TÉCNICO LISBOA

Atribuição Automática de Coordenadas Geoespaciais a Fotos Publicadas no Flickr

Alfredo Aniceto Andrade Tavares

Dissertação para obtenção do Grau de Mestre em

Engenharia Informática e de Computadores

Orientadores: Prof. Bruno Emanuel da Graça Martins e Prof. Alfredo Manuel dos Santos Ferreira Júnior

Júri

Presidente:	Prof. José Luís Brinquete Borbinha
Orientador:	Prof. Bruno Emanuel da Graça Martins
Vogal:	Prof. Manuel João Caneira Monteiro da Fonseca

Maio de 2015

Aos meus pais

Agradecimentos

A realização desta dissertação marca o final de uma etapa importante da minha vida. É com muito entusiasmo que expresso aqui o mais profundo agradecimento a todos aqueles que contribuíram para sua concretização.

Antes de mais, um especial agradecimento aos meus pais, pela total confiança e apoio oferecido a esta e todas as demais caminhadas necessárias à conclusão deste trabalho. Sem vocês isso não seria possível. Agradeço também aos meus irmãos pelo apoio e incentivo recebido ao longo destes anos.

Gostaria de agradecer também, ao meu orientador Professor Bruno Martins, e ao meu co-orientador, Alfredo Ferreira, pelo apoio, sugestões, críticas e disponibilidade, os quais foram fundamentais para o enriquecimento deste trabalho. Além disso, agradeço também as suas ajudas nas questões que fogem da própria área de orientação, desde o início do projecto de mestrado até a fase final.

Por fim, um agradecimento a todos os amigos e familiares que contribuíram de qualquer forma para a concretização desta dissertação, e que sempre acreditaram e me apoiaram, apesar das dificuldades encontradas.

Resumo

A maioria dos recursos multimédia produzidos no contexto das mais diversas aplicações encontra-se relacionado com algum tipo de contexto geográfico. Contudo, os métodos tradicionais de recuperação de informação multimédia simplesmente vêem esses recursos como um conjunto de termos ou de características visuais de baixo nível, ignorando outros aspectos. A disponibilização de recursos de informação multimédia georreferenciados pode ser útil no contexto de várias aplicações. Neste trabalho, propomos e avaliamos uma técnica simples para a atribuição automática de coordenadas geoespaciais de latitude e de longitude a novas fotografias, usando as anotações (i.e., *tags*) como a maior fonte de evidência. A técnica proposta atribui pesos para as diferentes *tags* de acordo com a frequência inversa da *tag* na colecção de fotografias, ou de acordo com a área geométrica da região que abrange um conjunto de fotografias. Em seguida, são atribuídas coordenadas geoespaciais a uma nova foto através de uma interpolação das coordenadas associadas a todas as fotografias de treino que partilham *tags* com a mesma, em que os pesos correspondem aos valores associados a cada *tag*. Experimentamos também com o uso de descritores visuais, usando-os para pesar as contribuições das diferentes tags de acordo com a similaridade das imagens. Avaliamos o método proposto usando uma colecção de fotografias georreferenciadas, recolhidas a partir do *Flickr* e disponibilizada no contexto do evento *MediaEval* 2013. A melhor atribuição de coordenadas geoespaciais foi alcançada com a configuração correspondente ao uso do método *convex-hull* para a obtenção de uma região geográfica que engloba um conjunto de fotos, obtendo um erro médio de 2348 km, e de um erro mediano de 558 km.

Palavras-chave: Georreferenciação de Imagens, Recuperação de Informação Geográfica, Indexação e Recuperação de Informação Multimédia, Anotação de Fotografias.

Abstract

Most multimedia resources can be said to be related to some particular geographic context, although traditional multimedia retrieval methods simply model these resources as bags of textual and/or low-level visual features, ignoring other aspects of the encoded information. Nonetheless, geospatial metadata associated to multimedia resources can be useful in the context of many different applications. In this work, we empirically evaluate a relatively simple technique for assigning geospatial coordinates of latitude and longitude to previously unseen photographs, using the associated textual tags as the main input evidence. The proposed technique assigns weights to the different tags according to the inverse frequency of the tag in the collection of photos, or according to the area of an encompassing geometric shape that covers the locations of all associated training photos, and it then computes a weighted geographic midpoint from all the locations of training photos that share tags with the photos that are to be geocoded. We also experimented with the usage of visual content descriptors, using them to weight the contribution of the different tags according to image similarity. We evaluate the proposed methods through a large collection of geo-referenced photos, gathered from Flickr and made available in the context of the 2013 MediaEval placing task. The best performing configuration uses weights based on a convex-hull to find an encompassing region for tagged photos, and it achieves an average prediction error of 2348 kilometers, and a median prediction error of 558 kilometers.

Keywords: Geocoding Images, Geographic Information Retrieval, Content-Based Multimedia Indexing and Retrieval, Photo Annotation.

Conteúdo

Lista de Tabelas

Lista de Figuras

Glossário

CBIR Recuperação de Imagens com base em Características Visuais é uma técnica de recuperar/procurar imagens digitais a partir de uma grande base de dados que contêm características visuais similares à imagem de interrogação.

GPS Sistema de Posicionamento Global é um sistema de posicionamento geográfico que fornece a um aparelho receptor móvel a sua posição, assim como informação horária, sobre todas condições atmosféricas, de um lugar na Terra.

TBIR Recuperação de Imagens com base em Características Textuais é uma técnica de recuperar/procurar imagens digitais a partir de uma grande base de dados que contêm características textuais similares à imagem de interrogação.

TF-IDF Term Frequency–Inverse Document Frequency é uma abordagem para quantificar a importância de um termo/palavra num documento ou em uma colecção de documentos.

Capítulo 1

Introdução

A quantidade de informação multimédia georreferenciada disponível na Web, e redes sociais tais como o *Flickr*[1], *Facebook*[2] ou o *Instagram*[3] tem vindo a crescer devido aos avanços tecnológicos em termos de dispositivos móveis com receptores *GPS* integrados. No entanto, a maioria das imagens disponíveis nestes serviços não estão ainda georreferenciadas, e isto implica a necessidade de métodos para a georreferenciação automática destas imagens. Por outro lado, a quantidade de informação georreferenciada está a proporcionar novas possibilidades no contexto do desenvolvimento de métodos automáticos para a descoberta e organização dessa informação. Temos, por exemplo, que colecções de fotos extraídas do *Flickr*, em conjunto com os seus meta-dados (i.e., anotações textuais, descritores visuais, e coordenadas geográficas), têm sido amplamente exploradas no contexto de sistemas para a Recuperação de Informação Geográfica (*RIG*).

Nesta dissertação, propomos e avaliamos uma nova técnica para a atribuição automática de coordenadas geoespaciais de latitude e de longitude a novas fotografias, usando as *tags* (i.e., anotações textuais) como a principal fonte de evidência. A técnica proposta atribui pesos para as diferentes *tags* de acordo com a frequência inversa da *tag* na colecção de fotografias, ou de acordo com a área geométrica de uma localização que abrange um conjunto de fotografias. Neste trabalho, esta área pode ser delimitada por uma *bounding box* (i.e., uma menor região rectangular contendo os pontos correspondentes às localizações de fotos), um polígono convexo (i.e., *convex-hull*, ou seja um menor polígono convexo contendo os pontos), ou um polígono côncavo (i.e., *concave-hull*, ou seja um menor poligno côncavo contendo os pontos). A atribuição das coordenadas geoespaciais a uma nova foto é efectuada usando uma interpolação das coordenadas associadas a todas as fotografias de treino que partilham *tags* com a mesma, usando como pesos os valores associados às *tags*.

[1] `http://flickr.com/`
[2] `http://facebook.com/`
[3] `https://instagram.com/`

1.1 Hipótese e Metodologia

Partindo do princípio da existência de grandes colecções de fotografias já georreferencias, podemos estimar as coordenadas geográficas de uma outra fotografia através da exploração da similaridade dos meta-dados da mesma para com os de uma colecção de fotos já georreferenciadas. Por exemplo, para estimar as coordenadas geográficas de uma dada fotografia, todas as coordenadas geográficas das *knn* fotografias mais similares existentes na colecção podem ser consideradas.

Para validar a abordagem proposta, usamos uma colecção de fotografias disponibilizada no contexto do evento *MediaEval* 2013 [Hauff et al., 2013]. *MediaEval* é uma iniciativa de benchmarking dedicada a avaliar novos algoritmos para o acesso e a recuperação de informação multimédia[4]. Nesta colecção, estão incluídas cerca de 9 milhões de fotografias publicadas no Flickr, das quais $8,5$ milhões estão associadas a coordenadas de latitude e longitude. Para além disso, esta colecção contém meta-dados descritivos, entre os quais se incluem características derivadas de anotações textuais (i.e., *tags*) e de descritores para os conteúdos visuais das fotos. Todas as fotografias foram anotadas com a mais alta precisão do *Flickr* (i.e., no *Flickr*, existem diferentes níveis de precisão, sendo o maior é de 16, o que significa que o local atribuído a fotografia é preciso ao nível da rua).

Para a validação dos resultados foi usada a média e a mediana das distâncias geoespaciais entre as coordenadas retornadas pelo método proposto e as coordenadas indicadas na colecção. As distâncias geoespaciais foram calculadas através das fórmulas de Vincenty[5]. Por fim, um valor indicativo da precisão foi determinado através do número de atribuições correctas de coordenadas, considerando um erro na distância abaixo de 1, 10 ou 100 km.

1.2 Contribuições

As principais contribuições do presente trabalho são as seguintes:

- Implementação de uma técnica eficiente para a atribuição das coordenadas geoespaciais de latitude e longitude a fotos, através da interpolação das coordenadas associadas às fotografias que partilhem pelo menos uma *tag*.

- Avaliação comparativa de diferentes técnicas para quantificar a importância das *tags*, incluindo uma técnica baseada no uso dos conteúdos visuais das imagens. Os resultados experimentais mostraram que a configuração correspondente ao uso de técnica baseada em calcular o *convex-hull* dos pontos correspondentes à localização das fotos associadas à *tag*, obteve melhores resultados.

[4] `http://www.multimediaeval.org/`
[5] `http://en.wikipedia.org/wiki/Vincenty's_formulae/`

1.3 Organização do Documento

Os restantes conteúdos desta dissertação estão organizados da seguinte forma: O Capítulo 2 apresenta os conceitos fundamentais necessários à compreensão do trabalho desenvolvido. O Capítulo 3 descreve os principais trabalhos anteriores relacionados com a georreferenciação de fotos. No Capítulo 4 detalhamos o trabalho desenvolvido, descrevendo pormenorizadamente a arquitectura do sistema proposto para a georreferenciação de fotos. No Capítulo 5 apresentamos a validação experimental e os resultados obtidos. Finalmente, no Capítulo 6 apresentamos as principais conclusões deste trabalho e as possíveis direcções para trabalho futuro.

Capítulo 2

Conceitos

Este capítulo apresenta os conceitos fundamentais para a compreensão do trabalho desenvolvido. Inicialmente, são abordados os principais conceitos na área de processamento de imagem. Em seguida, são apresentadas as características usadas para representação dos conteúdos das fotografias. Posteriormente, são descritas funções para o cálculo da similaridade entre *vectores de características*, populares na área de recuperação de imagens com base no conteúdo. Finalmente, é descrita uma estrutura de indexação para a pesquisa eficiente de documentos semelhantes, mais especificamente o índice invertido.

2.1 Representação de Imagens

Uma imagem pode ser definida como uma representação visual de um ou vários objectos. A necessidade de uma ferramenta para a manipulação das imagens, no sentido de melhorar as suas características visuais, é fundamental na área de processamento de imagem. Nesta área, as imagens são geralmente representadas como um conjunto de pontos finitos definidos por valores numéricos, denominando esta representação como imagem digital.

Uma imagem digital pode ser considerada como uma função bidimensional $f(x, y)$ da intensidade luminosa, em que o valor ou a amplitude de f nas coordenadas espaciais (x, y) correspondem a intensidade (brilho) ou nível de cinza da imagem naquele ponto [Gonzalez and Woods, 2001]. Em outras palavras, uma imagem digital pode ser representada através de uma matriz $n \times m$, onde cada elemento é identificado pelos índices da linha e da coluna, e o seu valor corresponde à intensidade ou nível de cinza $f(x, y)$ em um determinado ponto na imagem. Estes elementos desta matriz são denominados de pixéis, abreviado do inglês *picture elements* [Gonzalez and Woods, 2001]. Assim sendo, uma imagem digital é constituída por um número finito de elementos, na qual cada um possui uma posição e um valor (escalar ou vectorial), que representa uma determinada cor em um determinado ponto na imagem. Para imagens binárias (i.e., em preto e branco) os pixéis podem assumir valores 0 e 1. Para imagens em níveis de cinza, estes valores podem variar entre 0 a 255, considerando um byte por pixel. Já no caso de imagens coloridas, o valor de cada pixel pode ser representado por três componentes formando um

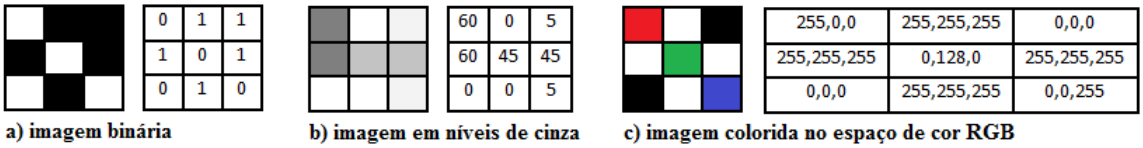

Figura 2.1: Exemplo de imagens e suas respectivas representações digitais.

espaço de cor. A Figura 2.1 apresenta um exemplo de três imagens constituídas por 3×3 pixéis e suas matrizes de representação digital correspondentes.

2.1.1 Propriedades de Imagens

Existem várias propriedades que podem ser extraídas a partir de uma imagem e elas podem representar características gerais ou de um domínio específico [Gudivada and Raghavan, 1995]. As características gerais são aquelas que podemos encontrar facilmente em qualquer imagem, tais como cor, textura, forma e, suas combinações. Já as características de um domínio específico são aquelas que dependem da aplicação, por exemplo, as características referentes a impressões digitais e faces humanas. Em seguida, serão abordadas algumas características de domínio geral, mais especificamente cor e textura.

Cor

A cor é considerada um conceito subjectivo do ser humano, uma vez que consiste na interpretação do sistema nervoso aos diferentes comprimentos de onda da radiação luminosa recebida [Sharma, 2002]. Numa imagem, a cor de cada pixel corresponde a um ponto em um sistema de coordenadas 3D, denominado de espaço de cor. Em seguida, são apresentados os espaços de cor incluídos nos descritores visuais usados nas experiências desta dissertação.

- **Espaço de cor RGB**

  O espaço de cor *RGB* (*Red*, *Green* e *Blue*) é o mais conhecido e o mais utilizado. Este espaço de cor é orientado ao *hardware*, e é conhecido como um cubo *RGB*, tendo como base as cores primárias vermelho (*R*), verde (*G*) e azul (*B*) nos vértices, conforme ilustrado na Figura 2.2. Estas cores base são combinadas de várias formas para reproduzir uma variedade de cores. Cada uma das três cores primárias é chamada de um componente da cor resultante, e cada uma delas pode ter um determinado valor de intensidade, resultando noutras cores. O valor da intensidade pode variar entre 0 e 1. O valor 1 corresponde à intensidade máxima que uma cor pode ser pode ser representada, e o valor 0 à intensidade mínima. Geralmente, as implementações do espaço de cor *RGB* nos sistemas gráficos utilizam valores inteiros entre 0 e 255 (i.e., um byte) para representar o valor da intensidade de cada componente de cor, em vez dos valores reais normalizados entre 0 e 1.

- **Espaço de cor HSV**

  O espaço de cor *HSV* (*Hue*, *Saturation* e *Value*) é um espaço mais intuitivo comparado com

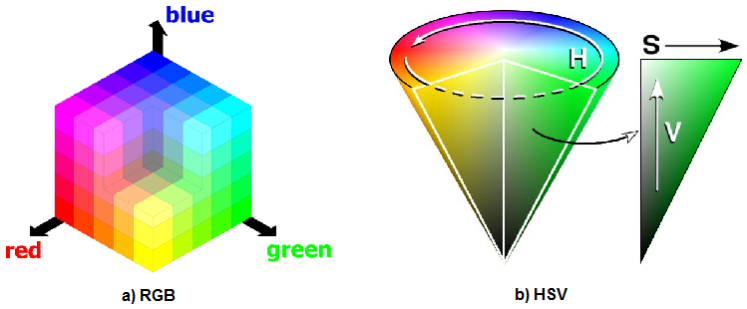

Figura 2.2: Espaços de cor RGB e HSV, Figuras adaptadas de figuras originais em Gonzalez and Woods [2001].

espaço de cor *RGB*. O *HSV* é baseado nas percepções humanas, i.e., explora as características utilizadas por humanos para distinguir uma cor da outra, através dos seus três componentes. O primeiro componente é o *Hue*, que nos permite distinguir o tipo de cor. O segundo é a *Saturation*, que indica o grau de pureza de uma cor, i.e., a quantidade de cinza que uma cor contém. Por último, o componente *Value* indica o brilho da cor, representando o quão clara ou escura uma cor é, em relação a sua cor padrão. A sua representação gráfica é apresentada na Figura 2.2.

- **Espaços de cor YIQ e YCbCr**

  Os espaços de cor *YIQ* e *YCbCr* foram desenvolvidos para permitir que as emissões dos sistemas de televisão a cores fossem compatíveis com os receptores a preto e branco. Os dois espaços são semelhantes e ambos baseiam-se na separação das componentes do espaço de cor *RGB* em uma componente de luminância, e duas componentes de crominância ou diferença de cor. Desta forma, o sinal de televisão a cores correspondente a luminância é transmitido da mesma forma que o sinal de televisão a preto e branco e, assim, os receptores a preto e branco podem receber as emissões da televisão a cores [Sproson, 1983]. O primeiro componente é o *Y*, que corresponde à luminância, i.e., à quantidade das cores pretas e brancas presentes. O seu valor é obtido de forma idêntica em ambos os espaços, e consiste na ponderação dos valores das componentes *RGB* de uma determinada cor, sendo definida por:

$$Y = 0,299R + 0,587G + 0,114B$$

  Assim sendo, a diferença entre os dois espaços de cor reside na definição das outras componentes, i.e., as de crominância. Estas componentes contêm informação de cor propriamente dita, e são combinadas de várias formas para reproduzir diversas cores. No espaço de cor *YIQ*, a componente em fase (*I*) e em quadrado (*Q*), são calculadas por diferenças ponderadas entre as componentes vermelha e azul da cor no espaço *RGB*, tal como apresentado de seguida:

$$I = 0,74(R - Y) - (B - Y), \quad Q = 0,48(R - Y) + 0,41(B - Y)$$

  Já no espaço de cor *YCbCr*, as componentes de crominância são definidos como diferença entre

a crominância de azul (i) e de vermelho (*Cr*) no espaço de cor *RGB*, definida por:

$$Cb = B - Y, \;\; Cr = R - Y$$

- **Espaço de cor oponente**

  O espaço de cor oponente (*OPP*) é baseado na teoria de processo oponente descrito no trabalho de Ewald Hering et al. [1964]. Com base nos resultados de misturas de cores, Ewald Hering et al. argumentou que a cor amarela deve ser considerada como uma cor primária, para além da cor vermelho, verde e azul. Ainda mais, que as combinações das cores verde e vermelho, como também a de azul e amarelo não produzem nenhuma cor, i.e., não somos capazes de ver a cor verde avermelhado ou a cor azul amarelado. Por conseguinte, ele afirmou a existência de 3 canais de cores opostos nomeadamente, um canal de verde até o vermelho (O_1), um de azul até amarelo (O_2), e mais um outro de preto até branco (O_3). Os dois primeiros canais representam a crominância, e o último a luminância. Nesta dissertação, foi usado o espaço *OPP* obtido pela seguinte transformação linear do espaço de cor *RGB* [van de Sande et al., 2010]:

$$\begin{pmatrix} O_1 \\ O_2 \\ O_3 \end{pmatrix} = \begin{pmatrix} \frac{R-G}{\sqrt{2}} \\ \frac{R+G-2B}{\sqrt{6}} \\ \frac{R+G+B}{\sqrt{3}} \end{pmatrix}$$

Textura

O conceito de textura refere-se ao aspecto de uma superfície, ou seja, a *pele* de uma forma, que permite identificá-la e distingui-la de outras superfícies. Quando tocamos ou olhamos para um objecto ou superfície, sentimos se a sua *pele* é lisa, rugosa, macia, áspera ou ondulada. Por isso, a textura é uma sensação visual ou táctil. Embora não exista uma definição exacta para o conceito de textura de uma imagem, Tuceryan and Jain [1998] definiram-na como mudanças na intensidade da imagem que formam determinados padrões repetitivos. Estes padrões podem ser estabelecidos pela diferença na superfície da estrutura, ou pela diferença de cor. Sendo assim, as texturas são aspectos presentes nas imagens, caracterizados pela relação entre pixéis, ao contrário da cor que é uma propriedade individual de cada pixel.

Existem três abordagens principais usadas em classificação de imagens para a descrição de texturas, nomeadamente a espectral, a estrutural, e a estatística [Gonzalez and Woods, 2001]. A abordagem espectral caracteriza as texturas como propriedades baseadas na *transformada de Fourier*, onde pode ser detectada a existência de padrões periódicos ou semi-periódicos. A abordagem estrutural representa uma textura como sendo formada pela repetição de padrões (i.e., relação espacial entre os pixéis de imagens) que obedeçam a alguma regra de posicionamento para a sua geração. Já na abordagem estatística, são utilizadas medidas estatísticas das características visuais de uma imagem para descrever as texturas. Esta última abordagem é geralmente recomendada para as texturas que não apresentem um elevado grau de regularidade.

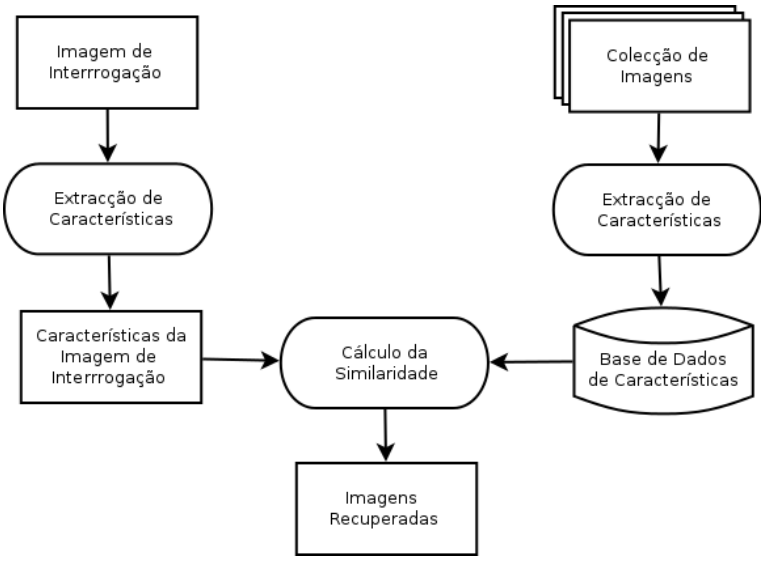

Figura 2.3: Arquitectura geral de um sistema de recuperação de imagem.

2.2 Recuperação de Imagens

A recuperação de imagem é a tarefa de recuperar/procurar imagens digitais a partir de uma grande base de dados. O processo base dos sistemas de recuperação de imagem é a extracção de características, o qual tem como objectivo obter os atributos de baixo nível de uma imagem. Estas características são normalmente os conteúdos textuais (e.g., data de captura, anotações ou palavras chave) e visuais (e.g., cor ou textura). Uma vez obtidas estas características, a tarefa de recuperar imagens consiste basicamente em, dada uma imagem de interrogação, calcular a sua similaridade para com as imagens armazenadas em uma base de dados, exibindo as mais similares por ordem de relevância. A similaridade é obtida comparando os conteúdos da imagem de interrogação e os das imagens na base de dados. Assim sendo, os sistemas de recuperação de imagem diferem no tipo de características utilizados. A Figura 2.3 apresenta de forma simples e genérica a arquitectura de um sistema de recuperação de imagem.

2.2.1 Conteúdos Textuais

Os conteúdos de uma imagem podem ser descritos através de anotações textuais (i.e., *tags*), as quais fornecem informações importantes para os sistemas de recuperação de imagens com base em características textuais similares. Estes sistemas são denominados de *Text-Based Image Retrieval* (*TBIR*) [Chang et al., 1988]. No entanto, as anotações textuais associadas às imagens são muitas vezes subjectivas, uma vez que os conteúdos visuais das imagens podem ser interpretados de diversas formas de acordo com a percepção de cada pessoa. De modo a resolver este problema, é necessário determinar quais os termos presentes nas anotações textuais que fornecem informações importantes para a descrição dos conteúdos visuais das imagens. Neste contexto, a área de Recuperação de Informação (*RI*) desenvolveu diversas abordagens para quantificar a importância de um termo/palavra num documento ou em uma colecção de documentos [Manning et al., 2008]. Algumas das mais impor-

tantes são as seguintes:

- *Term frequency* ($tf_{t,d}$): A importância de um termo t para um documento d é definida como o número de ocorrências de t em d. Quanto maior for a ocorrência de um determinado termo num documento, maior é a sua relevância para descrever o documento. Geralmente, esta abordagem é definida de forma normalizada, tal como ilustrada na equação seguinte, onde $freq_{t,d}$ corresponde ao número de ocorrências de *t* em *d*, e o parâmetro D_d indica número total de termos em d:

$$tf_{t,d} = \frac{freq_{t,d}}{D_d}$$

- *Inverse document frequency* (idf): Esta abordagem mede o quão infrequente é um termo em toda a colecção de documentos. De acordo com esta abordagem, os termos que aparecem com muita frequência em toda a colecção de documentos são considerados os menos discriminativos (e.g., *stop-words*). Assim sendo, a importância de um termo t é inversamente proporcional à frequência de t em uma colecção de documentos D, e é definida de acordo com a seguinte equação, onde df_t corresponde ao número de documentos pertencentes a D, em que t ocorre:

$$idf_t = \log \frac{\#D}{df_t}$$

- *Term frequency - inverse document frequency* ($tf - idf$): A importância de um termo t para um documento d é obtida através de uma combinação linear das abordagens anteriores, dada pela seguinte equação:

$$tf - idf_{t,d} = tf_{t,d} \times idf_t$$

A área de *RI* também desenvolveu diversos modelos para a representação de grandes colecções de documentos textuais. Neste trabalho, a nossa representação foi baseada no modelo ”saco de palavras”, do inglês, *bag-of-words* (*BOW*), i.e., cada imagem foi representada como um conjunto de pares na forma de (*tag*, peso). Este modelo ignora não só a ordem dos termos, como também as suas relações semânticas. Contudo, este modelo tem um custo de processamento baixo e, é facilmente aplicável a textos provenientes de diferentes fontes. Normalmente, neste modelo, um documento d_j é representado como um vector de pesos de termos, em um espaço de dimensionalidade m, correspondente ao número total de termos distintos presentes na colecção ou ao tamanho de um dicionário. Cada uma das dimensões deste vector, $w_{i,j}$, corresponde ao valor da importância do termo t_i no documento d_j. Temos então que:

$$d_j = w_{1,j}, w_{2,j}, \ldots, w_{m,j}$$

Já na representação de uma colecção de documentos, é usada uma matriz denominada de termo-documento M. Nesta matriz, cada linha i representa um vector de termo e assume a forma $t_i = w_{i,1}, w_{i,2}, \ldots, w_{i,n}$, enquanto que cada coluna j representa um vector de documento $d_j = w_{1,j}, w_{2,j}, \ldots, w_{m,j}$. O parâmetro n é o número total de documentos e o parâmetro m corresponde ao número total de termos distintos da colecção. O valor $w_{i,j}$ para cada posição (i,j), reflecte o valor da importância do termo

t_i no documento d_j, tal como ilustrado de seguida:

$$M = \begin{bmatrix} w_{1,1} & \dots & w_{1,j} & \dots & w_{1,n} \\ \vdots & \ddots & & & \vdots \\ w_{i,1} & \dots & w_{i,j} & \dots & w_{i,n} \\ \vdots & & & \ddots & \vdots \\ w_{m,1} & \dots & w_{m,j} & \dots & w_{m,n} \end{bmatrix}$$

Como podemos observar, uma matriz termo-documento permite um desenvolvimento de soluções simples e rápidas. Esta abordagem é amplamente utilizada em soluções para a indexação e pesquisa por termos de documentos, sendo que um exemplo de uma implementação ainda mais eficiente seria o índice invertido, descrito na Secção 2.4.1.

2.2.2 Conteúdos Visuais

Os sistemas de recuperação de imagens com base em características visuais similares são conhecidos como Sistemas de Recuperação de Imagens Baseados em Conteúdo (*CBIR*). Nos sistemas *CBIR*, as imagens são normalmente representadas através de *vectores de características*, uma vez que esta representação facilita o armazenamento e o processamento das mesmas. Temos que um *vector de características* é uma representação numérica sucinta de uma imagem, através de um vector de dimensionalidade n. Diversos mecanismos diferentes (i.e., diversos descritores dos conteúdos) podem ser usados na composição dos *vectores de características*, sendo que um exemplo seria um *histograma de cor*. Um *histograma de cor* é formado por contentores (*bins*), um para cada cor de um espaço quantizado de todas as cores existentes, que somam o número de ocorrências ou a percentagem de cada cor, numa zona da imagem correspondente (i.e., no caso de histogramas de cor locais) ou na totalidade da imagem (i.e., no caso de histogramas de cor globais).

Através dos descritores de imagens é possível obter quantificações objectivas das características visuais do conteúdo das mesmas, representando-se assim alguns atributos visuais elementares das imagens, tais como cor e textura, em vectores sucintos de características. Para as experiências deste trabalho, usamos o descritor visual *Simple Color Histogram*, o qual também foi disponibilizado no âmbito do *MediaEval* 2013. Este descriptor é um histograma de cor no espaço *RGB* quantizado em 512 cores, onde cada uma corresponde a um bin no histograma. Desta forma, o valor armazenado em cada bin corresponde ao número de pixéis da imagem com uma destas cores.

2.3 Medição de Similaridade Entre Vectores de Características

Uma função de similaridade $\mathrm{sim}(X, Y)$ compara dois *vectores de características* e retorna um valor $\in [0, 1]$, que expressa a similaridade entre os dois vectores. Quanto maior for esse valor, mais similares são os vectores. Por outro lado, uma função de distância $\mathrm{dist}(X, Y)$ compara dois *vectores de características* e retorna um valor $\in [0, +\infty]$, que expressa a distância entre os dois vectores. Quanto menor

for esse valor, mais similares são os vectores.

Uma função de similaridade pode ser definida como uma inversa de uma dada função de distância, enquanto que uma função de distância pode ser transformada numa função de similaridade, tal como ilustrado de seguida:

$$\mathsf{sim}(X,Y) = \frac{1}{1+\mathsf{dist}(X,Y)} \quad \text{ou} \quad \mathsf{sim}(X,Y) = e^{-\mathsf{dist}(X,Y)} \tag{2.1}$$

Toda função de distância $\mathsf{dist}(X,Y)$, num espaço métrico, deve obedecer às seguintes propriedades:

1. Não negatividade, isto é $\rightarrow \mathsf{dist}(X,Y) \geq 0$;
2. Identidade, isto é $\rightarrow \mathsf{dist}(X,Y) = 0 \Leftrightarrow X = Y$;
3. Simetria, isto é $\rightarrow \mathsf{dist}(X,Y) = \mathsf{dist}(Y,X)$;
4. Desigualdade triangular $\rightarrow \mathsf{dist}(X,Z) + \mathsf{dist}(Z,Y) \geq \mathsf{dist}(X,Y)$;

Em cada um dos pontos anteriores, X e Y representam dois vectores de dimensionalidade n. Qualquer função de distância é chamada de métrica se satisfizer todas estas propriedades. No entanto, uma função pode não satisfazer a Propriedade 4 e ser usada como forma de comparar *vectores de caracaterísticas*. A escolha da função de distância adequada para o domínio de aplicação é importante no desenvolvimento de ferramentas com vista à recuperação de imagens.

As funções de distância mais usadas em espaços vectoriais são as funções da família L_p ou *Minkowski*, que podem ser definidas genericamente como:

$$\mathsf{dist}(X,Y) = \sqrt[p]{\sum_{i=1}^{n} |X_i - Y_i|^p}$$

Na fórmula acima, X e Y são dois *vectores de características* de dimensão n, onde X_i representa o elemento na posição i do vector X, onde Y_i representa o elemento na posição i no vector Y, e onde $i \in [1,n]$. Diferentes valores do parâmetro p levam as diferentes interpretações, sendo que as mais populares são a *distância* L_2 (i.e., a *distância Euclidiana*), a *distância* L_1 (i.e., a *distância de Manhattan* ou *City Block*), a *distância* $L_{0,5}$ (i.e., um exemplo de uma *fractional* L_p *distance*), e a *distância* L_∞ (i.e., de *Chebychev*).

A distância *Euclidiana* é uma métrica que deriva da medida de *Minkowski* quando p for igual a 2, e é definida genericamente como:

$$\mathsf{dist}(X,Y) = \sqrt{\sum_{i=1}^{n} |X_i - Y_i|^2}$$

Esta função corresponde ao somatório das diferenças dos quadrados entre as coordenadas. Seu raio de abrangência é maior, quando comparada com a distância L_1, e a medida pode ser interpretada atraves de formação de uma circunferência centrada na coordenada de referência.

A *distância Euclidiana* aplica-se melhor a dados não padronizados, i.e., dados que não têm nenhum tipo de tratamento de adaptação de escala, dado que o resultado final é insensível a *valores extremos* (e.g., exceções, ou dados com uma diferença muito grande em relação à média do conjunto de dados). Contudo, as distâncias podem ser fortemente afectadas por diferenças de escala entre as dimensões do vector sob o qual as distâncias são calculadas.

Muitas vezes, com o objectivo de optimizar o tempo de cálculo, a distância L_2 é utilizada na sua forma quadrática (i.e, sem a extracção da raiz quadrada), sendo desta forma denominada de *distância Euclidiana quadrada*. Essa modificação preserva a distância relativa entre pontos.

A distância de *Manhattan* é uma métrica que deriva da medida de *Minkowski* quando p for igual a 1, e é definida genericamente como:

$$\text{dist}(X, Y) = \sum_{i=1}^{n} |X_i - Y_i|$$

Esta função corresponde ao somatório do módulo das diferenças entre as coordenadas. O seu raio de abrangência, formado pelos pontos equidistantes a um dado ponto central, forma um losango com as diagonais paralelas aos eixos das coordenadas. Nesta métrica, os *valores extremos* não são considerados, e portanto não há influência da escala sobre o resultado, já que não há o cálculo da raiz quadrada. Temos também que a distância L_1 é menos dispendiosa para calcular do que a L_2, uma vez que não é necessário calcular a raiz quadrada, o que acelera sensivelmente o tempo de computação.

A *distância de Chebyshev* é uma métrica que deriva da medida de *Minkowski* quando p tende para mais infinito, e é definida genericamente como:

$$\text{dist}(X, Y) = \max_{i=1}^{n} |X_i - Y_i|$$

Como podemos verificar, esta função corresponde ao maior dos módulos da diferença entre os elementos de dois vectores com o mesmo índice. O seu raio de abrangência, formado pelos pontos equidistantes a um dado ponto central, forma um quadrado com os lados paralelos aos eixos das coordenadas, e é maior quando comparado com as distâncias L_1 e L_2.

A *distância de Chebyshev* é suscetivel a variações de escala, mas pode ser apropriada em casos em que a velocidade de execução é crítica.

As distâncias da família *fractional* L_p *distances*, das quais a $L_{0,5}$ é um exemplo, derivam da medida de *Minkowski* quando p for igual a um valor fraccionário, tal como $0,5$. No caso da $L_{0,5}$, a métrica é definida genericamente como:

$$\text{dist}(X, Y) = \left(\sum_{i=1}^{n} |X_i - Y_i| \right)^2$$

Esta funcao não é considerada uma métrica porque viola a *desigualdade triangular*. Mesmo assim, trabalhos recentes têm sugerido a utilização de dissimilaridades fracionais (i.e., $0 < p < 1$), porque estas medidas podem ser usadas na procura dos K vizinhos mais próximos com bons resultados. Howarth

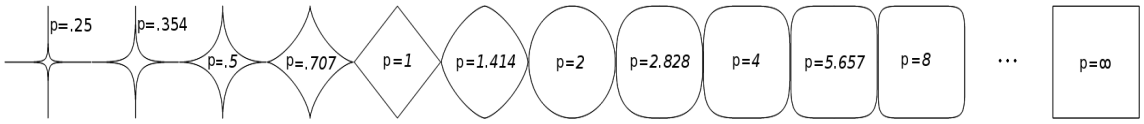

Figura 2.4: Representação das formas geométricas definidas por diferentes valores de p.

and Ruger [2005] verificaram que o desempenho em recuperação de imagens seria aumentado em muitas circunstâncias, quando $p = 0,5$.

Deve ser observado que numa *distância* L_p, à medida que o valor de p aumenta, a quantidade de operações a serem realizadas também aumenta, o que leva a um aumento do custo computacional. Temos ainda que a sensibilidade das métricas vai ser maior a distâncias maiores, com o aumento de p. A Figura 2.4 apresenta as diferentes formas geométricas das distâncias de família L_p definidas pelos diferentes valores de p, ilustrando que quanto maior for o valor de p, maior será o raio de abrangência correspondente.

Existem ainda varias outras medidas de distância e de similaridade, muito usadas em recuperação de imagens, e não baseadas nas distâncias L_p.

Um exemplo é a medida de *similaridade dos cossenos*. Dados dois vectores, X e Y, a *similaridade dos cossenos* é representada usando um produto escalar e a magnitude dos vectores, i.e., a similaridade é calculada através do cosseno do ângulo entre os vectores. O resultado da função de cosseno varia entre $[0 \dots 1]$, e quanto maior for esse valor, mais similares são os vectores:

$$\text{sim}(X,Y) = \cos(\theta) = \frac{X \cdot Y}{\parallel X \parallel \parallel Y \parallel} = \frac{\sum_{i=1}^{n} X_i \times Y_i}{\sqrt{\sum_{i=1}^{n} (X_i)^2} \times \sqrt{\sum_{i=1}^{n} (Y_i)^2}}$$

Temos também que uma das técnicas possíveis de ser usadas para calcular a similaridade entre duas imagens é a *intersecção de histogramas* de características representativas, obtidas por exemplo através da quantização de *vectores de características*. Sejam X e Y, dois histogramas de características correspondentes a duas imagens. Temos que a *intersecção de histogramas* é definida por:

$$\text{sim}(X,Y) = \frac{\sum_w min(X^w, Y^w)}{\sum_w max(X^w, Y^w)} \tag{2.2}$$

Na fórmula acima, tanto X^w como Y^w correspondem a um valor quantizado para o peso associado a cada uma das características das imagens.

2.4 Estruturas de Indexação para Conteúdos Textuais

A Recuperação de Informação *RI* é uma área da ciência da computação que lida com a procura de recursos (geralmente documentos) de natureza não-estruturada (normalmente texto), que satisfazem

a necessidade de uma determinada informação, a partir de grandes colecções de recursos [Manning et al., 2008]. Em colecção de documentos textuais, um Sistema de Recuperação de Informação (*SRI*) tem como principal objectivo retornar os documentos que possam ser úteis ou relevantes para a necessidade de informação do utilizador (interrogação), ordenando-os de acordo com o seu grau de relevância. Uma das estratégias simples a usar por um *SRI* consiste em determinar quais documentos em uma colecção contêm os termos da interrogação. Para isto, é necessário comparar os termos da interrogação com todos os termos dos documentos da colecção, o que é uma tarefa computacionalmente cara. Com o objectivo de acelerar este processo, surge a necessidade de limitar a comparação com todos os documentos da colecção, usando estruturas de indexação eficientes associadas aos documentos de uma colecção.

2.4.1 Índices Invertidos

A estrutura de indexação mais frequentemente utilizada em *SRI* é o índice invertido, o qual permite pesquisas eficientes, limitando a comparação com todos os documentos da colecção [Baeza-Yates and Ribeiro-Neto, 1999]. O índice invertido é composto por dois elementos fundamentais, nomeadamente o vocabulário e as ocorrências. O vocabulário corresponde ao conjunto de todos os termos distintos na colecção. Geralmente, antes da construção de um vocabulário, é necessário efectuar um pré-processamento dos termos, de modo a reduzir o tamanho do vocabulário, possibilitando, assim, um maior desempenho computacional nos *SRI*. Um exemplo seria a remoção de termos irrelevantes para a discriminação de um documento, denominados de *stopwords*, tais como por exemplo, artigos, preposições, pronomes, e ainda os termos muitos frequentes na colecção. Outras técnicas de pré-processamento podem ser encontradas nos trabalhos de Manning et al. [2008].

Já as ocorrências são um conjunto de listas, sendo cada uma definida para um termo do vocabulário, e composta pelos identificadores dos documentos nos quais o termo ocorre. Adicionalmente, juntamente com os identificadores, podem ser armazenadas outras informações dos termos do vocabulário, como por exemplo, as suas posições nos documentos ou os seus respectivos valores de importâncias, obtidos através das abordagens introduzidas na Secção 2.2.1. Assim sendo, as informações armazenadas num *SRI* variam de acordo com o tipo de aplicação, por exemplo, as posições dos termos nos documentos podem ser úteis em aplicações que utilizam a proximidade entre os termos para calcular a relevância dos resultados.

Seja a Tabela 2.1, a representação de uma colecção de documentos com seus respectivos conteúdos textuais, podemos então construir um índice invertido para uma rápida recuperação desses documentos, tal como ilustrado na Tabela 2.2. Observa-se neste exemplo que, cada termo do vocabulário foi associado à duas informações, i.e., cada termo foi mapeado à uma lista de pares composta por ocorrências e posições em cada documento da Tabela 2.1.

Dado um documento de interrogação $d_q = days\,in\,the\,pot$, para encontrarmos quais os potências

Doc. ID	Texto
1	pease porridge hot, pease porridge cold
2	pease porridge in the pot
3	nine days cold
4	some like it hot, some like it cold
5	some like it in the pot
6	nine days old

Tabela 2.1: Colecção de Documentos.

Nrº	Vocabulário	(Ocorrências; Posições)
1	cold	(1;6), (4;8)
2	days	(3;2), (6;2)
3	hot	(1;3), (4;4)
4	in	(2;3), (5;4)
5	it	(4;3,7), (5;3)
6	like	(4;2,6), (5;2)
7	nine	(3;1), (6;1)
8	old	(3;3), (6;3)
9	pease	(1;1,4), (2;1)
10	porridge	(1;2,5), (2;2)
11	pot	(2;5), (5;6)
12	some	(4;1,5), (5;1)
13	the	(2;4), (5;5)

Tabela 2.2: Índice invertido com as posições dos termos.

documentos semelhantes a d_q, usamos o índice invertido para encontrar quais documentos da colecção contêm pelo menos um dos termos contidos em d_q. Assim, serão calculadas as similaridades apenas entre d_j e os documentos 2, 3, 5 e 6, evitando deste modo a comparação com os documentos 1 e 4.

2.5 Sumário

Neste capítulo, foram apresentados os conceitos fundamentais necessários para entender o trabalho desenvolvido. A Secção 2.1 introduziu conceitos relacionados com o processamento de imagens digitais, destacando os atributos de cor e de textura usados no processo de caracterização de imagens. Na Secção 2.2, foram abordados conceitos básicos da área de recuperação de imagens. A recuperação de imagens pode ser baseada em dois tipos de características, nomeadamente textuais e visuais. Foram apresentadas com maior ênfase as características textuais, as quais foram o foco deste trabalho, descrevendo-se as principais técnicas para a quantificação das mesmas. Na Secção 2.3, foram descritas varias técnicas para o cálculo da similaridade entre vectores de características. Finalmente, com o intuito de acelerar o processo de recuperação de imagens nos *TBIRs*, a estrutura de indexação do índice invertido foi apresentada na Secção 2.4. O índice invertido possibilita uma pesquisa mais eficiente, uma vez que permite que sejam armazenadas informações adicionais que, acompanhadas de algoritmos adequados, facilitam a classificação e ordenação dos resultados.

Capítulo 3

Trabalho Relacionado

Vários trabalhos anteriores abordaram a tarefa da georreferenciação automática de imagens, explorando técnicas diferentes, tais como modelos de linguagem definidos com base em anotações textuais associadas às fotos, ou classificadores combinando diversos tipos de características. Este capítulo sumariza os principais trabalhos anteriores relacionados com o tema da georreferenciação de fotos, os quais têm algumas características semelhantes com o trabalho desenvolvido.

3.1 Estimar Localizações Geográficas a partir de uma Imagem

Hays and Efros estimaram distribuições a para localização geográfica de uma imagem, baseando o seu trabalho numa base de dados de imagens georreferenciadas com coordenadas *GPS* [Hays and Efros, 2008]. Estes autores começaram por aplicar um filtro nesta base de dados, resultando assim numa colecção com aproximadamente 6 milhões de imagens georreferenciadas do *Flickr*. Este filtro consiste em seleccionar apenas as imagens com pelo menos um nome de um país, continente, cidades com alta densidade de população, ou os lugares turísticos populares, que não tinham anotações textuais tais como datas de nascimento. Cada imagem é representada por um conjunto de características visuais, as quais incluem o *Tiny Images* [Torralba et al., 2007], o *Gist Descriptor* [Oliva and Torralba, 2006], um *Color Histogram*, um *Texton Histogram* [Martin et al., 2001], *Line Features* [Kosecka and Zhang., 2002], e um descritor baseado no *Geometric Context* [D. Hoiem and Hebert., 2005], respectivamente.

Dada uma imagem de teste, primeiro são extraídas todas as características visuais descritas acima, sendo então calculada a distância para com todas as outras características visuais das imagens da base de dados. Os autores usaram uma abordagem de pesquisa pelos k vizinhos mais próximos, para estimarem a localização geográfica, através de duas heurísticas. A primeira heurística consiste em utilizar as coordenadas geográficas do primeiro vizinho mais próximo (i.e., *1-NN*) como estimativa da localização. Já na segunda heurística, primeiro são construídos *clusters* a partir dos 120 vizinhos mais próximos (i.e., $120 - NN$), usando a abordagem *mean shift clustering* [Comaniciu and Meer, 2002]. Em seguida, a estimativa da localização é obtida a partir das coordenadas do *cluster* com maior cardinalidade.

Os autores testaram todas as características visuais, tanto em conjunto como também isoladamente para as duas heurísticas. Usando as características visuais isoladamente, para ambas as heurísticas, obteve-se melhores resultados em termos de precisão, do que quando as mesmas foram usadas em conjunto. As duas heurísticas tiveram resultados similares quando são usadas todas as características visuais em conjunto, mas considerando características visuais isoladamente os *clusters* tiveram melhores resultados. Considerando todos estes testes, cerca de 16% das imagens de teste foram estimadas com um erro de distância menor ou igual a 200 km em relação a localização correcta da imagem.

3.2 Associar as Fotografias do Flickr a uma Área Geográfica

Serdyukov et al. investigaram métodos genéricos para a georeferenciação de fotografias do *Flickr*, utilizando apenas as anotações textuais fornecidas pelos utilizadores [Serdyukov et al., 2009]. Estes autores usaram uma representação para o mundo baseada numa grelha, onde cada célula representa uma localização. Estas células têm uma resolução em termos de latitude variando de 0 km nos polos a 111 km no equador [Toyama et al., 2003], e uma resolução de cerca de 111 km por longitude para cada posição. Cada fotografia de treino (i.e., com informação de latitude e longitude) é associada à sua célula correspondente na grelha. Para cada uma das células, é construído um modelo de linguagem a partir das anotações textuais atribuídas pelos utilizadores às fotos. Este modelo de linguagem é baseado na abordagem *bag-of-tags*, onde a ordem dos termos das anotações textuais não influencia o resultado.

As células da grelha são organizadas numa estrutura representada por um grafo não direccionado, de modo a modelar as suas relações espaciais e semânticas. A ligação entre um par de células (i.e., localizações) existe somente se estão próximas entre si e a uma determinada distância pré-definida. Portanto, todas as células encontradas dentro deste raio de distância, em relação a uma dada célula, são consideradas vizinhas, i.e., têm maiores probabilidades de serem representadas por anotações textuais similares.

Dada uma fotografia de teste, esta é atribuída à célula cujo modelo de linguagem tem uma maior probabilidade de gerar as suas anotações textuais. A ideia geral deste processo consiste em obter uma ordenação das localizações (i.e., das células consideradas para a representação do espaço geográfico) através da probabilidade de obter uma dada localização L, dada uma anotação textual T:

$$P(L|T) = \frac{P(T|L)P(L)}{P(T)}$$

Uma vez que não se tem qualquer informação à priori acerca das localizações e das anotações textuais, que poderia influenciar a ordenação, os autores consideraram que $P(L)$ é distribuída de forma uniforme, e que $P(T)$ não influencia a ordenação [Peng et al., 2004]. Assumindo que cada termo $t_i \in T$ é gerado de forma independente, o termo $P(L|T)$ da equação anterior é calculado da seguinte forma:

$$P(T|L) = \prod_{i=1}^{|T|} P(t_i|L)$$

A probabilidade de cada termo individual t_i, dado o modelo de linguagem de uma localização L

(i.e., $P(t_i|L)$), é obtida através de uma combinação de duas abordagens, conforme apresentada na Equação 3.1. A primeira abordagem é a *estimativa por máxima verosimilhança* (*MLE*), que maximiza a probabilidade dos dados observados (i.e., busca os parâmetros que maximizem a função de verosimilhança) [Glen, 2010]. A segunda abordagem corresponde a um *Dirichlet smoothing* [Zhai and Lafferty, 2002], que executa uma interpolação linear com uma estimativa calculada a partir de um modelo de linguagem global.

$$P(t_i|L) = \frac{|L|}{|L|+\lambda} P(t_i|L)_{MLE} + \frac{\lambda}{|L|+\lambda} P(t_i|G)_{MLE} \tag{3.1}$$

Na fórmula acima, $P(t_i|L)_{MLE}$ e P $(t_i|G)_{MLE}$ representam as probabilidades de geração das anotações textuais, para a localização L e para o modelo de linguagem global G, usando a *MLE*. O parâmetro $|L|$ representa a dimensão das anotações textuais da localização L, e λ é o parâmetro do *Dirichlet smoothing*.

Serdyukov et al. introduziram ainda 4 extensões para melhorar este modelo de linguagem base (*ML*), baseadas em suavização das anotações textuais e das células representadas na grelha. Estas extensões são descritas de seguida.

Tag-based smoothing (TS): Considerando o facto de que algumas anotações textuais indicam uma área que ultrapassa os limites de uma célula específica, os autores usaram células vizinhas da grelha para suavizar as probabilidades do modelo de linguagem. Para tal, cada termo de uma anotação textual encontrada dentro de uma célula é gerado usando o seu modelo de linguagem e os modelos de linguagem das células vizinhas numa distância pré-determinada:

$$P(t_i|L) = \mu \frac{|L| \cdot P(t_i|L)_{MLE}}{|L|+\lambda} + (1-\mu) P(t_i|NB(L)) + \frac{\lambda \cdot P(t_i|G)_{MLE}}{|L|+\lambda}$$

$$P(t_i|NB(L)) = \sum_{L' \in NB(L)} \frac{|L'|}{|L'|+\lambda} \frac{P(t_i|L')_{MLE}}{(2d+1)^2 - 1}$$

Nas fórmulas acima, $NB(L)$ corresponde a todas as localizações vizinhas da localização L, a uma distância d, e o parâmetro μ é o coeficiente de suavização para a probabilidade do termo t_i ser gerado a partir da localização L.

Cell-based Smoothing (CS): Nesta extensão, os autores aumentaram a probabilidade de gerar as anotações textuais de uma determinada célula com as probabilidades das células vizinhas:

$$P(T|L) = \mu P(T|L) + (1-\mu) \sum_{L' \in NB(L)} \frac{P(T|L')_{MLE}}{(2d+1)^2 - 1}$$

Com o intuito de evitar o cálculo das probabilidades de todas as localizações vizinhas, os autores atribuíram pesos às localizações. A atribuição de pesos considera que as localizações muito distantes

não têm grande influência sobre a relevância de uma localização específica. Sendo assim, a suavização das probabilidades dos modelos de linguagem das células vizinhas é efectuada através da propagação de informações a partir das células com pesos mais baixos do que as células a serem suavizadas. Esta melhoria é denominada de suavização de células com propagação de pesos, na direcção dos vizinhos altamente relevantes (*CRS*).

Toponym based boosting (TB): Nesta extensão é efectuada a incorporação de informação relativa a localizações de topónimos, a partir de uma grande base de dados externa, a fim de ter conhecimentos preliminares sobre as localizações das anotações textuais no modelo. Os autores usaram uma lista de topónimos com nomes ingleses para localizações povoadas, retiradas da base de dados *GeoNames*[1]. Esta informação é introduzida no modelo a partir da seguinte fórmula:

$$P(t_i|L) = P(t_i|L)_{MLE}(1 + \beta P(Loc|t_i))/Z$$

Na fórmula acima, $P(t_i|L)_{MLE}$ corresponde a probabilidade do termo t_i estar na lista de topónimos, retornando um se o termo estiver nesta lista, e zero caso contrario. O parâmetro β é o coeficiente de suporte para a lista de topónimos, e o parâmetro Z corresponde a um coeficiente de normalização.

Ambiguity-aware tag specific smoothing (AS): Esta extensão usa as coordenadas geográficas conhecidas de todos os termos das anotações textuais nos dados de treino, caracterizando a sua ambiguidade espacial através do desvio padrão das suas coordenadas de latitude e longitude. Para tal, os autores alteraram o coeficiente de suavização λ da Equação 3.1, considerando as informações geográficas conhecidas de cada termo específico:

$$\lambda(t_i) = \lambda + \gamma(\sigma_{lat}(t_i) + \sigma_{lon}(t_i))$$

Na fórmula acima, γ é um coeficiente para controlar a influência do nível da ambiguidade na suavização.

Para a avaliação destes métodos, os autores recolheram 397.000 fotografias georreferenciadas do *Flickr*, de 180 países diferentes. Em seguida, aplicaram um filtro nestas fotografias, resultando assim numa colecção com 140.000 fotografias. Este filtro consiste em assegurar que em localizações de baixa granularidade espacial, exista no máximo uma fotografia por utilizador juntamente com o seu conjunto de anotações textuais. Estas fotografias foram então divididas de forma aleatória, considerando 120.000 fotografias para os dados de treino, 10.000 fotografias para afinação dos parâmetros, e as restantes 10.000 fotografias para os dados de teste, respectivamente. Estes autores testaram o método base LM, em comparação coma a adição de cada uma das extensões propostas. Todas as combinações apresentaram melhores resultados do que o método base. De entre todas as extensões, TB apresentou o melhor desempenho. Além disso, quando considerando as extensões TS, CS e CSR para a suavização com base nas células vizinhas, a extensão CRS apresentou melhores resultados.

[1] `http://www.geonames.org/`

3.3 Mapeando as Fotografias do Mundo

Crandall et al. desenvolveram um método para georreferenciar as imagens de uma colecção de aproximadamente 35 milhões de fotos extraídas do *Flickr*, incluindo seus meta-dados, usando uma combinação de características visuais, textuais e temporais [Crandall et al., 2009]. Estes autores definiram a localização de uma fotografia considerando dois níveis de granularidade, nomeadamente o nível da *escala metropolitana* (até cerca de 100 km) e o nível da *escala de landmark individual* (até cerca de 100 m). Os lugares populares onde as pessoas fazem fotografias, nestas escalas, são identificados usando o procedimento *mean shift clustering* [Comaniciu and Meer, 2002], e classificados com base no número de fotógrafos distintos. As experiências reportadas por estes autores estavam limitadas a um conjunto de 10 *landmarks* (i.e., pontos turísticos) e a um conjunto fixo de cidades, ou seja, não existem imagens nos seus conjuntos de teste que representam lugares fora deste conjunto.

Considerando apenas características visuais (i.e., descritores *SIFT* para as imagens [Lowe, 1999]) e as anotações textuais, a identificação da localização de onde uma foto foi tirada é efectuada da seguinte forma. Primeiramente, é construído um classificador para cada um dos dez *landmarks* da cidade onde a foto foi tirada. Para fazer isto, os classificadores são baseados na abordagem *Support Vector Machines* (*SVMs*) [Joachims, 1999]. Para cada um dos 10 classificadores, os exemplos positivos são as fotos nos *landmarks* desta cidade, e os exemplos negativos são as fotos noutros *landmarks*. Finalmente, para identificar a localização, é analisado cada um dos 10 classificadores sobre uma determinada foto de teste, escolhendo o *landmark* com a maior certeza na classificação.

Com o objectivo de melhorar o desempenho da classificação, os autores integraram informação temporal directamente no processo de classificação, da seguinte forma. Ao classificar uma foto, são analisadas todas as outras fotos tiradas pelo mesmo fotógrafo num intervalo temporal relativo aos 15 minutos antes e depois. Para cada uma destas fotos, calculam-se as distâncias de classificação para cada um dos 10 classificadores de *SVM*. Os valores de distâncias das diferentes fotos são somados para produzir um único vector e, em seguida, é tomada a decisão de classificação usando este vector. Os resultados experimentais mostraram que as características visuais e temporais melhoram a capacidade de estimar a localização de uma fotografia, em comparação com o uso apenas das características textuais.

Os mesmos autores apresentaram ainda uma técnica para escolher as imagens representativas dos *landmarks* e das cidades mais fotografadas. As imagens representativas são identificadas procurando por subconjuntos de fotos que são visualmente muito semelhantes, e escolhendo uma imagem mais saliente entre o subconjunto. Os autores começaram por construir um grafo onde cada nó representa uma foto, e entre cada par de nós existe uma aresta com um peso que indica o grau de semelhança visual entre as duas fotografias. Em seguida, é encontrado um grupo fortemente ligado de fotos que são muito semelhantes, através de uma técnica de agrupamento espectral [Shi and Malik, 2000]. Finalmente, escolhe-se como imagem representativa para cada grupo, a que corresponde ao nó com maior peso de semelhança associado. Os resultados demonstraram que as fotos representativas podem ser

seleccionadas automaticamente, apesar de grande parte das fotos de um determinado local não estar relacionada com qualquer *landmark* em particular.

3.4 Georreferenciação Automática de Fotos

Van Laere et al. apresentaram uma abordagem automática para prever as coordenadas geográficas de fotografias, baseada no algoritmo de agrupamento *k-medoides* e na classificação *Naive Bayes*, e usando apenas as anotações textuais [Van Laere et al., 2010]. Estes autores começaram por recolher fotografias georreferenciadas, publicadas no *Flickr*, das 55 grandes cidades Europeias. Estas cidades foram escolhidas através da intersecção entre um conjunto de 100 cidades Europeias com população mais densa[2], e um conjunto de 160 cidades Europeias mais importantes para o turismo[3]. Em seguida, aplicaram dois tipos de filtros nesta colecção de fotografias recolhidas, resultando assim numa colecção com aproximadamente 1 milhão de fotografias. O primeiro filtro consiste em remover todas as fotos cujas as coordenadas têm uma precisão menor ou igual a 13 (O *Flickr* fornece informação sobre a precisão das coordenadas como um número entre 1 (i.e., a nível mundial) e 16 (i.e., nível da rua)). O segundo filtro consiste em remover as fotos cujas as anotações textuais e o nome de utilizador são idênticas a uma foto já existente na colecção. Esta colecção foi dividida considerando 66% para os dados de treino e 33% para os dados de teste, de tal forma que todas as fotos de um dado utilizador ou estão nos dados de treino ou nos dados de teste.

O processo da previsão das coordenadas geográficas de uma dada fotografia de teste é efectuado da seguinte forma. Primeiramente, os autores dividiram cada uma das 55 cidades em conjuntos de áreas disjuntas, usando o algoritmo de agrupamento *k-medoides*. De seguida, com o objectivo de melhorar a robustez desta abordagem, os autores construíram um vocabulário para cada uma dessas áreas da seguinte forma. Primeiramente, removeram todas as anotações textuais usadas por menos do que dois utilizadores. Em seguida, removeram todas as anotações textuais que não são indicativos de uma determinada área em particular, através da aplicação da técnica de selecção de características denominada χ^2. Seja A um conjunto de áreas obtidas após a aplicação do algoritmo *k-medoides*. Para cada área a pertencente a A, e para cada anotação textual t que ocorra nas fotografias pertencentes à área a, o valor da componente χ^2 é calculado da seguinte forma:

$$\chi^2(a,t) = \frac{(O_{ta} - E_{ta})^2}{E_{ta}} + \frac{(O_{t\bar{a}} - E_{t\bar{a}})^2}{E_{t\bar{a}}} + \frac{(O_{\bar{t}a} - E_{\bar{t}a})^2}{E_{\bar{t}a}} + \frac{(O_{\bar{t}\bar{a}} - E_{\bar{t}\bar{a}})^2}{E_{\bar{t}\bar{a}}}$$

Na fórmula acima, O_{ta} é o número de fotografias na área a que contêm a anotação textual t, $O_{t\bar{a}}$ é o número de fotografias fora da área a que contêm a anotação t, $O_{\bar{t}a}$ corresponde ao número de fotografias da área a que não contêm a anotação t, e $O_{\bar{t}\bar{a}}$ é o número de fotografias fora da área a que não contêm a anotação t. Além disso, E_{ta} corresponde ao número de ocorrências da anotação textual t nas fotografias da área a, que poderia ser esperado quando a ocorrência de t for independente da localização da área a. Isto é, $E_{ta} = N \cdot P(t) \cdot P(a)$, onde N representa o número total de fotos, $P(t)$ é

[2] f-http://www.nga.mil/
[3] f-http://www.visiteuropeancities.info/

a percentagem de fotos que contêm a anotação textual t, e $P(a)$ é a percentagem de fotos que estão localizados em área a. Da mesma forma, $E_{t\overline{a}} = N \cdot P(t) \cdot (1 - P(a))$, $E_{\overline{t}a} = N \cdot (1 - P(t)) \cdot P(a)$, $E_{\overline{t}\overline{a}} = N \cdot (1 - P(t)) \cdot (1 - P(a))$. Finalmente, o vocabulário para cada uma das áreas é construído a partir das 25 anotações textuais que contêm maior valor de χ^2.

Dada uma fotografia de teste x, é determinada em que área a em que ela foi provavelmente tirada, i.e., a estimativa de $P(a|x)$. Esta estimativa é obtida através da classificação *Naive Bayes*, a qual também foi usada nas abordagens propostas nos trabalhos de Serdyukov et al. [2009] e de O'Hare and Murdock [2012a]:

$$P(a|x) = \frac{P(a) \cdot P(x|a)}{P(x)} \tag{3.2}$$

Uma vez que a probabilidade $P(x)$ de se observar as anotações textuais associadas à imagem x é fixa entre todas as áreas, ou seja, temos que $P(x)$ não influencia a ordenação. Além disso, considerando o pressuposto de que cada anotação textual é gerada de forma independente, a Equação 3.2 é substituída pela seguinte:

$$P(a|x) \propto P(a) \cdot \prod_{t \in x} P(t|a)$$

$$P(a) = \frac{|X_a|}{\sum_{b \in A} |X_b|} \qquad P(t|a) = \frac{N_t + 1}{\left(\sum_{y \in X_a} |y|\right) + |V|}$$

Nas fórmulas anteriores, $|X_a|$ corresponde ao número total de imagens na área a, e $\sum_{b \in A} |X_b|$ corresponde ao número total imagens em todas as áreas. O parâmetro N_t é o número de imagens na área a que contêm a anotação textual t, $\sum_{y \in X_a} |y|$ representa o número total de anotações textuais que ocorrem em todas as imagens na área a, e V corresponde ao vocabulário, como definido anteriormente.

Nas suas experiências, estes autores tentaram descobrir tanto em que cidade uma foto foi tirada, como também onde ela foi tirada dentro daquela cidade. Ou seja, a classificação *Naive Bayes* foi treinada ao nível da cidade, e ao nível das subáreas dentro da cidade, usando agrupamentos de 250, 500 e 1000 áreas. Os melhores resultados foram alcançados na tarefa de estimar a cidade onde uma foto foi tirada, com uma melhor precisão de cerca de 87%.

3.5 Estimar a Localização de uma Fotografia de Praia

Cao et al. desenvolveram um método para associar fotografias de praias a *clusters geográficos*, ou seja, agrupamentos de fotografias de praias com informação geográfica relativa ao local onde foram capturadas [Cao et al., 2012]. Estes autores, argumentaram que o uso de *clusters geográficos* fornece benefícios às duas fases (i.e., treino e de teste) do processo de identificação de localizações geográficas de fotografias. Na fase de treino, os *clusters geográficos* fornecem mais exemplos de treino

e, consequentemente, consegue-se obter uma melhor precisão na estimativa da localização geográfica de uma determinada fotografia. Na fase de teste, os *clusters geográficos* fornecem informações úteis para o planeamento de uma viagem, e para as aplicações de organização de fotografias.

Os autores começaram por recolher 35.000 fotografias de praias, do *Flickr*, onde cada fotografia contém um vector bidimensional com as coordenadas *GPS*. A partir das coordenadas das 33.900 fotografias, construíram *clusters geográficos*, usando o procedimento *mean shift clustering* [Comaniciu and Meer, 2002]. As restantes 1.100 fotografias foram consideradas como dados de teste. Em seguida, desenvolveram uma estratégia iterativa, onde simultaneamente treinam classificadores visuais (i.e., baseados em descritores *SIFT* [Lowe, 1999] para as fotografias) e fazem a refinação dos *clusters geográficos*. Em cada iteração, um classificador visual é treinado para cada *cluster geográfico*. Este treino, consiste em comparar as características visuais entre as fotografias dos *clusters*. Com base nos votos dos classificadores associados a cada *cluster*, calcula-se uma taxa de falso alarme (*FA*), juntamente com a taxa de erro de detecção (*MD*). Os clusters com elevadas taxas de *FA* (i.e., os que englobam grandes regiões) são divididos, e os *clusters* com baixas taxas de *MD* (i.e., os que englobam regiões não distinguíveis) são removidos. Em seguida, os *clusters* são refinados, e os classificadores são treinados para a próxima iteração. O processo termina quando os *cluters* são finalizados sem nenhuma divisão ou remoção. Finalmente, os classificadores são usados para associar uma nova fotografia de teste a um *cluster*, com base nas suas características visuais.

Os autores testaram o método iterativo, em comparação com um método *random* e um método baseado nos vizinhos mais próximos. Este método teve melhor precisão, ao associar cada fotografia de teste ao *cluster* correcto.

3.6 Modelos de Localização Baseados em Géneros a partir das Fotos do Flickr

O'Hare and Murdock exploraram as informações dos perfis dos utilizadores do *Flickr* para criar modelos de localização específicos para cada género, usando apenas as anotações textuais para prever a localização de uma foto [O'Hare and Murdock, 2012a]. Este trabalho é semelhante à abordagem proposta no trabalho de Serdyukov et al. [2009]. O'Hare and Murdock também representaram o mundo como uma grelha, criada pela quantificação dos valores de latitude/longitude para criar células com diferentes granularidades (i.e., 100 ou 1 km^2 de dimensão). Cada célula da grelha representa uma localização e é constituída por um conjunto de fotos com meta-dados textuais, incluindo título e descrição, em conjunto com as informações dos utilizadores do *Flickr*. Considerando as anotações textuais fornecidas pelos utilizadores, para cada célula, os autores construíram um modelo de linguagem. Comparando com Serdyukov et al. [2009], neste trabalho os modelos de linguagem são criados com um conjunto de dados muito maior, usando uma abordagem padrão dos modelo de linguagem para *Recuperação de Informação (RI)* [Ponte and Croft, 1998]. Especificamente, como base de dados de imagens, estes autores usaram um conjunto de 10 milhões de fotografias georreferenciadas recolhidas

do Flickr, as quais foram então divididas de tal forma que todas as fotos de um dado utilizador ou estão nos dados de treino ou nos dados de teste, da seguinte forma. Um total de 8 milhões de fotografias para os dados de treino, 1 milhões de fotografias para afinação dos parâmetros, e os outros restantes 1 milhões de fotografias para os dados de teste respectivamente.

Temos ainda que a ordenação das localizações para uma nova foto, i.e., a estimativa de $P(L|T)$, é obtida da mesma forma e considerando os mesmos pressupostos da Serdyukov et al. [2009]. Deste modo, para cada localização (i.e., para cada célula da representação), dado o seu modelo de linguagem, é calculada a probabilidade de obter a anotação textual, através da seguinte fórmula:

$$P(T|\theta_L) = \prod_{i=0}^{|T|} P(t_i|\theta_L)$$

Na fórmula acima, θ_L corresponde a um modelo de linguagem de uma determinada localização L, T é um conjunto de anotações textuais, e t_i é um termo de T. Estes autores demonstraram, em trabalhos anteriores, que estimar a probabilidade de cada t_i dado um θ_L , i.e., $P(ti|\theta_L)$, com base na frequência de utilizadores, tem maior precisão do que fazer o mesmo com base na frequência de termos [O'Hare and Murdock, 2012b]. Com base nesta observação, estes autores estimaram a probabilidade de um termo considerando a frequência de utilizadores, como definida na fórmula seguinte:

$$P_{user}(t|\theta_L) = \frac{C_{user}(t, L)}{\sum_{t_i \in L} C_{user}(t_i, L)} \tag{3.3}$$

Na fórmula acima, $c_{user}(t, L)$ é o número de utilizadores que usam o termo t na localização L. O denominador corresponde à soma das frequências dos utilizadores para todos os termos de uma localização.

Considerando a Equação 3.3 em conjunto com o género dos utilizadores, os autores criaram um modelo de linguagem baseado no género dos utilizadores. Em vez de $P(L|T)$, calcula-se $P(L|T, G)$, que é a probabilidade de uma localização, dada uma anotação textual e um género. Desta forma, os autores substituíram a Equação 3.3 pela seguinte:

$$P_{user}(t|\theta_{L,G}) = \frac{C_{user}(t, L, G)}{\sum_{t_i \in L} C_{user}(t_i, L, G)}$$

Na equação acima, $C_{user}(t, L, G)$ corresponde ao número de utilizadores de um dado género que utilizam o termo t na localização L. Os resultados experimentais mostraram claramente que as fotos anotadas por utilizadores do género masculino contem informação de localização mais útil do que aquelas anotadas por utilizadores do género feminino. Usar somente os dados de utilizadores do género masculino fornece resultados com melhor precisão.

3.7 Estimativa da Localização Geográfica das Fotografias do Flickr com base em Fontes Externas

Hauff and Houben investigaram até que ponto a georreferenciação de fotografias do *Flickr*, baseada em anotações textuais dos utilizadores, pode ser melhorada considerando também os *tweets* dos autores das fotos (i.e., mensagens publicadas no *Twitter*[4]), como uma fonte de informação textual adicional [Hauff and Houben, 2012]. De forma semelhante aos trabalhos de Serdyukov et al. [2009] e de O'Hare and Murdock [2012a], estes autores também usaram uma representação para o mundo baseada numa grelha, onde cada célula representa uma localização. Em comparação com estes trabalhos, onde foram usadas células com dimensões fixas, neste trabalho as células são de dimensões variáveis. As grelhas de células são construídas começando com uma única célula, que se estende por todo o mapa do mundo (i.e., se for visto como um gráfo, esta célula é o nó raiz). Em seguida, as fotografias de treino com meta-dados textuais são adicionadas nas células, uma de cada vez. Uma vez que o número de fotografias numa determinada célula exceder um determinado limite de divisão l_{split}, esta célula será dividida em quatro células de tamanhos iguais. Assim, cada uma destas células, com dimensão igual a um quarto da célula original, é adicionada na grelha, e as fotografias de treino são redistribuídas para estas células. De modo a evitar varias divisões em localizações onde existem grandes quantidades de fotografias de treino, estes autores consideraram que uma célula não pode ser mais dividida quando a diferença dos seus valores mínimo e máximo de latitude e longitude atingirem um determinado limite inferior l_{lat_lng}. Este processo produz células de pequenas dimensões em localizações onde existem grandes quantidades de fotografias de treino, e células de grandes dimensões onde existem poucas quantidades de fotografias de treino (e.g., os oceanos).

A partir das anotações textuais associadas às fotografias de treino em cada célula, Hauff and Houben construíram também os modelos de linguagens para cada célula, como nos trabalhos de Serdyukov et al. [2009] e de O'Hare and Murdock [2012a]. Dada uma fotografia de teste, a sua georreferenciação é realizada em duas etapas. A primeira etapa envolve obter uma ordenação das localizações θ_R, dada a anotação textual T de uma determinada fotografia de teste, i.e., a estimativa de $P(\theta_R|T)$. Se T for vazia, então, serão considerados os termos da localização do utilizador, incluídos nos meta-dados. Esta estimativa é obtida de forma análoga ao trabalho de Serdyukov et al. [2009], com a diferença de que neste trabalho, $P(\theta_R)$ é considerada dependente de uma região (i.e., regiões com densidades de população elevadas tem uma maior probabilidade de terem fotografias, do que as regiões com densidades baixas).

$$P(\theta_R|T) = \frac{P(T|\theta_R)P(\theta_R)}{P(T)} \propto P(\theta_R) \times \prod_{i=1}^{n} P(t_i|\theta_R)$$

Assim, na segunda etapa, serão consideradas apenas as fotografias da localização do topo da ordenação. Em seguida, é gerado um modelo de linguagem para cada uma das fotografias desta localização, e as mesmas são ordenadas de acordo com as suas probabilidades de geração das anotações textuais da fotografia de teste. Por fim, a fotografia de teste será atribuída às coordena-

[4] `https://twitter.com/`

das de latitude e longitude da fotografia do topo desta ordenação.

Com o objectivo de melhorar este método base, os autores apresentaram dois filtros básicos, baseados nos dados de treino, para aplicar às anotações textuais. O primeiro filtro denomina-se *Geographic Spread Filtering*, e o mesmo remove todos os termos pertencentes a T que são considerados como inúteis para a identificação de uma localização (i.e., ruídos geográficos). O segundo filtro é o *User Spread Filtering*, que remove todos os termos usados por menos do que um determinado número de utilizadores nos dados do treino. Caso a aplicação destes dois filtros produza anotações T vazias, então serão atribuídas, à fotografia de teste, as coordenadas de latitude e longitude da localização do topo da ordenação, obtida na primeira etapa.

Os mesmos autores melhoraram ainda mais o método base, com a introdução de *tweets* dos utilizadores nas anotações textuais. Dada uma fotografia de teste, com as anotações textuais T de um utilizador U, são extraídos os termos adicionais nos *tweets* de *U* e adicionados a *T*. Em relação à extracção dos termos, foram considerados todos os termos de todos os *tweets* (i.e., *URLs* e os termos dos nomes dos utilizadores são removidos) ou todos os *hashtags* (e.g., $\#$ecir2012, $\#$barcelona) disponíveis. Esta melhoria é combinada com os filtros descritos anteriormente, resultando num conjunto *T* de pequena dimensão. Importa aqui referir que foram usados apenas os *tweets* e as fotografias mais recentes nos dados de teste, de modo a evitar grandes diferenças temporais entre eles.

Para a avaliação destas abordagens, os autores usaram os dados disponibilizadas no âmbito do *MediaEval* 2011. Esta colecção de dados inclui um total de $3,2$ milhões de fotografias e 10.000 vídeos, em conjunto com os seus meta-dados textuais. Com base nesta colecção, os autores criaram um conjunto de 77.591 fotografias para os dados de treino e um outro conjunto de 8.306 fotografias para os dados de teste. Estas fotografias são apenas dos utilizadores que ligam explicitamente as suas contas do *Flickr* e do *Twitter*, obtidas a partir de um agregador social FriendFeed[5]. Resultados experimentais mostraram que, usando os dois filtros e também todos os *tweets* e *hashtags* disponíveis, as fotografias de teste com menos de três termos são as mais beneficiadas com a adição de informações do *Twitter*.

3.8 Participação da equipe *CEA LIST* no *MediaEval* 2013

Popescu apresentaram uma técnica para a georreferenciacão de fotografias do *Flickr*, usando apenas os conteúdos textuais [Popescu, 2013]. De forma semelhante aos trabalhos de O'Hare and Murdock [2012a], de Serdyukov et al. [2009] e de Hauff and Houben [2012], estes autores também usaram uma representação para o mundo baseada numa grelha, onde cada célula representa uma localização. Para cada célula os autores construíram um modelo de linguagem a partir dos meta-dados textuais associados às fotografias de treino. A probabilidade de um determinado conjunto de *tags T* pertencer

[5] `http://friendfeed.com/`

a uma localização L é calculada através da seguinte fórmula:

$$P(T|L) = \frac{\sum_{i=0}^{|T|} C_{user}(t_i, L)}{\sum_{i=o,j=0}^{|T|,|L|} C_{user}(t_i, Lj)}$$

Na fórmula acima, $C_{user}(t_i, L)$ é número de utilizadores que utilizaram a *tag* t_i na localização L, e t_i é um termo de T. O denominador corresponde ao número total de utilizadores que usaram as *tags* T em todas as localizações. Dada uma fotografia de teste, esta é atribuída a média das coordenadas de latitude e longitude das fotografias da célula com a maior probabilidade.

Os mesmos autores introduziram mais 4 extensões de modo a melhorar este modelo de linguagem, da seguinte forma:

- **External data**: Nesta extensão, os autores adicionaram um total de 90 milhões de fotografias aos dados de treino, a fim de melhorar a qualidade do modelo de linguagem inicial, o qual foi construído usando cerca de $8,5$ milhões de fotografias.

- **Machine tags processing:** Esta extensão usa as tags produzidas automaticamente por máquinas (e.g., "geo: long=123.456") para aprender as suas correlações para com as suas coordenadas correctas nos dados de treino, de modo a fornecer uma informação mais precisa de localização das fotos. Sempre que uma fotografia tiver as tags de máquinas associadas, em vez de se usar o modelo de linguagem, os autores exploram as correlações aprendidas para estimar as coordenadas.

- **Geographicity:** Nesta extensão é estimada o valor da geograficidade (*geo*) de cada uma das tags associadas às fotografias de treino. O valor da geo varia entre 0 (não discriminante) a 1 (total discriminante) e, é obtido através do cálculo da probabilidade de uma determinada tag aparecer nas células vizinhas da sua célula mais provável, calculado a partir do modelo de linguagem. Sempre que para todas as tags de uma dada fotografia o valor máximo $geo <= 0,2$, então é usada uma modelagem de utilizadores descrita abaixo.

- **User modeling:** Finalmente, com o intuito de complementar o modelo de linguagem sempre que a extensão acima tende em falhar, os autores adicionaram várias fotografias de todos os autores das fotografias de teste nos dados de treino. Em seguida, usaram o valor da geograficidade em conjunto com os meta-dados temporais da seguinte forma. Se duas fotografias foram tiradas pelo mesmo autor num intervalo temporal de 24 horas, e têm uma diferença entre os seus valores da geograficidade de pelo menos $0,2$, então são transferidas as coordenadas da foto com a maior geograficidade para a outra.

Os autores testaram o método base, comparando-o com a adição de cada uma das extensões propostas. Os melhores resultados foram alcançados com a combinação do modelo de linguagem e de user modeling.

3.9 Identificação da Localização de uma Imagem com uma Função de Densidade de Probabilidade Multimodal

Davies et al. desenvolveram um método para atribuir coordenadas geográficas às fotografias do *Flickr*, usando as características visuais e textuais [Davies et al., 2013]. A ideia base consiste em estimar uma função de densidade de probabilidade (*PDF*) contínua ao longo da superfície da Terra usando um conjunto de pontos (i.e., latitude e longitude), obtidos a partir das características visuais e textuais associadas às fotografias. Neste trabalho, os autores usaram quatro características distintas, e cada uma fornece um número fixo de pontos em função da fotografia de teste, obtidos da seguinte forma:

- *Location prior*: Uma característica constante construída inicialmente pela recolha de 1000 coordenadas a partir dos dados do treino.

- *Tags*: Cada tag de uma dada fotografia de teste é associada as coordenadas geográficas das fotografias de treino em que ela aparece. Se a uma tag da fotografia de teste não ocorrer nos dados de treino, então ela não contribui com nenhum ponto.

- *PQ-CEDD*: Os autores construíram um índice a partir das características visuais *CEDD* usando uma abordagem baseada em uma quantização dos *vectores de características* [Jegou et al., 2011], de modo a proporcionar uma pesquisa eficiente pelos vizinhos mais próximos como também uma baixa precisão e uma alta *recall*. Dada uma fotografia de teste, o seu conjunto de pontos será constituído pelas coordenadas das 100 fotografias mais similares.

- *LSH-SIFT*: Com o objectivo de obter uma alta precisão e uma baixa *recall*, os mesmos autores construíram ainda um outro índice, sendo que desta vez foi utilizado o descritor *SIFT* e uma variante da abordagem proposta no trabalho de Hare et al. [2013]. Inicialmente, os vectores de características da *SIFT* são indexados usando a técnica *LSH*. Em seguida, é construído um grafo na qual cada nó corresponde a uma foto, e entre cada par de nós existe uma aresta com um peso com base no número de colisões. Finalmente, para uma determinada fotografia de teste, será retornada as coordenadas de todos os nós conectados directamente com a mesma.

A estimativa das coordenadas geográficas de uma dada fotografia de teste é efectuada a partir da localização do melhor modo de *PDF*, i.e., a localização em que a *PDF* tem o valor mais elevado. Em particular, estes autores utilizaram o algoritmo de agrupamento *mean shift* em conjunto com uma estrutura *KD-Tree* para acelerar o processo da procura do vizinho mais próximo para associar pontos a localizações. Desta forma, os pontos pertencem ao mesmo *cluster* só se eles convergem para a mesma localização. Os resultados experimentais mostraram que, usando os conteúdos textuais em conjunto com os conteúdos visuais, obtêm-se piores resultados em comparação com o uso de apenas os conteúdos textuais.

Trabalho	Características			Fontes de Conhecimento Externas
	Textuais	Visuais	Temporais	
Hays and Efros [2008]		✓		
Serdyukov et al. [2009]	✓			✓
Crandall et al. [2009]	✓	✓	✓	
Van Laere et al. [2010]	✓			✓
Cao et al. [2012]		✓		
O'Hare and Murdock [2012a]	✓			
Hauff and Houben [2012]	✓			✓
Popescu [2013]	✓		✓	✓
Davies et al. [2013]	✓	✓		✓

Tabela 3.1: Comparação entre os trabalhos relacionados.

3.10 Sumário

Este capítulo apresentou alguns dos trabalhos anteriores mais importantes focados na georreferenciação automática de foto. Em alguns dos trabalhos apresentados, os autores exploraram modelos de linguagem, isoladamente ou em conjunto com fontes de conhecimento externas. Noutros trabalhos, os autores exploraram as principais características visuais de uma imagem. A Tabela 3.1 apresenta uma comparação entre os trabalhos relacionados que foram apresentados. A maioria dos trabalhos apresentados baseia-se no uso de modelos de linguagem. O nosso trabalho basea-se numa interpolação das coordenadas geoespaciais associadas as imagens que partilham *tags*.

Capítulo 4

Atribuição de Coordenadas Geoespaciais a Fotos

Neste capítulo, apresentamos uma abordagem simples para a atribuição automática de coordenadas geoespaciais a novas fotos não georreferenciadas, publicadas em serviços como o *Flickr* e usando as *tags* descritas como a principal fonte de evidência. As coordenadas geoespaciais são atribuídas com base numa interpolação das coordenadas geográficas associadas às fotografias de treino, as quais partilhem *tags* com a foto de teste.

4.1 O Método Proposto

A abordagem proposta para a georreferenciação automática de fotografias é baseada na ideia da pesquisa pelas coordenadas geoespaciais das fotos incluídas nos dados de treino que compartilham *tags* com a foto de teste. Em seguida, atribuímos as coordenadas geográficas à foto de teste através do cálculo do ponto médio geográfico ponderado das coordenadas associadas a estas fotos, tendo como pesos os valores de importância das *tags*. Esta proposta envolve duas etapas, nomeadamente um processo da indexação e um processo de georreferenciação, apresentados em seguida.

4.1.1 Indexação dos Dados de Treino

O processo de indexação consiste em indexar os meta-dados de uma colecção de fotografias georreferenciadas, disponibilizadas no contexto do *MediaEval* 2013, da seguinte forma.

Começamos por construi um vocabulário a partir de todas as *tags* presentes nos dados de treino. Durante este processo, usamos uma lista de *stop-words* inglesa[1], e algumas expressões regulares (e.g., padrões envolvendo nomes de máquinas fotográficas) para remover as *tags* que não identificam nenhuma localização. As versões maiúsculas das *tags* originais estão também incluídas no vocabulário.

[1] `http://www.ranks.nl/stopwords`

Cada *tag* no vocabulário está associada a um peso que indica a sua importância ao descrever uma determinada localização. Os valores dos pesos de cada tag podem ser definidos de diversas formas. Para a presente dissertação, experimentamos três diferentes abordagens:

- *Inverse tag frequency* (*itf*) numa colecção de fotografias, tentando capturar o poder discriminativo de cada *tag* de modo a determinar localizações geográficas especifica, através da intuição de que as *tags* que aparecem com muita frequência na colecção provavelmente são menos discriminativas. Esta abordagem é semelhante à heurística da frequência inversa em documentos no contexto da área de *RI* [Manning et al., 2008], sendo definida pela seguinte equação, onde N é o número total de fotografias da colecção, e onde I_t corresponde ao número de fotografias que contêm a *tag* t:

$$itf_t = 1 + \log\left(\frac{N}{I_t}\right)$$

- *Inverse area frequency* (*iaf*) de um polígono calculado a partir de um conjunto de pontos correspondentes às coordenadas geoespaciais associadas a todas as fotos que contêm uma determinada *tag*. De forma análoga ao método *itf*, este método é baseado na ideia de que *tags* associadas a uma pequena área geográfica são provavelmente as mais discriminativos na tarefa de georreferenciação. O seu valor é definido pela seguinte equação, onde A corresponde ao valor da área do polígono obtido a partir das coordenadas geográficas associadas a todas as fotografias que contêm a *tag* t:

$$iaf_t = \frac{1}{1 + \sqrt{A_t}}$$

- Uma combinação linear entre itf_t e iaf_t, dada pela seguinte equação:

$$comb_t = itf_t \times iaf_t$$

No âmbito deste trabalho, a área geométrica associada à uma *tag* t é delimitada por uma *bounding box* (i.e., uma menor área rectangular contendo os pontos), um polígono convexo (i.e., *convex-hull*, ou seja um menor polígono convexo contendo os pontos), ou um polígono côncavo (i.e., *concave-hull*, ou seja um menor polígono côncavo contendo os pontos). O cálculo das últimas duas técnicas são detalhadas na Secção 4.1.4.

Além de indexar o vocabulário (i.e., as associações entre *tags* e pesos), construímos também outros índices associando *tags* às suas fotos de treino correspondentes, associando fotos de treino às suas coordenadas geoespaciais correspondentes, associando os utilizadores às suas fotos correspondentes, e por fim, associando fotos aos seus respectivos utilizadores.

4.1.2 Georreferenciação Usando Índices

Este processo consiste em atribuir automaticamente as coordenadas geográficas a uma dada foto de teste, usando os índices descritos acima de modo obter uma lista das coordenadas geoespaciais as-

sociadas às fotografias que partilhem pelo menos uma *tag* com a foto de teste. As coordenadas geográficas são atribuídas a partir de uma interpolação das coordenadas desta lista.

Inicialmente, construímos um vocabulário V com todas as *tags* presentes nos dados de treino, calculando os valores dos pesos através do método itf. Dada uma fotografia de teste, com as *tags* T de um utilizador U, é efectuada uma pesquisa pelas imagens similares dependendo do método do cálculo do peso usado. Se usarmos o método itf, inicialmente esta pesquisa é realizada usando apenas T. Caso não seja encontrada nenhuma imagem similar, então usamos as *tags* adicionais associados a todas as fotos do utilizador U.

Já ao usarmos os métodos iaf ou o $comb$, primeiro aplicamos um filtro ao vocabulário V, e o mesmo exclui todas as *tags* que ocorrem em uma área maior do que um determinado limiar pré-definido (A_{max}). Após a aplicação deste filtro são gerados dois conjuntos com pares de *tag* e de peso correspondente, nomeadamente um com todas as *tags* do vocabulário (VA), e o outro resultante da aplicação do filtro (VF). Em seguida, usando VF e T, é efectuada uma pesquisa pelas fotografias similares. Se este processo ainda não retornar imagens similares, é realizada uma nova pesquisa pelas imagens similares, sendo que desta vez é usado o conjunto VA. Se este processo continua ainda sem retornar imagens similares, usamos todas as *tags* associadas as fotografias do utilizador U.

Finalmente, é realizada uma interpolação a partir de uma lista de pares de coordenadas polares e do valor do pesagem correspondente à *tag* em comum. Esta interpolação é baseada no cálculo do ponto médio geográfico a partir de um conjunto de coordenadas [Jenness, 2008], tal como descrito na Secção 4.1.3.

Esta abordagem também pode ser adaptada de modo a usar os conteúdos visuais das imagens, envolvendo as seguintes etapas:

1. Seleccionar um dos métodos para calculo dos valores dos pesos das *tags* das fotografias de treino, de forma a gerar os vocabulários necessários, tal como descrita anteriormente.

2. Usar os índices para obter uma lista de pares de coordenadas polares e de valor do peso da *tag*, associadas às fotografias que partilhem *tags* com a fotografia de teste.

3. Para cada elemento da lista obtido na Etapa 2, substituir o valor do peso da *tag* em comum pelo valor da similaridade dos conteúdos visuais para com a fotografia de teste.

4. Finalmente, interpolar as coordenadas da lista gerada na Etapa 3, também com base na abordagem detalhada na Secção 4.1.3.

Dado que o tempo e o custo computacional necessário para o processamento de todos os descritores visuais são elevados, neste trabalho, usamos apenas o descritor *Simple Color Histogram*. Além disso, em vez de usarmos as coordenadas geoespaciais associadas à todas as fotografias que partilham *tags* com a fotografia de teste, obtidas na Etapa 2, usamos no máximo as 100 coordenadas associadas às fotografias com valores dos pesos das *tags* mais elevados.

4.1.3 Interpolação das Coordenadas

Nesta Secção, apresentamos o processo utilizado para o cálculo do ponto médio geográfico de um conjunto de coordenadas baseado no trabalho de Jenness [2008].

Seja uma fotografia de teste representada como $I_{test} = (t_1, w_1), \ldots, (t_n, w_n)$, com n *tags* respectivamente, onde cada w_i representa o peso associado à *tag* t_i. Podemos definir um conjunto formado por k pares de coordenadas polares (i.e., latitude e longitude) e de valor de peso, denotada por $C = \{(c_1, w_1), \ldots, (c_k, w_k)\}$. O parâmetro w_i representa o valor da similaridade/peso da *tag* partilhada, entre a imagem associada às coordenadas geoespaciais c_i e uma imagem de interrogação. O processo da interpolação envolve as seguintes etapas:

1. Converter cada uma das coordenadas polares das k imagens mais similares em coordenadas cartesianas (i.e., X, Y, Z):

$$X = \cos(latitude \times \frac{\pi}{180}) \cdot \cos(longitude \times \frac{\pi}{180})$$

$$Y = \cos(latitude \times \frac{\pi}{180}) \cdot \sin(longitude \times \frac{\pi}{180})$$

$$Z = \sin(latitude \times \frac{\pi}{180})$$

2. Calcular a média ponderada das coordenadas cartesianas, na qual cada uma das componentes é influenciada pelo valor da respectiva similaridade/peso w:

$$X_c = \frac{1}{\sum_{i=1}^{k} w_i} \cdot \sum_{i=1}^{k}(x_i \times w_i), \quad Y_c = \frac{1}{\sum_{i=1}^{k} w_i} \cdot \sum_{i=1}^{k}(y_i \times w_i), \quad Z_c = \frac{1}{\sum_{i=1}^{k} w_i} \cdot \sum_{i=1}^{k}(z_i \times w_i)$$

3. Projectar a média das coordenadas cartesianas obtidas na Etapa 2 para a superfície da Terra, i.e., convertendo-as novamente em coordenadas polares:

$$\text{latitude} = \arctan 2(Z_c, \sqrt{X_c^2 + Y_c^2}) \cdot \frac{180}{\pi}, \quad \text{longitude} = \arctan 2(Y_c, X_c) \cdot \frac{180}{\pi}$$

Por fim, as coordenadas geoespaciais retornadas na Etapa 3 são então atribuídas à foto de interrogação.

4.1.4 Cálculo do *Convex-Hull* e *Concave-Hull*

O cálculo do *convex-hull* e *cancave-hull* para um conjunto finito de pontos é um processo fundamental para diversas aplicações em diferentes áreas, especialmente em geometria computacional [Vishwanath et al., 2013]. Estas técnicas podem ser usadas para definir a forma de um conjunto de pontos ou encontrar uma área mínima que envolve o conjunto. Muitos algoritmos têm sido propostos para calcular o *convex-hull* [O'Rourke, 1998]. No âmbito do presente trabalho, o cálculo do *convex-hull* é efectuado da seguinte forma. Inicialmente, é efectuada uma triangulação do conjunto de pontos, resultando assim num conjunto de triângulos, os quais constituem um polígono convexo. O processo da triangulação

foi efectuada utilizando uma biblioteca desenvolvida em Python, denominada de *matplotlib*[2], na qual implementa vários algoritmos da área de geometria computacional. Em seguida, é calculada a área de cada um dos triângulos através das fórmulas de Hero's[3], tal como ilustrada na equação seguinte, onde a, b e c são os comprimentos dos lados, e o parâmetro s corresponde ao semiperímetro, i.e., a metade o perímetro definido por $s = \frac{(a+b+c)}{2}$:

$$A = \sqrt{s(s-a)(s-b)(s-c)}$$

Já o cálculo do *concave-hull* foi baseado numa generalização do *convex-hull*, denominada de *aplha-convex-hull* [Vishwanath et al., 2013]. A ideia base deste processo consiste em primeiro construir um polígono convexo tal como descrito anteriormente e, em seguida, remover os triângulos estreitos nos extremos deste polígono, convertendo-o assim num polígono côncavo. Os triângulos são removidos de acordo com uma restrição estabelecida ao raio do círculo circunscrito (i.e., o círculo que passa pelas três vértices do triângulo) de cada triângulo, dada pela seguinte fórmula:

$$\frac{abc}{4 \cdot A} < \frac{1}{\alpha} \tag{4.1}$$

Na fórmula acima, o parâmetro α é um numero real positivo maior do que zero, e quanto maior for o seu valor menor será a área do polígono. Desta forma, a área do *concave-hull* é definida pela a soma das áreas de todos os triângulos que satisfaçam a restrição definida na Equação 4.1.

4.2 Sumário

Nesta dissertação, propomos uma solução simples para a atribuição automática de coordenadas geo-espaciais, usando as *tags* como a principal fonte de evidência. Esta solução atribui pesos às *tags* de acordo com a frequência inversa da *tag* nos dados de treino, ou de acordo com a área geométrica de uma localização que abrange um conjunto de fotografias. A georreferenciação é efectuada a partir de uma interpolação das coordenadas associadas a todas as imagens que partilhem pelo menos uma *tag* com a foto a ser georreferenciada.

[2] `http://matplotlib.org/`
[3] `http://en.wikipedia.org/wiki/Heron%27s_formula/`

Capítulo 5

Validação Experimental

Neste capítulo, é apresentada a metodologia de validação usada nas experiências, discutindo os resultados obtidos pelos métodos propostos. Inicialmente, são descritas as colecções de fotografias de treino e de teste usadas nas experiências. Em seguida, são apresentadas as métricas usadas para quantificar a precisão dos resultados obtidos pela nossa proposta. Finalmente, são descritos os testes experimentais e os seus respectivos resultados, e é também feita uma comparação com alguns dos melhores resultados dos trabalhos propostos no contexto do *MediaEval* 2013.

5.1 Conjunto de Dados Usado nos Testes

Como dados de treino, usamos uma grande colecção de fotografias disponibilizada no *MediaEval* 2013 [Hauff et al., 2013]. O *MediaEval* é uma iniciativa de *benchmarking* dedicada a avaliar novos algoritmos para o acesso e a recuperação de informação multimédia[1]. Nesta colecção, estão incluídas cerca de 9 milhões de fotografias publicadas no *Flickr*[2], das quais $8.539.050$ estão associadas a coordenadas de latitude e longitude. Para além disso, esta colecção contém meta-dados descritivos, entre os quais se incluem características derivadas de anotações textuais e de 13 descritores para os conteúdos visuais das fotos. Todas as fotografias foram anotadas com a mais alta precisão do *Flickr* (i.e., no *Flickr*, existem diferentes níveis de precisão, sendo que o maior é de 16, o que significa que o local é preciso no nível da rua). Para validar os métodos propostos, usamos uma colecção de teste com um total de 262 mil fotografias, a qual foi disponibilizada também no âmbito do *MediaEval* 2013.

Antes de realizar a avaliação experimental, removemos todas as fotografias que têm pelo menos um descritor visual inválido (e.g., características representadas por caracteres alfabéticos ao invés de números nos descritores visuais), produzindo assim uma colecção com um total de $8.538.096$ fotografias para os dados de treino. A Tabela 5.1 apresenta uma caracterização estatística para o conjunto de dados considerados.

[1] `http://www.multimediaeval.org/`
[2] `http://flickr.com/`

Estatística	Treino	Teste
Num. Fotos	8.538.096	262.000
Num. Fotos sem *Tags*	1.271.347	35.052
Num. *Tags*	59.127.129	1.756.770
Avg. *Tags* por Foto	8,137	7,741
St.Dev. *Tags* por Foto	8,097	6,955

Tabela 5.1: Caracterização estatística da coleccão de dados usada na avaliação experimental.

		iaf			comb		
Precisão	itf	boundingBox	convexHull	concaveHull	itf-boundingBox	itf-convexHull	itf-concaveHull
1	0,44	**10,62**	10,02	2,22	**10,62**	9,80	2,28
10	1,50	**21,01**	20,65	6,16	20,91	20,34	6,30
100	5,65	30.68	**31,64**	14,10	30,54	30,98	14,30
Média	3183	2777	**2348**	3046	2762	2318	3034
Mediana	2511	1096	**558**	1969	1106	629	1951

Tabela 5.2: Resultados da georreferenciação de fotos usando diferentes técnicas para o cálculo do peso das *tags*.

5.2 Metodologia de Avaliação

Para a validação dos resultados foi usada a média e a mediana das distâncias geoespaciais entre as coordenadas retornadas pelo método proposto e as coordenadas indicadas na colecção *MediaEval* 2013. As distâncias geoespaciais foram calculadas através das *fórmulas de Vincenty*[3]. Um valor indicativo da precisão foi também determinado através do número de atribuições correctas de coordenadas, considerando um erro na distância abaixo de 1, 10 ou 100 km.

5.3 Resultados e Discussão

Com o intuito de validar a nossa proposta, realizamos uma avaliação experimental com diferentes tipos de configurações em termos da técnica usada para calcular o peso das *tags*.

A Tabela 5.2 apresenta os resultados para os diferentes métodos do cálculo dos valores dos pesos das *tags*, e usando o parâmetro $A_{max} = 62.500$ respectivamente. Os resultados mostraram que o método correspondente à utilização da técnica *convex-hull* produziu melhores resultados, com uma melhor precisão de cerca de $31,64\%$ na tarefa de encontrar as coordenadas correctas considerando um erro de distância abaixo de 100 km. Considerando apenas estas fotos (i.e., cerca de $31,64\%$ deles), produz uma distância média para com as coordenadas correctas de apenas $16,07$ km. A melhor atribuição de coordenadas geoespaciais também foi alcançada com a configuração correspondente ao uso do método *convex-hull*, obtendo um erro de 2348 km em termos da média, e de 558 km em termos da mediana.

Na Figura 5.1, apresentamos em escala logarítma a distribuição dos erros produzidos pela configuração corresponde ao uso do método *convex-hull*, em termos da distância entre as coordenadas estimadas e as coordenadas reais geográficas. De acordo com esta figura, podemos observar que este método atribui coordenadas para a maioria dos exemplos com um pequeno erro em termos de distância, muito

[3] `http://en.wikipedia.org/wiki/Vincenty's_formulae/`

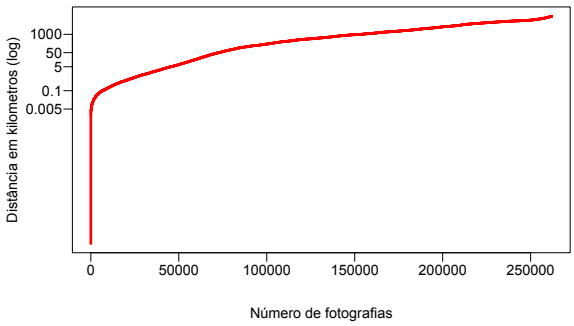

Figura 5.1: Gráficos de distribuição dos erros usando a técnica *convex-hull* no cálculo dos pesos das *tags*.

		iaf			comb		
	itf	boundingBox	convexHull	concaveHull	itf-boundingBox	itf-convexHull	itf-concaveHull
Média	2947	1471	**1295**	2848	1472	1412	2838
Mediana	2282	**24**	29	1731	24	44	1723

Tabela 5.3: Resultados da georreferencição de fotos usando diferentes técnicas para o cálculo do peso das *tags*, e apenas as *tags* da foto georreferenciada.

embora cerca de 130.000 fotografias tenham um erro maior do que 500 km.

Na Tabela 5.3, apresentamos os resultados correspondentes ao uso de diferentes técnicas para o cálculo do peso das *tags*, usando apenas o vocabulário V ou VF, i.e, sem usar as *tags* adicionais do mesmo autor ou as que ocorrem numa região geográfica com área superior a A_{max}. Pelos valores desta tabela, observa-se que os resultados são muitos melhores tanto em termos da média como também da mediana, em comparação com o uso das informações adicionais.

Na Tabela 5.4, apresentamos uma comparação entre o método correspondente ao uso da técnica *convex-hull* e o método correspondente ao uso dos descritores visuais, especificamente o descritor *Simple Color Histogram*. Pela análise dos resultados, pode-se observar que não houve uma melhoria significativa em comparação com o uso apenas das contribuições das *tags*.

A Tabela 5.5 apresenta um resumo dos nossos melhores resultados, obtidos a partir do método correspondente ao uso de da técnica *convex-hull*, em comparação com os resultados obtidos pelos melhores trabalhos propostos no âmbito do MediaEval 2013, especificamente nos trabalhos de Davies et al. [2013] e de Popescu [2013], respectivamente. Nestas experiências, foram usados apenas os conteúdos textuais, e os mesmos conjuntos de dados. Os resultados mostraram que não temos uma grande margem de diferença em termos dos valores de precisão em comparação com os melhores trabalhos propostos no âmbito do *Medieval* 2013.

	iaf	
Precisão	convexHull	visual features
1	**10,83**	9,37
10	**21,05**	19,62
100	30,35	**32,27**
Média	**2453**	2700
Mediana	**747**	880

Tabela 5.4: Comparação entre a técnica convex-hull e a técnica correspondente ao uso do descritor visual *Simple Color Histogram*.

Abordagem	1 Km	10 Km	100 Km	Median Km
Método Proposto	10,02	21,65	31,64	558,340
Davies et al. [2013]	22,97	37,42	43,49	451,89
Popescu [2013]	**26,84**	**42,77**	**50,04**	**98,71**

Tabela 5.5: Comparação entre os resultados da nossa proposta e os trabalhos de Davies et al. [2013] e de Popescu [2013].

5.4 Sumário

Neste capítulo, descrevemos a avaliação experimental da nossa proposta para a atribuição automática de coordenadas geográficas a fotografias publicadas no *Flickr*, usando as *tags* como a principal evidência de suporte. De uma forma geral, os resultados experimentais mostraram que usando a técnica *convex-hull* obteve-se melhores resultados, em comparação com os resultados obtidos a partir das outras técnicas. A melhor precisão, em termos do número de atribuições correctas de coordenadas, foi atingida quando consideramos um erro de distância abaixo de 100 km. A melhor atribuição de coordenadas geoespaciais foi alcançada com a configuração correspondente ao uso da técnia *convex-hull*, obtendo um erro de 2348 km em termos da média, e de 558 km em termos da mediana.

Capítulo 6

Conclusões e Trabalho Futuro

Nesta dissertação, avaliamos uma nova técnica para a atribuição automática de coordenadas geoespaciais de latitude e de longitude a novas fotografias, usando as *tags* associadas às mesmas como a principal evidência de suporte. Os resultados experimentais mostraram que a identificação automática da localização geoespacial de uma fotografia, com base apenas em suas *tags*, pode ser efectuada com uma alta precisão utilizando um método relativamente simples, que é também computacionalmente eficiente.

6.1 Sumário das Contribuições

As principais contribuições do trabalho desenvolvido são as seguintes:

- Desenvolvimento de uma técnica eficiente para estimar as coordenadas geoespaciais de latitude e longitude a fotos, usando uma interpolação das coordenadas associadas às fotografias com *tags* em comum.

- Comparação entre diferentes técnicas para determinar a importância das *tags*, concluindo que os melhores resultados são alcançados com o uso da técnica *convex-hull* para calcular os pontos correspondentes à localização das fotos associadas à *tag*.

6.2 Trabalho Futuro

Para trabalho futuro, gostaríamos de experimentar com o uso de métodos de detecção de *outliers* espaciais, antes do cálculo dos pesos associados às *tags*, ou antes da selecção das coordenadas para a interpolação. Também gostaríamos de experimentar com o uso de métodos de agrupamento de palavras, por exemplo, o algoritmo de agrupamento k-*means*, em conjunto com as representações de palavras produzidas pela ferramenta *word2vec*[1]. Desta forma, complementaríamos os nossos vocabulários

[1] `https://code.google.com/p/word2vec/`

com outras representações que capturam a semelhança entre o conteúdo sintáctico e semântico de uma *tag*, e o resto do das *tags* na colecção.

Bibliografia

R. A. Baeza-Yates and B. Ribeiro-Neto. *Modern Information Retrieval*. Addison-Wesley Longman Publishing Co., Inc., 1999.

L. Cao, J. R. Smith, Z. Wen, Z. Yin, X. Jin, and J. Han. Bluefinder: estimate where a beach photo was taken. *Proceedings of the 21st International Conference on World Wide Web*, 2012.

S. K. Chang, C. W. Yan, D. C. Dimitroff, and T. Arndt. An Intelligent Image Database System. *IEEE Transactions on Software Engineering*, 14(15), 1988.

D. Comaniciu and P. Meer. Mean shift: A robust approach toward feature space analysis. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 24(5), 2002.

D. J. Crandall, L. Backstrom, D. Huttenlocher, and J. Kleinberg. Mapping the world's photos. *Proceedings of the 18th International Conference on World Wide Web*, 2009.

A. A. E. D. Hoiem and M. Hebert. Geometric context from a single image. *IEEE International Conference on Computer Vision*, 2005.

J. Davies, J. S. Hare, S. Samangooei, J. Preston, N. Jain, D. Dupplaw, and P. H. Lewis. Identifying the Geographic Location of an Image with a Multimodal Probability Density Function. In *Proceedings of the MediaEval 2013 Multimedia Benchmark Workshop*, 2013.

H. Ewald Hering, L. Hurvich, and D. Jameson. *Grundzüge Der Lehre Vom Lichtsinn. Outlines of a Theory of the Light Sense ... Translated by Leo M. Hurvich and Dorothea Jameson. With Illustrations.* Harvard University Press, 1964.

A. G. Glen. Maximum likelihood estimation using probability density functions of order statistics. *Computers and Industrial Engineering*, 58(4), 2010.

R. C. Gonzalez and R. E. Woods. *Digital Image Processing*. Addison-Wesley Longman Publishing Co., Inc., 2nd edition, 2001.

V. N. Gudivada and V. V. Raghavan. *Finding the Right Image: Content-based Image Retrieval Systems*. IEEE Computer Society Press, 1995.

J. S. Hare, S. Samangooei, D. P. Dupplaw, and P. H. Lewis. Twitter's Visual Pulse. In *Proceedings of the 3rd ACM Conference on International Conference on Multimedia Retrieval*, 2013.

C. Hauff and G.-J. Houben. Geo-location estimation of Flickr images: social web based enrichment. *Proceedings of the 34th European Conference on Advances in Information Retrieval*, 2012.

C. Hauff, B. Thomee, and M. Trevisiol. Working Notes for the Placing Task at MediaEval 2013. In *Proceedings of the MediaEval 2013 Multimedia Benchmark Workshop*, 2013.

J. Hays and A. A. Efros. Im2GPS: estimating geographic information from a single image. *Proceedings of the 8th IEEE Conference on Computer Vision and Pattern Recognition*, 2008.

P. Howarth and S. Ruger. Fractional distance measures for content-based image retrieval. *Proceedings of the 27th European Conference on Advances in Information Retrieval*, 2005.

H. Jegou, M. Douze, and C. Schmid. Product Quantization for Nearest Neighbor Search. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 33(1), 2011.

J. Jenness. Calculating areas and centroids on the sphere. *Proceedings of the 28th ESRI International User Conference*, 2008.

T. Joachims. *Advances in kernel methods*. Massachusetts Institute of Technology (MIT) Press, 1999.

J. Kosecka and W. Zhang. Video Compass. *Proceedings of the 7th European Conference on Computer Vision*, 2002.

D. G. Lowe. Object recognition from local scale-invariant features. *Proceedings of the 7th the International Conference on Computer Vision*, 1999.

C. D. Manning, P. Raghavan, and H. Schutze. *Introduction to Information Retrieval*. Cambridge University Press, 2008.

D. Martin, C. Fowlkes, D. Tal, and J. Malik. A database of human segmented natural images and its application to evaluating segmentation algorithms and measuring ecological statistics. *IEEE International Conference on Computer Vision*, 2001.

N. O'Hare and V. Murdock. Gender-based models of location from Flickr. *Proceedings of the 20th ACM Multimedia Workshop on Geotagging and Its Applications in Multimedia*, 2012a.

N. O'Hare and V. Murdock. Modeling locations with social media. *Information Retrieval*, 2012b.

A. Oliva and A. Torralba. Building the gist of a scene: The role of global image features in recognition. *Progress in Brain Research*, 2006.

J. O'Rourke. *Computational Geometry in C*. Cambridge University Press, 2nd edition, 1998.

F. Peng, D. Schuurmans, and S. Wang. Augmenting naive bayes classifiers with statistical language models. *Information Retrieval*, 7(3-4), 2004.

J. M. Ponte and W. B. Croft. A language modeling approach to information retrieval. *Proceedings of the 21st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, 1998.

A. Popescu. CEA LIST's Participation at MediaEval 2013 Placing Task. In *Proceedings of the MediaEval 2013 Multimedia Benchmark Workshop*, 2013.

P. Serdyukov, V. Murdock, and R. van Zwol. Placing Flickr photos on a map. *Proceedings of the 32nd International ACM SIGIR Conference on Research and Development in Information Retrieval*, 2009.

G. Sharma. *Digital Color Imaging Handbook*. Chemical Rubber Company Press, Inc., 2002.

J. Shi and J. Malik. Normalized cuts and image segmentation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 22(8), 2000.

W. Sproson. *Colour Science in Television and Display Systems*. Adam Hilger Limited, 1983.

A. Torralba, R. Fergus, and W. T. Freeman. Tiny images. *Technical Report MIT-CSAIL-TR-2007-024*, 2007.

K. Toyama, R. Logan, and A. Roseway. Geographic location tags on digital images. *Proceedings of the 11th ACM International Conference on Multimedia*, 2003.

M. Tuceryan and A. K. Jain. *Texture Analysis*. The Handbook of Pattern Recognition and Computer Vision, 2nd edition, 1998.

K. van de Sande, T. Gevers, and C. Snoek. Evaluating color descriptors for object and scene recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 32(9), 2010.

O. Van Laere, S. Schockaert, and B. Dhoedt. Towards automated georeferencing of Flickr photos. *Proceedings of the 6th Workshop on Geographic Information Retrieval*, 2010.

A. V. Vishwanath, R. Arun Srivatsan, and M. Ramanathan. Minimum Area Enclosure and Alpha Hull of a Set of Freeform Planar Closed Curves. *Computer-Aided Design*, 45(3), 2013.

C. Zhai and J. Lafferty. Two-stage language models for information retrieval. *Proceedings of the 25th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, 2002.