Este permite a descrição de objetos simples, com ou sem buracos (figura 2.11), mas também permite a descrição de objetos mais complexos, que contém múltiplas regiões sem ligação.

As principais características deste descritor são [3]:

- Fornece uma forma compacta e eficiente de descrever várias regiões disjuntas;
- Quando, no processo de segmentação de um objeto, ocorrem sub-regiões sem ligação, o objeto ainda pode ser recuperado, desde que a informação de quais as regiões que foram divididas seja mantida e usada na extração do descritor;



Figura 2.11: Exemplo de formas de objetos que podem ser descritas eficazmente pelo descriptor baseado em região. *Retirada de [3]*.

- Apresenta uma boa robustez à segmentação de ruído.

2.3.5.2 Descriptor de Forma Baseado no Contorno

O descriptor de forma baseado no contorno (CSD - *Contour-based Shape Descriptor*) [3] fundamenta-se na representação da curvatura espaço-escala (CSS - *Curvature Scale-Space*) do contorno. O contorno é uma propriedade importante na identificação de objetos semanticamente semelhantes. É também bastante eficiente em aplicações onde a forma de objetos são muito variáveis, ou quando por exemplo, existem deformações de perspetiva. Esta apresenta um boa eficiência mesmo perante a existência de ruído nos contornos.

As principais características deste descriptor são [3]:

- Consegue distinguir objetos que apresentem formas semelhantes mas que a forma do contorno apresenta propriedades bem diferenciadoras;
- Tem a capacidade de encontrar formas que são semanticamente similar para os seres humanos, mesmo quando existe uma significativa variabilidade intra-classe;
- É eficiente mesmo em casos de deformações não rígidas;
- É eficiente mesmo em casos de distorções do contorno devido a variações de perspetiva, sendo uma situação muito comum em imagens e vídeo.

2.3.6 Descritores Locais

Um descriptor local permite a localização das estruturas locais de uma imagem de forma repetitiva. Estas são codificadas de modo a que sejam invariantes a transformações das imagens, tais como a translação, rotação, mudanças de escala ou deformações. Assim, estes descritores podem ser utilizados para representar uma imagem e podem ser utilizados para diversos fins, tais como, reconhecimento de objetos, reconhecimento de cenas, perseguição de movimento, correspondência entre imagens ou mesmo obtenção de estruturas 3D de múltiplas imagens. Para a extração das características deste descriptor é necessário utilizar um processo com as seguintes etapas [26]:

- Encontrar um conjunto de pontos chave;

- Definir uma região em torno de cada ponto-chave numa escala invariante;
- Extrair e caracterizar o conteúdo da região;
- Calcular o descritor da região normalizada;
- Combinar os descritores locais.

Em seguida são apresentados duas das técnicas mais representativas dos descritores locais, o SIFT e o SURF.

2.3.6.1 SIFT

O descritor SIFT (*Scale-Invariant Feature Transform*) [27, 4] é um descritor local que transforma uma imagem numa grande coleção de vetores de características locais invariantes a translação, rotação, mudanças de escala, e parcialmente invariante a mudanças de iluminação. Pode assim ser utilizado para detetar correspondência entre imagens com diferentes visões de objetos ou cenas. Este descritor apresenta como característica interessante o facto de compartilhar uma série de propriedades em comum com as respostas dos neurónios do lobo temporal na visão dos primatas. Os pontos-chave SIFT derivados de uma imagem são usados na indexação numa abordagem de vizinho mais próximo, para encontrar objetos candidatos.

Como foi indicado anteriormente, são necessários levar a cargo uma lista de etapas bem definidas para obtenção de descritores locais, sendo que as etapas para o descritor SIFT as seguintes:

1. **Deteção de extremos:** A deteção de extremos (máximos e mínimos) é conseguida através da procura realizada em várias escalas e localizações, de modo a serem extraídos pontos de interesse invariáveis à escala e rotação, através da diferença de filtros gaussianos. Estes pontos de interesse ou pontos chave correspondem a estes extremos para várias escalas. Um filtro gaussiano passa baixo é dado pela convolução entre uma imagem I e a função G :

$$L(x, y, \sigma) = G(x, y, \sigma) * I(x, y) \quad (2.17)$$

onde ,

$$G(x, y, \sigma) = \frac{1}{2\pi\sigma^2} e^{-\frac{(x^2+y^2)}{2\sigma^2}} \quad (2.18)$$

em que o o filtro varia à escala através do parâmetro σ

A função *DoG* ("Difference of Gaussian") é dada pela diferença entre as imagens filtradas em escalas próximas separadas por uma constante k e pode ser definida como:

$$DoG(x, y, \sigma) = G(x, y, k\sigma) * G(x, y, \sigma) \quad (2.19)$$

Assim, a convolução de uma imagem I com o filtro DoG é dado por:

$$D(x, y, \sigma) = (G(x, y, k\sigma) - G(x, y, \sigma)) * I(x, y) = L(x, y, k\sigma) - L(x, y, \sigma) \quad (2.20)$$

que corresponde à diferença entre as imagens filtradas pelo filtro gaussiano em escalas σ e $k\sigma$. Este filtro provoca uma perda de nitidez nas imagens (efeito desfoque) e torna-se assim capaz de detetar variações de intensidade nas imagens, como por exemplo, nos contornos. A Figura 2.12 mostra como é realizado o processo de obtenção das Diferenças Gaussianas e consequente formação das oitavas.

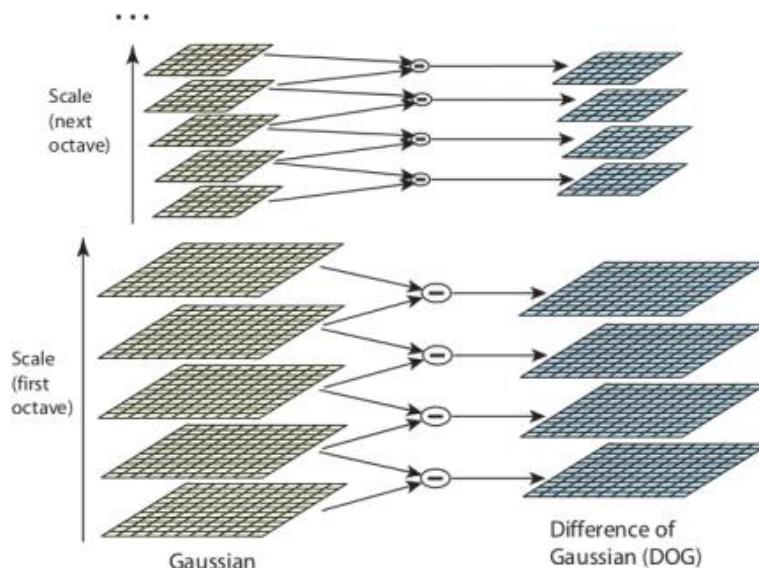


Figura 2.12: Construção das Diferenças Gaussianas e formação de oitavas *Retirada de [4]*.

Segundo Lowe [4] é necessário atingir a escala 2σ para ser possível a construção de um descriptor local invariável à escala, logo

$$k = 2^{(1/s)}$$

onde s é o número de intervalos entre imagens obtidas por DoG e $D(x, y, \sigma)$ corresponde à primeira imagem e $D(x, y, 2\sigma)$ à última de todo o conjunto de imagens geradas. Também deverão ser assim obtidas $s + 3$ imagens na pilha de imagens filtradas para cada oitava. Cada oitava contém as imagens de Diferença Gaussiana, sendo as que ficam entre as escalas superiores e inferiores designadas de intervalo.

Por fim é realizado o processo de deteção de extremos, onde um pixel é comparado com os seus oito vizinhos na imagem atual e com os nove pixels vizinhos das imagens de escalas adjacentes, numa região de 3x3. A figura 2.13 ilustra este processo.

O próximo processo passa pela localização dos ponto chave.

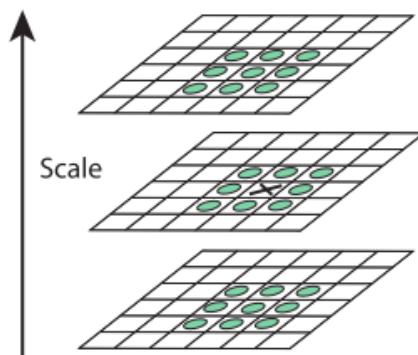


Figura 2.13: Ilustração do processo de deteção de máximos e mínimos das imagens de Diferença Gaussiana. O pixel candidato está marcado com X e os vizinhos com um círculo. *Retirada de [4]*

2. **Localização de pontos chave:** quando detetado um extremo, esse ponto é considerado um candidato a ponto chave ou ponto de interesse. Este ponto foi encontrado através da comparação de um pixel com os seus vizinhos como referido anteriormente. sendo necessário realizar um cálculo de ajuste detalhado da localização e escala gaussiana de cada um destes pontos. É utilizada então a série de Taylor para obter uma localização mais exata dos extremos, sendo rejeitado caso a intensidade de um extremo seja inferior a um limiar previamente definido.

Como DoG tem uma boa resposta a arestas e como estas fazem com que os pontos sejam instáveis com ruído, estes necessitam de ser removidos. É assim através da utilização de uma matriz Hessiana 2x2 possível calcular as curvaturas principais. Caso o rácio seja superior a um limiar previamente definido, o ponto é considerado uma aresta e assim o ponto chave é descartado.

Após a remoção de todos os pontos considerados não sendo de interesse, fica-se com todos os pontos chave, sendo necessário a atribuição das suas orientações.

3. **Atribuição da orientação dos descritores:** este processo tem como principal finalidade possibilitar a representação de um descritor em relação a sua orientação, permitindo assim que este seja invariante a rotações. Para realizar esta tarefa é utilizada a escala Gaussiana σ para a escolha da imagem filtrada L com a escala mais próxima e com a oitava referente ao ponto avaliado, tornando assim invariante também à escala.

Assim são calculados os gradientes para cada imagem $L(x, y, \sigma)$ de intervalo, referentes às escalas e oitavas utilizadas.

A magnitude e orientação são calculados da seguinte forma:

$$m(x, y) = \sqrt{\left((L(x+1, y) - L(x-1, y))^2 + (L(x, y+1) - L(x, y-1))^2 \right)} \quad (2.21)$$

Figura 2.14: Imagem

$$\theta(x,y) = \tan^{-1} \left(\frac{(L(x,y+1) - L(x,y-1))}{(L(x+1,y) - L(x-1,y))} \right) \quad (2.22)$$

Assim, é criado um histograma das orientações para pixels numa região em redor do ponto chave, em que de todas as orientações obtidas para um ponto, apenas o maior pico e aquelas acima de 80% do valor desse pico são utilizadas para definir a orientação de cada ponto chave.

Por fim, é possível a construção dos descritores para os pontos chave definidos como o ponto a seguir apresenta.

4. **Construção do descriptor local:** este é o último passo em que é criado o descriptor local para cada ponto de interesse. Para esse processo, é considerado um bloco 16x16 em redor do ponto chave e posteriormente dividido em 16 sub-blocos de 4x4. Por cada sub-bloco é criado um histograma com 8 picos relativos à orientação. Isto faz que no final seja extraído um vetor de 128 posições para cada ponto chave. Para além deste retorno, são também tomadas várias medidas de modo a que exista robustez suficiente para que esse ponto de interesse seja invariante a mudanças de iluminação e rotação.

O algoritmo SIFT é regularmente utilizado em reconhecimento de objetos ou cenas, tendo sido utilizado em deteção de objetos em *frames* de vídeo [28, 29] com utilização de técnicas mais avançadas de deteção de objetos e cenas utilizando vocabulários visuais como é apresentado na secção 2.3.7.

2.3.6.2 SURF

Outro algoritmo importante de extração de descritores locais é o SURF (*Speeded Up Robust Features*). Este é baseado no SIFT apresentado na secção 2.3.6.1 e também é utilizado, por exemplo, em reconhecimento de objetos ou mesmo na reconstrução 3D. Segundo os autores [5] o SURF é mais rápido (cerca de dez vezes) e robusto do que o SIFT, sendo que, segundo a comparação realizada em [30] comprovou-se que o SIFT é mais lento e não muito bom a mudanças de iluminação, mas apresenta melhores resultados a variações de rotação, mudanças de escala e transformações na imagem.

Este usa a técnica da imagem integral, onde cada pixel de uma imagem recebe um valor igual à soma dos pixels da sua esquerda e acima, incluindo o próprio. Este também utiliza um filtro Haar em formato de caixa numa sub-região 4x4 em redor de um ponto de interesse, como se pode ver na Figura 2.15, tornando o processo computacionalmente eficiente. Isto é realizado calculando a soma das respostas dos filtros e a soma do módulo das respostas dos filtros nas direções horizontal e vertical, gerando assim 4 valores por cada sub-região. Logo, o SURF retorna um vetor de 64 dimensões, metade do retornado pelo algoritmo SIFT.

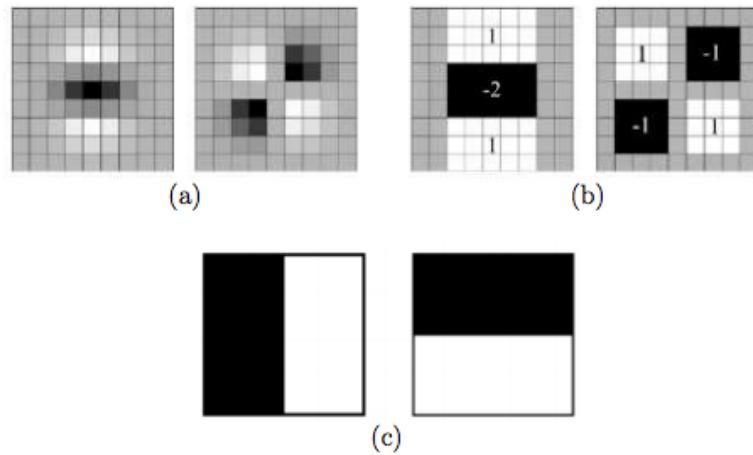


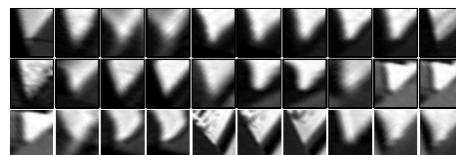
Figura 2.15: (a) Filtros Gaussianos de segunda ordem nas direções yy e xy ; (b) aproximação por filtros de caixa; (c) filtros de Haar; As regiões a cinzento têm valor igual a zero. *Retirada de [5]*.

2.3.7 Descritores Baseado em Vocabulário Visual

No reconhecimento de documentos de texto é utilizado o conceito de vocabulário, sendo muitas vezes designado por *BoW* (*Bag Of Words*) que em português significa, saco de palavras, onde existe um conjunto de palavras armazenadas pré definidas. Um texto é assim caracterizado, através da análise das palavras que possui, sendo contabilizado a frequência de palavras que estejam simultaneamente no texto e no *bag of words*, que assim atribuem um significado ao texto. É também utilizado uma lista de palavras que não acrescentam significado a expressões tais como, "o" ou "um", entre outras, que são removidas do texto durante a análise para não influenciarem os resultados. Um sistema de extração de informação de texto apresenta um número padrão de etapas [31].

Recentemente esta técnica foi adotada em aplicações de extração de informação visual, como por exemplo é nos mostrado em [28, 29]. Os autores recorrem a descritores locais invariantes a escala e rotação, para criar um vocabulário visual. A utilização de descritores locais como o SIFT, referido na secção 2.3.6.1, permite a extração de pontos de interesse nas imagens, invariantes a rotação e mudanças de escalas.

Quando efetuado este processo repetidamente com um grande conjunto de imagens, é possível criar *clusters* de regiões de imagens muito semelhantes entre si, como se pode ver na Figura ???. Todas estas regiões são candidatas a palavras visuais, sendo selecionada a que representa melhor esse *cluster*, isto é, é selecionado o centroide desse conjunto de regiões semelhantes entre si, recorrendo ao algoritmo *k-means* referido na secção 2.1.1. Esse centroide passa então a ser considerado uma palavra visual e é adicionado a um vocabulário com outras palavras visuais.



Por fim é necessária a realização de uma indexação de cada palavra visual às imagens, sendo utilizado um vetor para cada imagem que indica, o número de vezes em que uma determinada palavra visual se repete numa imagem. Neste caso o vetor funciona como em *text mining*, em que existe a contagem da frequência de palavras que ocorrem num determinado documento.

Outro processo possível para indexação do conteúdo visual ou texto, é atribuição de um peso ou ponderação para cada palavra visual numa determinada imagem. Aqui é utilizado a ponderação padrão conhecida como tf-idf ('*term frequency-inverse document frequency*') [28].

Considerando um vetor com k palavras visuais $Vd = (t_1, \dots, t_i, \dots, t_k)$, em que a ponderação de cada palavra é dada pela equação 2.23

$$t_i = \frac{n_{id}}{n_d} \log \frac{N}{n_i} \quad (2.23)$$

onde n_{id} é o numero de ocorrências da palavra i num documento d , n_d o número total de palavras no documento d , n_i é o numero de ocorrências da palavra i em todos documentos e N é o numero de documentos existentes.

Conclui-se assim que este descritor permite a identificação de imagens semelhante ou reconhecimento de objetos, através da comparação dos vetores de cada imagem, com a informação relativa ao vocabulário posteriormente criado. Este demonstra ser bastante eficiente, e segundo Nistér e Stewénius [32] é possível escalar este processo para enormes quantidades de imagens sem perda de performance utilizando para isso um vocabulário em forma de árvore, isto é, com a hierarquização das palavras visuais de um vocabulário visual, recorrendo para isso ao *clustering* hierárquico.

Capítulo 3

Módulo da Informação Visual

Neste capítulo será introduzido o módulo de informação visual desenvolvido, sendo apresentadas a estrutura e organização deste sistema. Como referido no capítulo 1, este projeto tem como objetivo estender a ferramenta TweeProfiles descrita na secção 2.2, dando-lhe uma dimensão de conteúdo diferente à que possui. Em conjunto com a informação espacial e temporal, O conteúdo das imagens partilhadas em tweets passa a ser uma fonte de informação e não o texto. Para que isto seja realizável, foi necessário o desenvolvimento de um módulo que efetue a recolha os dados com a devida filtragem, extraia e processe a informação visual e armazene essa informação de modo a que fosse possível a sua integração com o TweeProfiles.

3.1 Recolha dos Dados

O desenvolvimento deste módulo apenas era realizável com um conjunto de imagens partilhadas no serviço de microblogging Twitter, sendo fundamental a recolha dos dados necessários para esse processo, neste caso, os Tweets. Nesta secção é feita uma descrição do conjunto dos dados disponíveis, das filtragens que foram necessárias realizar e uma apresentação do conjunto de dados finais obtidos e utilizados no desenvolvimento deste projeto de dissertação.

3.1.1 Descrição dos Dados

O primeiro passo para a realização deste projeto de dissertação foi a recolha dos dados necessários. Estes dados foram recolhidos através de uma base de dados MongoDB previamente criada usando a plataforma Socialbus, anteriormente designada por TwitterEcho [11]. Este dados são estruturados sobre a forma de objetos JSON e possuem informação relativa a cada tweet. Estes objetos encontram-se todos num só documento MongoDB que se caracteriza pelas características apresentadas na Tabela 3.1

Estes tweets foram recolhidos entre o dia 17 e 19 de Junho de 2013 com conteúdo partilhado somente contendo texto escrito em português do Brasil, filtrados pela ferramenta Socialbus [11]

	Total	Com imagem	Twitter	TwitPic	Instagram
Nº Tweets	1704273	86349	202	6100	79210

Tabela 3.1: Descrição em números do total de tweets com indicação, nos que contém URL para imagem, do número de tweets por serviço de partilha de imagem

já referida anteriormente. Estas datas coincidiram com um evento ocorrido no Brasil, mais especificamente, as manifestações do ano passado do povo brasileiro contra o seu governo. Este foi um dos motivos da escolha desta base de dados, pois apresentava tweets que poderiam ser interessante para encontrar padrões ou eventos através das imagens partilhadas pelos brasileiros nas ruas, aliadas sempre às dimensões espaço-temporais.

3.1.2 Filtragem dos Dados

Após estar definido o conjunto de dados a utilizar, foi necessário realizar uma filtragem dos dados de modo a apresentarem a informação necessária para a realização deste projeto. O primeiro passo desta filtragem foi recolher todos os tweets que contivessem no seu objeto um URL para uma imagem, sendo que esse URL teria de pertencer a um dos seguintes serviços:

- Twitter
- TwitPic
- Instagram

O URL teria que ser válido, isto é, foi feita uma verificação prévia se a imagem estaria ainda disponível através do endereço existente.

Para ser mais fácil posteriormente uma seleção mais cuidadosa dos tweets foi criada uma base de dados local (SQLite) com a seguinte tabela:

```

1 create table if not exists IMAGENS (
2     id integer PRIMARY KEY AUTOINCREMENT,
3     id_tweet text ,
4     servico text ,
5     url text ,
6     tipo text ,
7     retweet text
8 );

```

Em que se descreve cada coluna da seguinte forma:

id - id da linha da tabela;

id_tweet - id do tweet na base de dados Mongodb;

servico - nome do serviço de alojamento da imagem;

url - endereço url para a imagem fonte;

tipo - este atributo identifica se a imagem pertence a um tweet ou retweet;

retweet - caso a imagem pertença a um retweet, este atributo pode assumir o valor "primeiro" no caso de ser o primeiro retweet, do tweet original, na base de dados MongoDB, ou caso contrário, assume o valor NULL.

Isto permitiu realizar de uma forma rápida alguns teste no *download* de algumas imagens pelos diferentes serviços, para além de permitir fazer uma seleção fácil e rápida de tweets, retweets ou por exemplo, do primeiro retweet no caso de existir vários retweets de um determinado tweet.

Após a criação desta base de dados, ficou decidido selecionar todos os objetos da base de dados MongoDB do tipo tweet em que o seu serviço fosse o Instagram. Esta decisão deveu-se ao facto dos retweets se referirem a imagens já existentes ocorrendo duplicação de dados. No caso da escolha do Instagram, este deveu-se ao facto de apresentar um número superior de imagens relativamente aos outros serviços, visto que ao se excluir os retweets o número de imagens dos restantes serviços passou a um valor pouco significativo.

Para além destes critérios é importante salientar que foi necessário selecionar apenas os dados com georreferenciação.

3.1.3 Conjunto de Dados Final

Por fim armazenou-se um ficheiro JSON com todos os dados que serão utilizados e foi realizado o *download* de todas as imagens relativas a cada tweet e armazenadas localmente em formato JPEG, que se apresentava como formato de origem das imagens descarregadas.

Relativamente aos objetos de cada tweet presentes no ficheiro JSON, optou-se por não armazenar todas as instâncias para reduzir o tamanho do ficheiro, tendo sido apenas incluídas as representadas no seguinte exemplo de um objeto de um tweet:

```

1 {
2     "_id": {
3         "$oid": "52c6d0f08ef20d397e42b516"
4     },
5     "coordinates": {
6         "coordinates": [
7             -8.61136747,
8             41.14668427
9         ]
10    },
11    "created_at": "Tue Jun 18 17:02:09 +0000 2013",
12    "entities": {
13        "urls": [
14            {
15                "display_url": "instagram.com/p/atTgTbEu0S/"
16            }
17        ]
18    },
19    "id_str": "347036525172772864",

```

```

20   "text": "#ogiganteacordou #DilmaNAO #brasilnraua #foradilma #
21   verasqueumfilhoteunaofogealuta #vamospararuas \u2026 http://t.co/AF0hcTxUBX"
22   ,
23   "user": {
24     "id_str": "1072354428",
25     "name": "Carlos Roma"
26   }
27 }
```

No caso das imagens armazenadas, de modo a que fosse associada cada uma das imagens ao seu respetivo tweet, foi atribuído o campo *id_str* presente no objeto JSON ao nome do ficheiro de imagem JPEG. Assim ao utilizar uma das imagens, para saber a que tweet pertence apenas é necessário procurar o objeto JSON que possua o atributo *id_str* igual ao nome do ficheiro da imagem.

Recolhidos todos os identificadores dos tweets pertencentes ao serviço Instagram que apenas fossem do tipo tweet e que os respetivos tweets apresentassem a informação de geolocalização no campo *coordinates*, avançou-se com o processo de *download* de todas as imagens, tendo sido efetuado com sucesso o *download* de 5964 imagens em 7195 possíveis.

Em suma, o conjunto de dados final utilizado para o desenvolvimento deste projeto foi de 5964 objetos relativos aos tweets contidos num ficheiro JSON e as respetivas 5964 imagens.

3.2 Extração, Processamento e Armazenamento da Informação Visual

O passo seguinte no desenvolvimento do módulo da informação visual foi a implementação de um sistema capaz de extrair, processar e armazenar a informação visual através das imagens armazenadas localmente. Para o desenvolvimento deste modulo foi utilizada a linguagem Python, pela sua capacidade de integração de diferentes bibliotecas desde manipulação de estruturas de dados, até mesmo a bibliotecas de manipulação e processamento de imagem. Para a concretização do modelo foram tidos como linhas guia, o processo desenvolvimento de uma ferramenta de pesquisa de imagens apresentado em [33]. A arquitetura deste módulo desenvolvido é apresentada na Figura 3.1.

Foram desenvolvidos três sub-módulos diferentes: o primeiro é responsável pela extração dos descritores locais, o segundo utiliza o primeiro para gerar um vocabulário visual, e o terceiro utiliza os dois anteriores para armazenar um histograma descritor de cada imagem numa base de dados indexada à sua respetiva imagem. Em seguida é apresentada uma subsecção com a descrição para cada sub-módulo, como representado na Figura 3.1

3.2.1 Extração dos Pontos de Interesse e Descritores Locais

O primeiro passo no desenvolvimento deste módulo passou pela extração da informação visual. Esta informação visual deveria representar uma imagem eficientemente e de forma a possi-

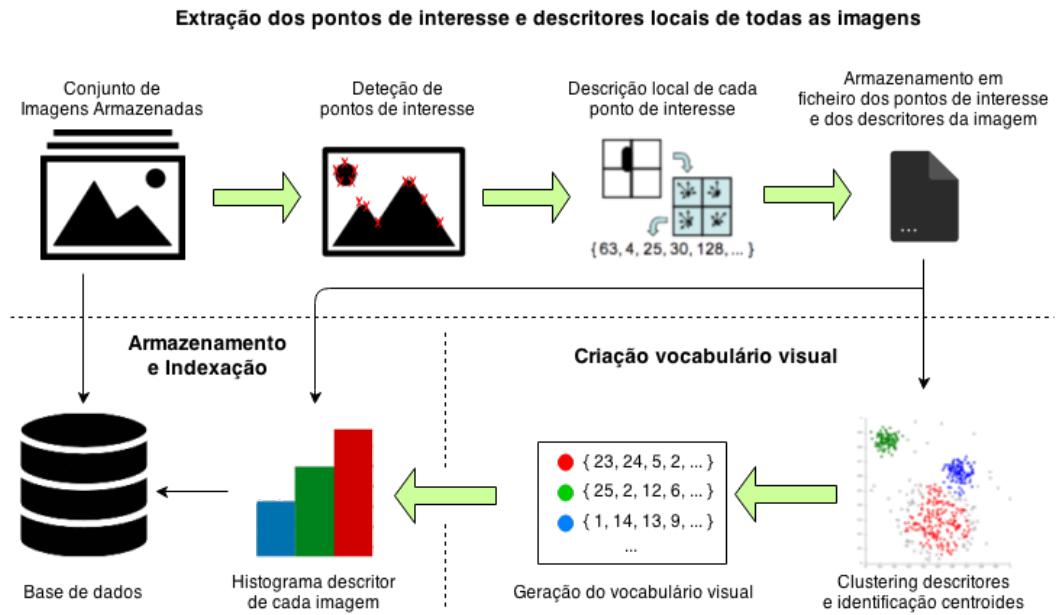


Figura 3.1: Arquitetura do módulo de extração, processamento e armazenamento da informação visual. *Adaptada de [6]*

bilitar a criação de um vocabulário visual. Uma das formas possíveis e apresentadas na secção 2.3 é a utilização de um descriptor local. Neste caso foi utilizado o descriptor SIFT [27, 4], pois como referido na secção 2.3.6, apesar de ser mais lento e menos eficiente a mudanças de iluminação, este apresenta melhores resultados a variações de rotação, mudanças de escala e transformações na imagem. Para além disso, foi utilizada a ferramenta e biblioteca *open source* VLFeat [34] que integra alguns dos algoritmos mais utilizados em visão computacional, e que inclui o algoritmo SIFT. Este, apesar de não possuir uma biblioteca para Python, permite a sua utilização através da linha de comandos.

Para o conjunto de imagens existentes foi necessário criar uma cópia de cada imagem em escalas de cinzento e em formato *.pgm* para ser utilizada pelo VLFeat. Utilizando assim estas imagens, o VLFeat armazena num ficheiro com o formato *.sift* os pontos de interesse e os descritores de uma imagem, sendo necessário criar um ficheiro para cada imagem. Nesses ficheiros os dados são armazenados em formato ASCII. A informação armazenada apresenta-se da seguinte forma

```

1 318.861 7.48227 1.12001 1.68523 0 0 0 1 0 0 0 0 0 11 16 0 ...
2 318.861 7.48227 1.12001 2.99965 11 2 0 0 1 0 0 0 173 67 0 0 ...
3 54.2821 14.8586 0.895827 4.29821 60 46 0 0 0 0 0 0 99 42 0 0 ...
4 155.714 23.0575 1.10741 1.54095 6 0 0 0 150 11 0 0 150 18 2 1 ...
5 42.9729 24.2012 0.969313 4.68892 90 29 0 0 0 1 2 10 79 45 5 11 ...
6 229.037 23.7603 0.921754 1.48754 3 0 0 0 141 31 0 0 141 45 0 0 ...
7 232.362 24.0091 1.0578 1.65089 11 1 0 16 134 0 0 0 106 21 16 33 ...
8 201.256 25.5857 1.04879 2.01664 10 4 1 8 14 2 1 9 88 13 0 0 ...
9 ... ...

```

onde cada linha contém as coordenadas, escala e ângulo de rotação para cada ponto de interesse, nos primeiros 4 valores respetivamente, correspondendo os restantes ao vetor descritor de tamanho 128 como referido na secção 2.3.6. Como o objetivo era utilizar as imagens originais, foram eliminadas as imagens temporárias com o formato *.pgm*. O passo seguinte passou pelo desenvolvimento do submodelo responsável pela criação do vocabulário visual, que é apresentado na subsecção seguinte.

3.2.2 Criação do Vocabulário Visual

Este módulo é o responsável pela criação do vocabulário visual. Para a sua concretização foi necessário utilizar os ficheiros de extensão *.sift* reproduzidos através da ferramenta VLFeat [34] no módulo anterior. Os passos seguintes basearam-se nos descritos em [33].

As palavras visuais não são nada mais do que um conjunto de vetores de características de imagens. Assim um vocabulário visual é o conjunto destas palavras visuais. Como todas as imagens possuem muitos descritores locais, sendo que muitos podem ser semelhantes, é necessário agrupar todos os descritores de um conjunto de imagens e detetar aqueles que possam representar um conjunto de descritores semelhantes, e assim formar várias palavras visuais, sendo uma palavra visual um centroide de um grupo.

Para criar um vocabulário visual foi então necessário utilizar um algoritmo de *clustering* por partição, tendo sido escolhido o *k-means* por ser um dos mais utilizados e eficiente, como foi referido no capítulo 2 na secção 2.1.1. O algoritmo foi aplicado aos descritores de um subconjunto de imagens aleatoriamente selecionadas do conjunto de imagens armazenadas localmente. Neste caso foi utilizado aproximadamente 8% das imagens para não comprometer a nível de tempo de processamento. Como este algoritmo implica a pré definição do número de *clusters*, foi atribuído a *k* o valor 1000. Isto significa que são gerados cerca de 1000 *clusters*, logo serão retornados aproximadamente 1000 centroides, o que significa que o nosso vocabulário visual possuirá cerca de 1000 palavras visuais.

Para utilizar este vocabulário visual é necessário armazena-lo e indexar cada palavra visual a cada imagem. O sub-módulo responsável por este processo é descrito na secção a seguir.

3.2.3 Armazenamento da Informação Visual

Com o vocabulário visual criado foi necessário armazenar esta informação e indexar cada palavra visual as imagens, de modo a seja possível a comparação entre imagens diferentes. Para armazenar este vocabulário desenvolveu-se módulo que execute esta tarefa, sendo que, foi necessário seguir alguns passos tais como, a criação de um histograma descritor do vocabulário visual de cada imagem e criação de uma base de dados com a informação necessária para a utilização do histograma descritor de cada imagem.

A criação do histograma foi o primeiro passo do desenvolvimento deste sub-módulo, em que foi utilizado o vocabulário disponibilizado através do sub-módulo apresentado anteriormente. Como o vocabulário é constituído por vetores descritores que representam cada uma das palavras

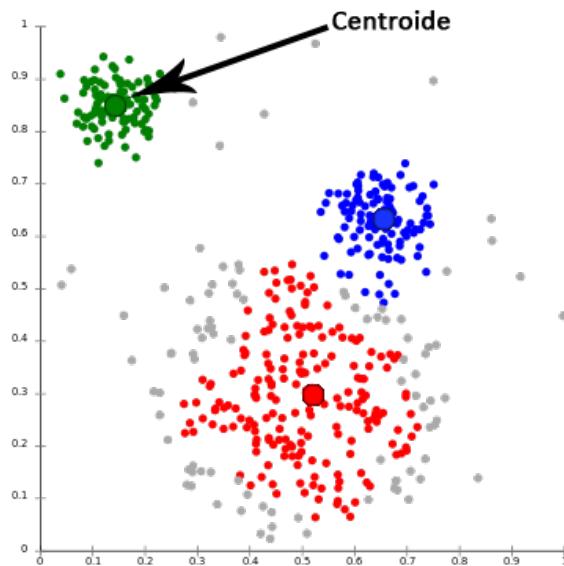


Figura 3.2: Exemplo de projeção de centroides após tarefa de *clustering* num espaço a duas dimensões.

visual, para a criação do histograma para cada imagem foi necessário utilizar novamente os ficheiros com os descritores locais SIFT de cada imagem e assim projetar num espaço utilizando novamente o algoritmo *k-means* de modo a atribuir a cada ponto chave de uma imagem a sua palavra visual respetiva. Foi então possível após este processo criar um histograma com a contagem das palavras visuais em cada imagem.

Assim o próximo passo passou pelo armazenamento dos histograma e a indexação a sua respetiva imagem. Como todo este módulo foi operado em modo *offline*, optou-se pela utilização de uma base de dados local SQLite. Esta funciona de modo semelhante a uma base de dados MySQL ou PostgreSQL, sendo que pode ser desenvolvida e acedida sem recurso a um servidor. Para a sua concretização foi criado um esquema muito simples apenas com três tabelas como ilustrado na tabela 3.2. A tabela *imlist* contém o nome de todas as imagens através do atributo *filename*, a tabela *imwords* contém índice das palavras visuais através do atributo *wordid*, a identificação do vocabulário com o atributo *vocname* e o índice das imagens em que as palavras visuais aparecem com o atributo *wordid*. Por fim, a tabela *imhistograms* contém o histograma de palavras visuais completo para cada imagem com o atributo *histogram*.

3.3 Matriz de Distâncias entre Imagens

Com o módulo descrito na secção anterior 3.2, a informação que descreve cada uma das imagens armazenadas localmente está alojada e é acessível através da base de dados local criada. Esta

imlist	imwords	imhistograms
rowid	imid	imid
filename	wordid	histogram

Tabela 3.2: Esquema da base de dados SQLite

informação está no formato de um histograma que descreve a imagem através de um vocabulário visual. Este histograma é um vetor de tamanho único e igual para todas imagens, o que permite uma fácil comparação entre eles, o que é equivalente a dizer que estes permitem calcular a distância entre eles. Como cada histograma está associado a uma imagem, calcular a distância entre dois histogramas equivale a calcular a distância entre duas imagens.

O TweeProfiles utiliza o algoritmo DBSCAN apresentado no Capítulo 2 na secção 2.1.3 para realizar a tarefa de *clustering* ao conteúdo e às dimensões espacial e temporal. O *clustering* baseado em densidades, neste caso mais concreto, com o recurso ao algoritmo DBSCAN, pode utilizar uma matriz distância entre objetos para realizar a tarefa de *clustering*, sendo que retorna um número indefinido de *clusters* dependentes dos dados. A matriz distância trata-se apenas de uma matriz NxN em que N é o número total de objetos. No nosso caso, os objetos são as imagens e o valor de N é igual a 5964. A figura 3.3 apresenta a estrutura de uma matriz distância, em que 0 é a distância de um objeto a ele mesmo.

$$\begin{bmatrix} 0 & & & \\ d(2,1) & 0 & & \\ d(3,1) & d(3,2) & 0 & \\ \vdots & \vdots & \vdots & \\ d(n,1) & d(n,2) & \dots & \dots & 0 \end{bmatrix}$$

Figura 3.3: Estrutura de uma matriz de distâncias. *Retirada de [7]*

Tendo existido algumas limitações a nível de *hardware* para o cálculo de uma matriz desta dimensões, optou-se por utilizar o mesmo procedimento utilizado no TweeProfiles [10] e dividir os dados em três partes distintas. Esta divisão foi efetuada tendo em conta uma divisão temporal, isto é, foram em primeiro lugar ordenados os tweets por ordem cronológica e apenas posteriormente foi dividido em três partes iguais de 1986 tweets cada conjunto. Após este processo foi então calculadas as três matrizes.

Para calcular as matrizes de distâncias utilizou-se o histograma descritor de cada imagem, calculando a distância entre histogramas através da função distância euclidiana apresentada no capítulo 2 na secção 2.1.5.

Com as matrizes distâncias calculadas, estavam as condições reunidas para a integração do modelo no TweeProfiles. O próximo capítulo irá introduzir a integração deste modelo com o TweeProfiles e o desenvolvimento da ferramenta Olhó-passarinho e os seus resultados ilustrativos.

Capítulo 4

Olhó-passarinho

Neste capítulo será abordado a ferramenta desenvolvida com a descrição da arquitetura sistema implementado, do processamento da informação espaço-temporal e da sua integração com a informação visual de modo a aplicar a tarefa de *clustering*. Por fim será apresentado a visualização dos resultados ilustrativos e consequentemente a sua discussão.

4.1 Arquitetura do Sistema

A arquitetura do sistema desenvolvido é apresentado na figura 4.1. Este apresenta uma divisão entre os serviços externos e o modelo desenvolvido. Este modelo foi desenhado de modo a que existisse uma separação entre o tratamento de toda a parte de processamento dos dados e a visualização, existindo assim um *back-end* com todos os ficheiros e módulos desenvolvidos e um *front-end* que representa a aplicação web para visualização dos resultados. No *back-end* existe também uma divisão entre dois módulos fundamentais, o módulo de processamento da informação visual, responsável por tratar a informação das imagens como descrito no Capítulo 3, de modo a que essa informação possa ser utilizada pelo módulo responsável pelo processo de *Data Mining* já desenvolvido no TweeProfiles [10].

Os serviços externos correspondem à base de dados Mongodb para a recolha dos tweets e os serviços Twitter e Instagram para a recolha das imagens através do URL. No caso do modelo desenvolvido, a parte de *back-end* possuí os ficheiros JSON com os dados e as imagens necessárias, tanto para o módulo de processamento da informação visual como para a extração e processamento do dados espaço-temporais, explicados na próxima secção 4.2. Os dados processados no módulo da informação visual e os dados espaço-temporal extraídos dos tweets são assim utilizados no processo de *Data Mining*, onde é aplicada a tarefa de *clustering* como explicado na secção 4.3 apresentada mais adiante. De este processo resultam os *clusters* calculados através de vários parâmetros, sendo estas informações armazenadas em ficheiros. Por fim, foi utilizada a *microframework* Flask para desenvolvimento de aplicações web em Python, que permitiu o desenvolvimento da aplicação Olhó-passarinho para visualização dos resultados.

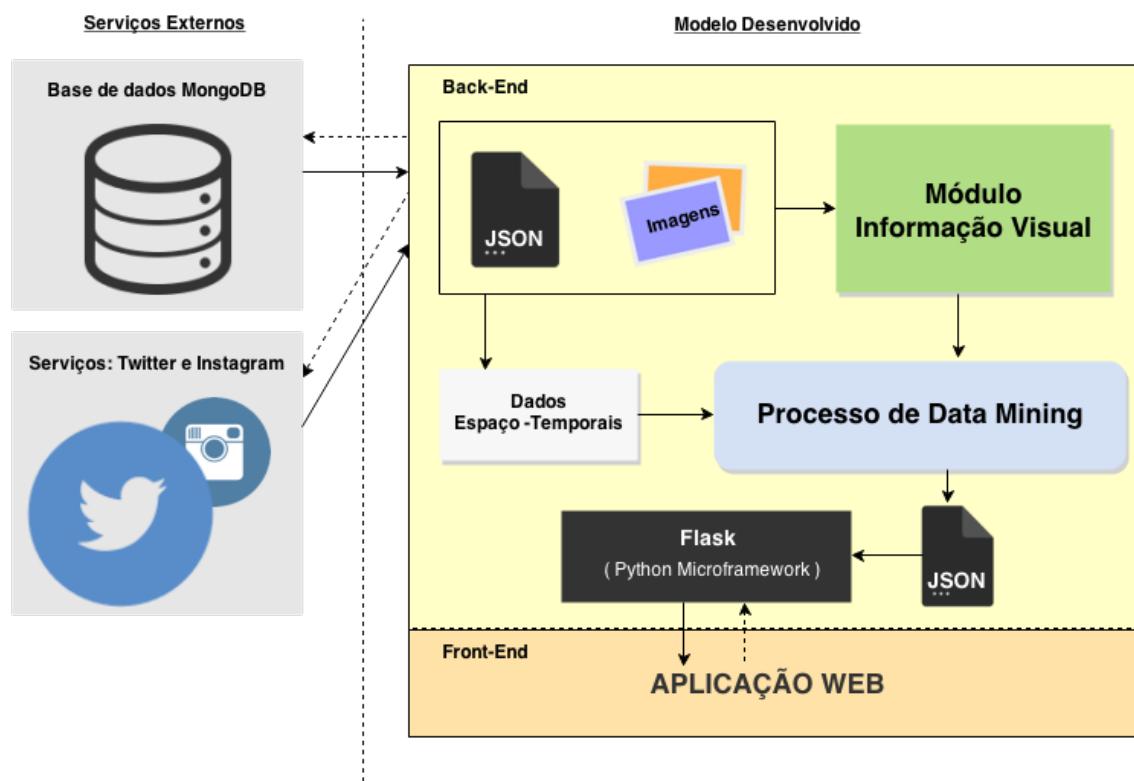


Figura 4.1: Arquitetura do sistema completo

4.2 Informação Espaço-Temporal

Uma das características principais tanto do TweeProfiles como do Olhó-paddarinho é a integração das dimensões espaço-temporais com o conteúdo. Tal como no TweeProfiles, também aqui foi utilizado estas dimensões, tendo sido então recolhido a informação de tempo e espaço dos tweets para o cálculo das respetivas matrizes de distância entre tweets.

Em primeiro lugar foi recolhida a informação espacial. Neste caso os dados possuem a informação de latitude e longitude do ponto onde foi enviado o tweet. Para calcular a distância entre tweets utilizou-se a função distância Haversine abordada no capítulo 2 na subsecção 2.3.3.1. Neste caso foi calculada a distância em quilómetros, tendo sido considerado o valor do raio da Terra igual a 6371 Km.

Posteriormente foi então recolhida a informação temporal dos tweets. Esta informação apresenta-se no seguinte formato:

Tue Jun 18 17:02:09 +0000 2013

Para o cálculo da distância entre datas foi utilizada a função distância euclidiana, que se pode resumir ao módulo da diferença entre o tempo de dois tweets, como pode ser visto no capítulo 2 na subsecção 2.1.5.3. Utilizando as bibliotecas *datetime* e *dateutil* este cálculo é direto sem necessitar ser realizada uma conversão do formato da data.

4.3 Clustering da Informação Visual, Espacial e Temporal

A tarefa de *clustering* é o passo final para a obtenção dos resultados finais. Este é o processo que engloba os dados resultantes de todo o processamento da informação tratada e discutida anteriormente.

Para a tarefa de *clustering* optou-se pela utilização do algoritmo implementado no TweeProfiles [10], o DBSCAN. Este apresenta algumas vantagens na utilização de dados recolhidos de redes sociais, pois este não necessita de uma predefinição do número de *clusters* que se pretende obter, sendo o número de *clusters* é definido através da distribuição em densidade dos objetos, como referido no capítulo 2 na subsecção 2.1.3.

Mas antes de fornecer os dados, neste caso as matrizes já calculadas com as distâncias entre tweets para as dimensões temporal, espacial e de conteúdo visual, foi necessário recorrer a sua normalização de modo a que fosse possível a combinação entre as várias matrizes. Foi então utilizada a normalização através da média e do desvio padrão.

Após a normalização, realizou-se a combinação entre as matrizes, com atribuição de vários pesos a cada matriz, de forma a que a soma dos diferentes pesos fosse igual a 1. Inicialmente, a distribuição dos pesos foi feita com passos de 0.25 pontos, isto é, os valores de cada matriz podiam assumir os pesos: 0, 0.25, 0.5, 0.75 e 1. Isto permitiu uma fácil divisão dos pesos, mas não permitia atribuir pesos iguais às três matrizes, pelo número de matrizes ser ímpar. Assim, optou-se por dividir os pesos em fator de 0.333, em que os valores de cada matriz podem assumir os pesos: 0, 0.333, 0.666, 1. Isto faz que a soma dos pesos possa não dar exatamente 1, mas sim 0.999. Por outro lado, isto possibilitou atribuir o mesmo peso às três matrizes diferentes de cada subconjunto.

A tabela 4.1 apresenta as diferentes combinações possíveis de pesos para as diferentes dimensões

Combinação	Visual	Espacial	Temporal
1	1.000	0.000	0.000
2	0.666	0.333	0.000
3	0.666	0.000	0.333
4	0.333	0.333	0.333
5	0.000	1.000	0.000
6	0.333	0.666	0.000
7	0.000	0.666	0.333
8	0.000	0.000	1.000
9	0.333	0.000	0.666
10	0.000	0.333	0.666

Tabela 4.1: Tabela com as possíveis distribuições de pesos entre as várias dimensões

4.4 Visualização

A aplicação web desenvolvida é a responsável pela visualização dos resultados obtidos pela tarefa de *clustering*. Esta apresenta três secções principais de visualização dos *clusters* pelas diferentes dimensões utilizadas.

O primeiro é um mapa, onde são apresentados a distribuição dos tweets, como pode ser visto na figura 4.2 e a respetiva distribuição dos *clusters* geograficamente. Este foi desenvolvido recorrendo à API Javascript do Google Maps v3 [35] disponibilizada pela própria Google, sendo toda ela controlada através da linguagem de programação Javascript. Esta permite utilizar um mapa e controlar de modo a adicionar componentes a esse mapa. Os tweets foram assim representados por pequenos círculos azuis, e os *clusters* como círculos vermelhos com transparência.



Figura 4.2: Distribuição de tweets na dimensão espacial

A segunda secção é responsável pela representação da distribuição dos *clusters* na dimensão temporal e foi utilizado a ferramenta Google Charts, mais especificamente a API Timeline [8]. Esta, tal como a API do Google Maps, também é desenvolvida em javascript, e permite reproduzir um gráfico com barras de duração temporal, como podemos ver na figura 4.3.



Figura 4.3: Exemplo ilustrativo da ferramenta Timeline. *Retirada de* [8]

Por último, a secção de visualização de imagens que pertencem a um *cluster*. Nesta é apresentado uma matriz com nove imagens presentes nos *clusters* escolhidas aleatoriamente, havendo ainda a possibilidade de ir modificando as imagens visíveis. É possível também clicar numa das imagens e visualizar a imagem numa dimensão maior e ver a informação relativa à imagem como o utilizador que fez a partilha e o texto partilhado em conjunto. Para além disto, é possível aceder ao tweet original através de um botão com essa indicação. Esta secção também indica o nome do *cluster* e a informação temporal em texto do mesmo.

A aplicação apresenta também os controlos para escolher o subconjunto que se pretende visualizar, e controlos para definir o peso que se pretende atribuir a cada dimensão.

4.5 Resultados Ilustrativos

Nesta secção serão apresentados alguns resultados ilustrativos visíveis através ferramenta Olhó-passarinho. Neste relatório não é possível apresentar todos os resultados devido ao número de diferentes possíveis combinações.

A Figura 4.4 mostra um exemplo da visualização da distribuição dos *clusters* no espaço geográfico. Este exemplo é relativo a uma divisão igual dos pesos entre as três dimensões, espacial, temporal e as fotografias, tendo cada uma um peso de 33.33% nos resultados dos *clusters*.

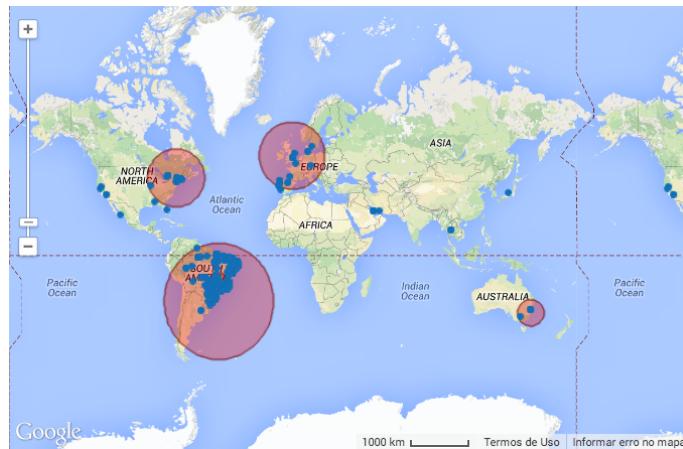


Figura 4.4: Exemplo ilustrativo da visualização da distribuição espacial dos *clusters* para o caso em que a distribuição dos pesos entre as três dimensões é igual

A dimensão dos círculos que representam os *clusters* é proporcional ao número de tweets presentes nesse mesmo *clusters* isto é, quanto maior o número de tweets maior é a respetiva circunferência.

Já Figura 4.5 apresenta a distribuição dos cinco diferentes *clusters* da mesma situação anterior, sendo possível visualizar a existência de *clusters* com uma margem de maior duração, enquanto outros como o caso do *Clusters 3* que apresenta uma duração de menos de 4 horas. A distribuição no tempo do conjunto de tweets utilizados para este exemplo, como podemos ver na Figura 4.5, é

de aproximadamente 24 horas, em que teve início no dia 17 de Junho de 2013 e terminou no dia 18 de Junho de 2013.

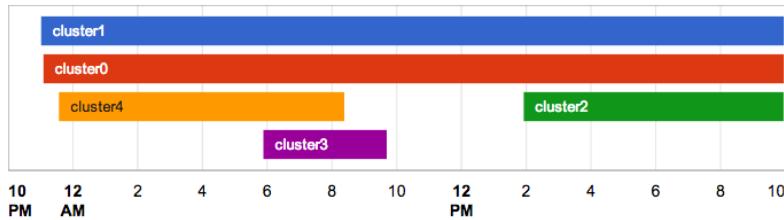


Figura 4.5: Exemplo ilustrativo da visualização da distribuição temporal dos *clusters* para o caso em que a distribuição dos pesos entre as três dimensões é igual

A Figura 4.6 traz-nos uma perspetiva global da aplicação web, mostrando mais detalhe sobre um dos *clusters*, neste caso o *Cluster 2*. Na Figura 4.6 é possível a visualização de nove imagens aleatórias do *Cluster 2*. Já no caso da Figura 4.7, é apresentado a visualização mais pormenorizada de uma das imagens, com a indicação do nome do autor do tweet e a mensagem associada à imagem partilhada pelo mesmo. Também é possível aceder ao tweet original.

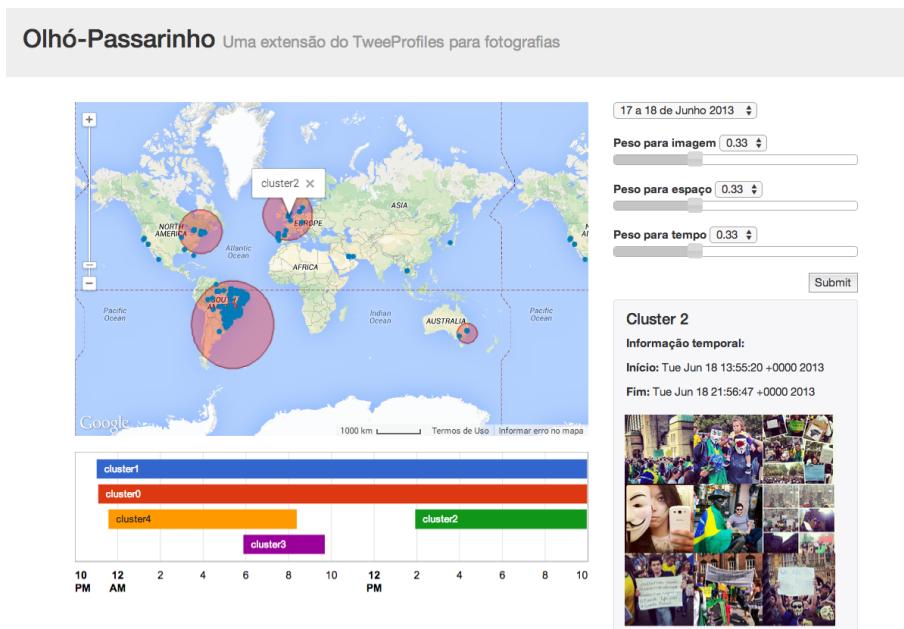


Figura 4.6: Exemplo ilustrativo da visualização completa da aplicação web com a visualização pelas várias dimensões

É interessante salientar que no exemplo ilustrado na Figura 4.6 existe uma semelhança geral entre as nove imagens, sendo isto possível, devido a combinação das três dimensões, tendo sido agrupado tweets com semelhança nas imagens, mas também no espaço e tempo.

Outro exemplo interessante, é quando a atribuição do peso de 100% apenas para a dimensão do conteúdo visual. A Figura 4.8 demonstra um exemplo desses, e como é visível, o *cluster* possui pelo menos nove imagens muito semelhantes, que apesar de não fazerem parte de um evento

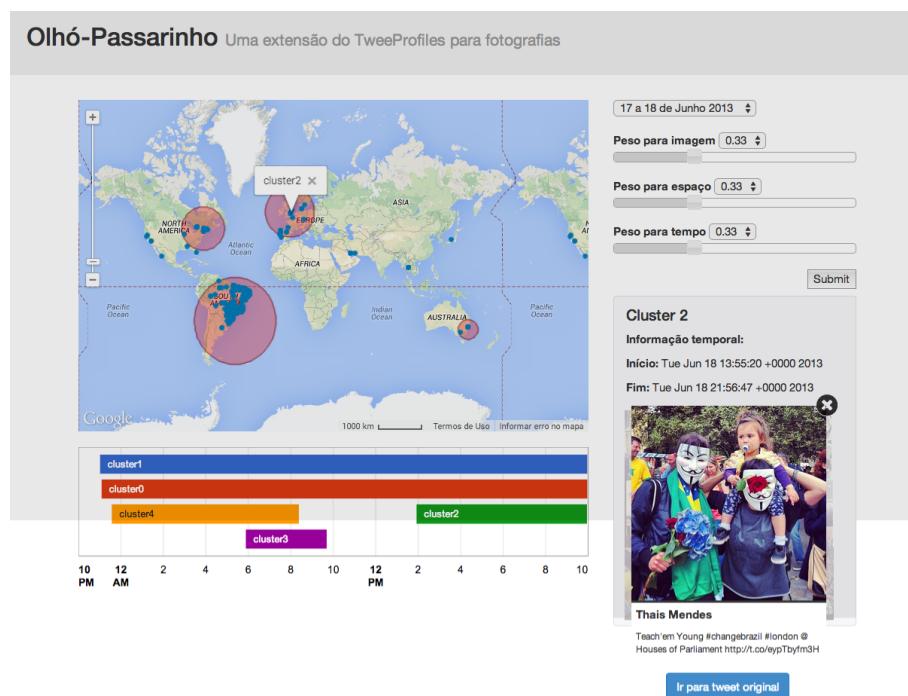


Figura 4.7: Exemplo ilustrativo da visualização completa da aplicação web com a visualização mais detalhada de uma das imagens

que inclua pessoas, este permitiu encontrar a partilha de várias partilhas de um mesmo texto por várias pessoas através de uma imagem. Este resultado também vem demonstrar que o modelo responsável pela extração da informação visual apresentou um bom desempenho, tal como era o esperado.



Figura 4.8: Exemplo ilustrativo da visualização de um conjunto de imagens pertencentes a um *Cluster* e em que foi atribuído o peso total apenas à dimensão do conteúdo visual

Capítulo 5

Conclusões e Trabalho Futuro

Neste capítulo são expostas algumas conclusões retiradas do desenvolvimento desta dissertação e são apresentadas sugestões para um trabalho futuro com indicação de melhorias a implementar e sugestões de

5.1 Resumo do Trabalho Realizado

Durante o período dedicado à realização do projeto de dissertação, foram seguidas uma sequência definida de etapas que culminou num sistema capaz de reproduzir a visualização no espaço, tempo de *clusters* e visualizar e navegar por fotografias partilhadas no serviço de *microblogging* Twitter contidas num determinado *cluster*.

Inicialmente foi feita uma recolha dos dados necessários ao desenvolvimento deste projeto de dissertação. Este dados foram recolhidos através de base de dados Mongodb e possuíam a informação relativa a tweets partilhados na rede social Twitter. Como o objetivo era a descoberta de padrões através de fotografias, foi deitado o *download* de todas as imagens pertencentes a tweets e partilhadas no Twitter através do serviço Instagram. Os dados desses tweets também foram armazenados no formato JSON.

O próximo passo foi o desenvolvimento de um módulo responsável pela extração, processamento e armazenamento da informação visual. Este foi desenvolvido para que representasse as imagens de uma forma mais eficiente e compacta, e que tornasse assim possível a comparação entre imagens para a criação de uma matriz distância para ser utilizada no processo de *Data Mining*.

Proseguí-se com a produção das matrizes de distância entre tweets pelas dimensões temporais, de forma a combinar esta informação com a informação visual, as fotografias. Com esta integração concluída utilizou-se essa informação no processo de *Data Mining* para a obtenção dos *clusters*, com a atribuição de diferentes pesos às diferentes dimensões.

Após a obtenção dos diferentes *clusters*, foi desenvolvida a aplicação web em Python, recorrendo a microframework Flask, para visualização dos resultados através do conteúdo dos tweets e das diferentes dimensões já referidas, resultando assim num sistema completo e funcional.

5.2 Trabalho Futuro

Após a finalização deste projeto de dissertação foi feita uma análise a todo o processo realizado, tendo sido concluído que os objetivos principais propostos foram atingidos. Apesar disso, alguns objetivos mais ambiciosos não foram possíveis ser atingidos devido a vários fatores, e que devem ser tidos em conta num trabalho futuro.

Um dos pontos de partida que num trabalho futuro deve ser tido em conta é a possibilidade de aceder a uma base de dados maior, pois apesar de existirem muitas imagens partilhadas no Twitter através de vários serviços, o número de tweets georeferenciados ainda é bastante reduzido. Para além disso, era interessante utilizar uma base de dados com uma extensão temporal superior e consequentemente, com conteúdos mais diversificados.

O desenvolvimento do modelo responsável pelo tratamento da informação visual, mais concretamente, da criação de um vocabulário visual, apresentou-se como uma boa opção com resultados muito satisfatórios, mas num trabalho futuro também seria interessante a integração de outros descritores, como por exemplo, descritores de cor, adicionando assim a componente cor. Isto iria permitir uma melhor descrição das imagens e possibilitaria identificar cenários mais específicos onde a cor é um fator determinante de distinção, como por exemplo, fotografias de praias, alimentos ou mesmo locais com vegetação, como jardins ou parques naturais onde predomina a cor verde.

Já na parte da aplicação, existe diferentes abordagens a poderem ser seguidas para diferenciarem a visualização dos resultados, e consequentemente melhorarem a capacidade do utilizador compreender melhor o que está a visualizar. Uma das possibilidades, seria a síntese de uma imagem que fosse representativa do *cluster* a que pertence, isto é, um sistema que analisasse todas as imagens contidas num *cluster*, e fosse capaz de reproduzir uma imagem modelo, através da informação de todas as imagens do *cluster*, e até mesmo, através de informação existente numa base de dados de imagens exterior.

Por fim, a utilização de uma ferramenta com este intuito tornar-se-ia mais interessante se, a informação disponível para visualização fosse constantemente atualizada. Para isso seria necessário a utilização de hardware com capacidade suficiente de analisar as imagens em tempo real e exportar a informação necessária ao processo de *Data Mining*. A utilização do vocabulário visual iria permitir a utilização de um serviço assim, sendo que seria necessário realizar algumas alterações, como por exemplo utilizar um vocabulário visual disposto de forma hierárquica, o que permitiria uma mais rápida descrição de uma imagem. Para além disso, o processo de *Data Mining* teria de estar constantemente em funcionamento de forma a atualizar os *clusters* sempre que existissem alterações nos mesmos ou mesmo no aparecimento de novos.

Anexo A

Exemplo objeto JSON de um tweet

```
1 {
2     "_id": {
3         "$oid": "52c6d0f28ef20d397e42c54a"
4     },
5     "contributors": null,
6     "coordinates": null,
7     "created_at": "Tue Jun 18 17:08:15 +0000 2013",
8     "entities": {
9         "hashtags": [],
10        "symbols": [],
11        "urls": [
12            {
13                "display_url": "twitpic.com/cxvh38",
14                "expanded_url": "http://twitpic.com/cxvh38",
15                "indices": [
16                    104,
17                    126
18                ],
19                "url": "http://t.co/P7nwF6GOFc"
20            }
21        ],
22        "user_mentions": [
23            {
24                "id": 1181030630,
25                "id_str": "1181030630",
26                "indices": [
27                    3,
28                    14
29                ],
30                "name": "Bruna",
31                "screen_name": "holdmenian"
32            }
33        ]
34    },
35    "favorite_count": 0,
```

```

36     "favorited": false ,
37     "filter_level": "medium",
38     "geo": null ,
39     "id": 347038057221980162,
40     "id_str": "347038057221980162",
41     "in_reply_to_screen_name": null ,
42     "in_reply_to_status_id": null ,
43     "in_reply_to_status_id_str": null ,
44     "in_reply_to_user_id": null ,
45     "in_reply_to_user_id_str": null ,
46     "lang": "pt",
47     "metadata": {
48         "client": "192.168.102.195",
49         "emoticons": [],
50         "hashtags": [],
51         "language": "pt",
52         "mentions": [
53             "@holdmenian"
54         ],
55         "tokenized": "RT @holdmenian : \" O comercial da fiat ' Vem pra rua ' saiu do ar ap\u00f3f3s virar m\u00f3fasica tema dos protestos \" mas http://t.co/P7nwF6GOFc",
56         "topic": "protestos_brasil",
57         "urls": [
58             "http://t.co/P7nwF6GOFc"
59         ]
60     },
61     "place": null ,
62     "possibly_sensitive": false ,
63     "retweet_count": 0,
64     "retweeted": false ,
65     "retweeted_status": {
66         "contributors": null ,
67         "coordinates": null ,
68         "created_at": "Tue Jun 18 13:09:39 +0000 2013",
69         "entities": {
70             "hashtags": [],
71             "symbols": [],
72             "urls": [
73                 {
74                     "display_url": "twitpic.com/cxvh38",
75                     "expanded_url": "http://twitpic.com/cxvh38",
76                     "indices": [
77                         88,
78                         110
79                     ],
80                     "url": "http://t.co/P7nwF6GOFc"
81                 }
82             ],
83         }
84     },
85 
```

```
83         "user_mentions": []
84     },
85     "favorite_count": 7,
86     "favorited": false,
87     "geo": null,
88     "id": 346978014636146688,
89     "id_str": "346978014636146688",
90     "in_reply_to_screen_name": null,
91     "in_reply_to_status_id": null,
92     "in_reply_to_status_id_str": null,
93     "in_reply_to_user_id": null,
94     "in_reply_to_user_id_str": null,
95     "lang": "pt",
96     "place": null,
97     "possibly_sensitive": false,
98     "retweet_count": 98,
99     "retweeted": false,
100    "source": "<a href=\"http://messaging.nokia.com/\" rel=\"nofollow\">
Social by Nokia</a>",
101   "text": "\"O comercial da fiat 'Vem pra rua' saiu do ar ap\u00f3s virar
m\u00e1fica tema dos protestos\" mas http://t.co/P7nwF6GOFc",
102   "truncated": false,
103   "user": {
104     "contributors_enabled": false,
105     "created_at": "Fri Feb 15 03:50:19 +0000 2013",
106     "default_profile": false,
107     "default_profile_image": false,
108     "description": "make a wish \u221e",
109     "favourites_count": 12,
110     "follow_request_sent": null,
111     "followers_count": 934,
112     "following": null,
113     "friends_count": 718,
114     "geo_enabled": false,
115     "id": 1181030630,
116     "id_str": "1181030630",
117     "is_translator": false,
118     "lang": "en",
119     "listed_count": 0,
120     "location": "Charlie \u2665",
121     "name": "Bruna",
122     "notifications": null,
123     "profile_background_color": "FFFFFF",
124     "profile_background_image_url": "http://a0.twimg.com/
profile_background_images/344918034409728850/6
c945a8006333f3847476148aa68d7a0.png",
125     "profile_background_image_url_https": "https://si0.twimg.com/
profile_background_images/344918034409728850/6
c945a8006333f3847476148aa68d7a0.png",
```

```

126     "profile_background_tile": false ,
127     "profile_banner_url": "https://pbs.twimg.com/profile_banners
128 /1181030630/1371263818",
129     "profile_image_url": "http://a0.twimg.com/profile_images
130 /344513261580011975/f105a548f1b2198d325864dc5313e06b_normal.png",
131     "profile_image_url_https": "https://si0.twimg.com/profile_images
132 /344513261580011975/f105a548f1b2198d325864dc5313e06b_normal.png",
133     "profile_link_color": "B40B43",
134     "profile_sidebar_border_color": "FFFFFF",
135     "profile_sidebar_fill_color": "DDEEF6",
136     "profile_text_color": "333333",
137     "profile_use_background_image": true ,
138     "protected": false ,
139     "screen_name": "holdmenian",
140     "statuses_count": 9291,
141     "time_zone": "Mid-Atlantic",
142     "url": null ,
143     "utc_offset": -7200,
144     "verified": false
145   },
146   "source": "web",
147   "text": "RT @holdmenian: \"O comercial da fiat 'Vem pra rua' saiu do ar ap\u00f3s virar m\u00f3veis\u00fasica tema dos protestos\" mas http://t.co/P7nwF6GOFc",
148   "truncated": false ,
149   "user": {
150     "contributors_enabled": false ,
151     "created_at": "Wed Mar 02 16:18:50 +0000 2011",
152     "default_profile": false ,
153     "default_profile_image": false ,
154     "description": "... o fim virou come\u00e7o. E eu me permiti come\u00e7ar.",
155     "favourites_count": 6,
156     "follow_request_sent": null ,
157     "followers_count": 405,
158     "following": null ,
159     "friends_count": 360,
160     "geo_enabled": true ,
161     "id": 259792830,
162     "id_str": "259792830",
163     "is_translator": false ,
164     "lang": "pt",
165     "listed_count": 4,
166     "location": "Pau dos ferros - RN",
167     "name": "Fernando Cassio",
168     "notifications": null ,
169     "profile_background_color": "759AAD",
170     "profile_background_image_url": "http://a0.twimg.com/
profile_background_images/396888075/bg-meio.jpg",

```

```
169     "profile_background_image_url_https": "https://si0.twimg.com/
170     profile_background_images/396888075/bg-meio.jpg",
171     "profile_background_tile": true,
172     "profile_banner_url": "https://pbs.twimg.com/profile_banners/
173     /259792830/1361109491",
174     "profile_image_url": "http://a0.twimg.com/profile_images/3687937981/
175     bd4525bc45aa159a1c8dabcb5eab3ef4_normal.jpeg",
176     "profile_image_url_https": "https://si0.twimg.com/profile_images/
177     /3687937981/bd4525bc45aa159a1c8dabcb5eab3ef4_normal.jpeg",
178     "profile_link_color": "888E94",
179     "profile_sidebar_border_color": "D1D3DE",
180     "profile_sidebar_fill_color": "E2EEF0",
181     "profile_text_color": "131314",
182     "profile_use_background_image": true,
183     "protected": false,
184     "screen_name": "fernandocassio_",
185     "statuses_count": 5296,
186     "time_zone": "Santiago",
187     "url": "http://www.facebook.com/fernando.cassio.948",
188     "utc_offset": -14400,
189     "verified": false
190   }
191 }
```


Referências

- [1] Max Bramer. *Principles of Data Mining*. 2007.
- [2] Peng Wu, YM Ro, CS Won, e Yanglim Choi. Texture descriptors in MPEG-7. *Comput. Anal. Images* . . , páginas 21–28, 2001.
- [3] M. Bober. MPEG-7 visual shape descriptors. *IEEE Trans. Circuits Syst. Video Technol.*, 11(6):716–719, Junho 2001. doi:10.1109/76.927426.
- [4] David G. Lowe. Distinctive Image Features from Scale-Invariant Keypoints. *Int. J. Comput. Vis.*, 60(2):91–110, Novembro 2004. URL: <http://link.springer.com/10.1023/B:VISI.0000029664.99615.94>, doi:10.1023/B:VISI.0000029664.99615.94.
- [5] Herbert Bay, Tinne Tuytelaars, e Luc Van Gool. Surf: Speeded up robust features. *Comput. Vision–ECCV 2006*, 2006. URL: http://link.springer.com/chapter/10.1007/11744023_32.
- [6] LM Bueno. Análise de descritores locais de imagens no contexto de detecção de semi-réplicas. 2011. URL: <http://www.dca.fee.unicamp.br/~dovalle/recod/works/lucasBueno2001mscDissertation.pdf>.
- [7] Jiawei Han, Micheline Kamber, e Jian Pei. *Data Mining, Second Edition: Concepts and Techniques*. 2006.
- [8] Google. Google chart - timeline. <https://developers.google.com/chart/interactive/docs/gallery/timeline>, 2014.
- [9] Matthew A Russell. *Mining the Social Web*, volume 54. 2011. doi:10.1081/E-ELIS3-120043522.
- [10] TDS Cunha. *TweeProfiles: detection of spatio-temporal patterns on Twitter*. Tese de doutoramento, Faculdade de Engenharia da Universidade do Porto, 2013. URL: http://paginas.fe.up.pt/~ei08142/files/mieic_en.pdf.
- [11] M Boanjak e Eduardo Oliveira. TwitterEcho: a distributed focused crawler to support open research with twitter data. *Proc. 21st* . . , 2012.
- [12] Dr. Matthew A North. *Data Mining for the Masses*. Global Text Project, 2012.
- [13] Usama Fayyad, Gregory Piatetsky-shapiro, e Padhraic Smyth. From Data Mining to Knowledge Discovery in Databases. páginas 37–54, 1996.
- [14] Bing Liu. *Web Data Mining: Exploring Hyperlinks, Contents, and Usage Data*. Springer, 2011.

- [15] Wei Wang, Jiong Yang, e Richard R. Muntz. STING: A Statistical Information Grid Approach to Spatial Data Mining. páginas 186–195, Agosto 1997.
- [16] Rakesh Agrawal, Johannes Gehrke, Dimitrios Gunopulos, e Prabhakar Raghavan. Automatic subspace clustering of high dimensional data for data mining applications. *ACM SIGMOD Rec.*, 27(2):94–105, Junho 1998. doi:10.1145/276305.276314.
- [17] J. Montavont e T. Noel. IEEE 802.11 Handovers Assisted by GPS Information. Em *IEEE Int. Conf. Wirel. Mob. Comput. Netw. Commun. 2006.*, páginas 166–172. IEEE, 2006. URL: <http://ieeexplore.ieee.org/lpdocs/epic03/wrapper.htm?arnumber=1696358>, doi:10.1109/WIMOB.2006.1696358.
- [18] David A. Forsyth e Jean Ponce. *Computer Vision: A Modern Approach*. Pearson Education, Limited, 2011.
- [19] Mark S. Nixon e Alberto S. Aguado. *Feature Extraction and Image Processing*. 2002.
- [20] BS Manjunath e JR Ohm. Color and texture descriptors. *Circuits Syst. . . .*, 11(6):703–715, 2001.
- [21] Charilaos Christopoulos, Daniel Berg, e Athanassios Skodras. The colour in the upcoming MPEG-7 standard. *Invit. Pap. Eur. . . .*, páginas 1–4, 2000.
- [22] Leszek Cieplinski. MPEG-7 Color Descriptors and Their Applications. 7:11–20, 2001.
- [23] Leszek Cieplinski (mitsubishi Electric Ite-vil. The MPEG-7 Color Descriptors Jens-Rainer Ohm (RWTH Aachen, Institute of Communications Engineering).
- [24] Vinay Modi. Color descriptors from compressed images. 2008.
- [25] H Shao, J Ji, Y Kang, e H Zhao. Application Research of Homogeneous Texture Descriptor in Content-Based Image Retrieval. *Inf. Eng. . . .*, (2008515):2–5, 2009.
- [26] Kristen Gauman e Bastian Leibe. *Visual Object Recognition*. Morgan & Claypool Publishers, 2010. URL: <http://books.google.com/books?id=fYZgAQAAQBAJ&pgis=1>.
- [27] D.G. Lowe. Object recognition from local scale-invariant features. *Proc. Seventh IEEE Int. Conf. Comput. Vis.*, páginas 1150–1157 vol.2, 1999. URL: <http://ieeexplore.ieee.org/lpdocs/epic03/wrapper.htm?arnumber=790410>, doi:10.1109/ICCV.1999.790410.
- [28] Josef Sivic e Andrew Zisserman. Video Google: A text retrieval approach to object matching in videos. *Comput. Vision, 2003. Proceedings. . . .*, (Iccv):2–9, 2003.
- [29] Josef Sivic e Andrew Zisserman. Video Google: Efficient visual search of videos. *Towar. Categ. Object Recognit.*, 4170:127–144, 2006. doi:10.1007/11957959_7.
- [30] Luo Juan e O Gwun. A comparison of sift, pca-sift and surf. *Int. J. Image Process.*, (4):143–152, 2009. URL: <http://www.cscjournals.org/csc/manuscript/Journals/IJIP/volume3/Issue4/IJIP-51.pdf>.
- [31] R Baeza-Yates e B Ribeiro-Neto. *Modern information retrieval*. 1999. URL: ftp://mail.im.tku.edu.tw/seke/slides/baeza-yates/chap10_user_interfaces_and_visualization-modern_ir.pdf.

- [32] D. Nister e H. Stewenius. Scalable recognition with a vocabulary tree. ... *Vis. Pattern Recognition, 2006...*, 2, 2006. URL: http://ieeexplore.ieee.org/xpls/abs_all.jsp?arnumber=1641018.
- [33] Jan Erik Solem. Programming Computer Vision with Python. 2012.
- [34] A. Vedaldi e B. Fulkerson. VLFeat: An open and portable library of computer vision algorithms. <http://www.vlfeat.org/>, 2008.
- [35] Google. Api javascript do google maps v3. <https://developers.google.com/maps/documentation/javascript/>, 2013.