BUSCA VISUAL EM UM ESPAÇO SEMÂNTICO: UMA ESCOLHA ENTRE IDENTIDADE E POPULARIDADE

MARIANE MOREIRA DE SOUZA

BUSCA VISUAL EM UM ESPAÇO SEMÂNTICO: UMA ESCOLHA ENTRE IDENTIDADE E POPULARIDADE

Tese apresentada ao Programa de Pós-Graduação em Ciência da Computação do Instituto de Ciências Exatas da Universidade Federal de Minas Gerais como requisito parcial para a obtenção do grau de Doutor em Ciência da Computação.

ORIENTADOR: ADRIANO VELOSO

Belo Horizonte

Fevereiro de 2018

MARIANE MOREIRA DE SOUZA

FASHION RETRIEVAL IN A SEMANTIC SPACE: BALANCING IDENTITY AND FASHIONABILITY

Thesis presented to the Graduate Program in Computer Science of the Federal University of Minas Gerais in partial fulfillment of the requirements for the degree of Doctor in Computer Science.

Advisor: Adriano Veloso

Belo Horizonte

February 2018

Abstract

Online social networks, such as Facebook and Instagram, are becoming major sources of clothing inspiration, with users sharing their latest outfits and looking for others with similar apparel style. This typical behavior made possible the emergence and popularization of fashion bloggers, considered, today, the great dictators of fashion trends, with wide audiences searching for looks that are in conformity with their fashion sense. However, a substantial time is generally spent searching for specific looks in fashion blogs or social networks. The main problem we investigate in this thesis is how to facilitate and improve the retrieval of relevant looks posted in fashion blogs. We tackle this problem by using a content-based retrieval (CBIR) approach – given a query image, we find images with similar meanings in a large database of images posted in online social networks.

In our solution, we approximate the meaning of an outfit through the pieces of clothes that compose it, using a Convolution Neural Network (CNN) for representation learning and classification. In few words, the CNN model takes as input the pixels of an image and transforms them into a multi-dimensional feature vector, where each dimension corresponds to the probability associated with the corresponding clothing item. Since the model has learned a representation for the images, we are able to compare them in the resulting semantic space. That is, given an arbitrary query image, the CNN model is able to predict the pieces of clothes of the look in that image, and the ranking model is able to retrieve a sorted group of images, from the most to the least similar images, considering the distance between their feature vectors in the semantic space.

When searching for looks, a user is, implicitly, searching for something that matches her or his identity. Besides, a user wants, most of times, to be inspired by looks with high levels of popularity in terms of fashion, i.e. fashionability. Considering the fact that identity and fashionability are, most of times, in non-conformity, this thesis also analyses the trade-off between these two concepts, in order to improve the results of the search, according to the user's needs. We produce a second ranking

function, considering the balancing of identity and fashionability, in which the user is able to prioritize the similarity of candidate images or their popularity in terms of fashion. In this analysis we also consider the variation of fashionability, according to the user's location, which reflects culture and lifestyle of the people.

The results achieved by this thesis show the improvement of the state-of-the-art in fashion retrieval and also show it is possible to build the balanced ranking with a little loss in terms of NDCG. The results also show the impact of culture and lifestyle in different countries, making it necessary that the ranking is composed with posts related to the same location of user's.

Keywords: visual search, fashion retrieval, CBIR, CNN, fashionability, fashion applications.

List of Figures

List of Tables

Contents

Chapter 1

Introduction

Online social networks, such as Facebook and Instagram, allow their users to express themselves in many different ways by creating and sharing content. A particular way of expression being increasingly adopted by members of these sites is to post photos that show their latest looks.[1] Typically, comments about the clothes appear shortly after the image is posted, showing that online social networks are becoming major sources of clothing inspiration (Lin et al. [2015]), with users looking for others with similar apparel style[2] and fashion sense,[3] usually to facilitate the choice of their own looks. This typical behavior made possible the emergence and popularization of fashion bloggers, considered, today, the great dictators of fashion trends (Eytan [2016], Sedeke [2012]), generally posting photos of looks with high levels of fashionability (Simo-Serra et al. [2015]).[4] As a result, there is an increasing number of fashion blogs and fashion profiles in social networks, with wide audiences searching for looks that are in conformity with their fashion sense.

Most of times, a substantial time is spent searching for specific outfits. Indeed, a user may navigate for hours, and there is no guarantee of finding the desired content, since fashion blogs and the corresponding networks usually present a huge amount of available data and no efficient way for users to find the information they want. Besides, it is a difficult task to define, precisely, the meaning of an outfit to be searched, since the main aspects we could consider to define it are all subjective, e.g. season, style and occasion (Lurie [2000]). Considering the aspect season or climate, for example, if there is an image showing that it is raining, how can we be sure it is hot or cold?

[1] The set of clothes and accessories that a person uses. Also known as outfit (Callan [2007]).

[2] The way someone uses to dress herself.

[3] The knowledge or expertise in fashion field.

[4] A quality of being well dressed, many times using clothes or accessories considered to be fashion trends. A fashion expert tends to dress looks with high levels of fashionability.

Figure 1.1. Semantic space - similar looks according to style

Also, different looks, with different colors, shapes, textures and accessories could be related to the same style, as shown in Figure 1.1. Thinking about the occasions, we have countless of them, making it impossible to connect specific types of looks for each occasion and season. Actually, these connections become even more complicated to be predicted when we analyze the cultural differences between countries, leading to myriad preferences of clothing style around the world.

1.1 The Problem

The main problem we investigate in this thesis is how to facilitate and improve the retrieval of relevant looks posted in fashion blogs. We tackle this problem by using a content-based retrieval (CBIR) approach – given a query image, we find images with similar meanings in a large database of images posted in fashion blogs. Also, when searching for relevant looks, it is common and crucial to consider the user's identity (i.e. user's preferences). Another important fact that must be considered is that when a person is looking for clothing inspiration, generally, he or she wants to be inspired by looks with high levels of fashionability (Simo-Serra et al. [2015], Lurie [2000]). In practice, however, user's identity and fashionability are a trade-off, that is, a user can provide an image of a look that encodes her preferences, but if this look has a similar meaning when compared to others with low levels of fashionability, there are chances

the results of the search will not be really satisfactory to the user. So, as another important contribution of this thesis, we investigate the relationship between identity and fashionability, aiming to balance them, and consequently improving the results according to the user's needs.

1.2 Our Solution

In view of the aforementioned challenges, we propose a solution to fashion retrieval which we briefly describe next.

1.2.1 Representation Learning: The Principle of Compositionality

We propose to approximate the meaning of a look through the pieces of clothes that compose it, thus based on the principle of compositionality.[5] Compositionality allows us to learn feature vectors for accurately representing outfits based solely on the occurrences of clothing items, and this has a fundamental motivation since it is relatively easy to obtain outfits labeled with their constituent items (e.g., hat, glasses, bag, pants, shoes and so on). By contrast, there may be debate on whether an outfit should be associated with a style or other, or if the outfit is suitable or not to certain occasions. Further, while low-level visual features, such as color, shape and texture, leave a lot to be desired when it comes to carry enough semantics to find outfits with similar meanings properly (Moreira et al. [2014]), compositionality allows us to match semantically close outfits that may not be visually similar, since no visual features are used to directly learn the appearance of outfits.

Convolutional neural networks (or simply CNNs) have long and widely been applied to object recognition in images (Krizhevsky et al. [2012a]). Still, recognizing clothing items and accessories in images is particularly hard. Clothing items and accessories are frequently subject to deformations and occlusion, to different lighting conditions, and often exhibit serious variations when they are taken under different scenarios. Thus, instead of recognizing clothing items, we employ a CNN model to learn outfit representations. That is, the CNN model takes as input the pixels of an image and transforms them into a multi-dimensional feature vector, where each dimension corresponds to the probability associated with the corresponding clothing item.

[5] In mathematics, semantics, and philosophy of language, the principle that says the meaning of a complex expression is determined by the meanings of its constituent expressions and the rules used to combine them.

After some optimization iterations the learned representation is expect to converge to the most likely probabilities for each clothing constituent. For instance, if an image is showing a person using pants, boots and a t-shirt, we expect the dimensions associated with these items will have a higher probability value than the others in the vector related to that image.

1.2.2 Ranking Fashion Looks in the Semantic Space

After representing looks through feature vectors, we are able to compare them using the resulting semantic space. In this thesis, we assume that relevance information is estimated using the context information, e.g. season, style and occasion related to a look. That is, when two looks share the style and are appropriate for the same season and occasion, it is a perfect match, and they are close to each other, considering the semantic space. By contrast, a totally irrelevant look does not share any of these attributes with the other. Then, given an arbitrary query image, the CNN model is able to predict the pieces of clothes of the look in that image, and the ranking model is able to retrieve a sorted group of images, from the most to the least similar images, considering the distance between their feature vectors in the semantic space.

1.2.3 Balancing Identity and Fashionability in a World-Wide Scale

When searching for looks, a user is, implicitly, searching for something that matches her or his identity. In this thesis, we consider identity as the personal choices related to colors, shapes, pieces of clothes and accessories that belong to the user, and can be inferred through the query image he or she provides. In few words, for a single search, the identity of a user can be estimated through the multi-dimensional feature vector generated by our CNN model, corresponding to the most relevant clothing items in the look presented in the query image.

Besides the identity matching, a user wants, most of times, to be inspired by looks with high levels of fashionability, since he or she generally wants to learn how to make good combinations of clothes and accessories or use outfits considered to be fashion trends. We may assess how fashionable is a look by taking into account the number of likes in its respective post or the number of followers related to the user who posted the look.

The problem, in this case, is that identity and fashionability are, most of times, in non-conformity. The ideal scenario is when the user is inspired by images of looks that

match, at a certain acceptable level, her or his identity and present high popularity in terms of fashion. Otherwise, fashionability can vary according to many aspects. We consider the user's location the most important aspect, which reflects culture and lifestyle. For instance, in Brazil, a casual dress, chosen to go to the church in a Sunday morning, would be considered a look with high fashionability, but in The United States, a formal dress would be a better choice for the same occasion. In this context, we decided to investigate the relationship between identity and fashionability, considering the difference between ranks built with posts from the same location of the user and without this concern, discovering, among others, why it is important to conduct the search by country. So, another contribution of this thesis is a good solution for a multi-objective function, considering the aspects identity and fashionability in a world-wide scale.

1.3 Thesis Statement

Fashion Retrieval is posed as a representation learning problem, in the sense that outfits can be placed in a semantic vector space, thus enabling the retrieval of semantically similar outfits. The main hypothesis of this thesis is that the principle of compositionality, which states that the meaning of a whole is a function of the meanings of its parts together with the manner in which these parts were combined, allows us to learn feature vectors for accurately representing outfits based solely on the occurrences of clothing items. The aim of this thesis is to build a visual search model which works by comparing outfits in the semantic space. We claim that our compositional approach, based on a deep CNN architecture, is a determining factor for improving representation learning, and thus, the retrieval effectiveness. We also claim that it is relevant to analyze the relation between fashionability and visual identity, aiming to detect behavior patterns and check the relevance of considering both variables during the search. Finally, we claim that this analysis should be conducted considering the user's location, in order to analyze the impact of culture and lifestyle of a country in the choice of looks.

1.4 Contributions

Some of the specific contributions of this thesis include:

- We represent outfits in a semantic level, following a compositional approach in which dimensions correspond to the likelihood of occurrence of clothing items.

A deep CNN model computes the probabilities for each clothing item. The final result is that outfits are placed on a semantic space, enabling the search for outfits that are semantically related.

- We formulate the search procedure as a simple multi-objective problem in which outfits are ranked based on a proper balance between visual identity and fashionability. The user may employ a control function in order to set the appropriate trade-off between these two objectives, and the final ranking will emphasize outfits that balance fashionability and visual identity.

- We built a new dataset for fashion retrieval. Images of fashion looks were collected from a fashion social network called Chictopia.[6] Chictopia is a fashion social network founded in 2008, that has a growing base of 1.5MM visitors and 13MM page views monthly. It is a platform for fashion bloggers to share their looks for inspiration seekers and for brands to sell their products. We collected approximately 68,000 fashion images along with information such as the clothing items that compose the look, user location, number of likes and followers, season, style and occasion.

- The world-wide analysis of two important concepts related to the fashion area: identity and fashionability, aiming to discover a configuration of values that meets the users' needs, improving the final ranking.

- We conducted comparisons over representative fashion retrieval models, and demonstrate that the model proposed in this thesis outperforms methods that use low-level descriptors, and also recent fashion retrieval models based on dense representations.

1.5 Thesis Outline

This thesis is structured in six chapters, as follows:

Chapter 2 Presents the basic definitions and techniques concerning this thesis. The concepts of Convolutional Neural Networks, Content-Based Image Retrieval, among others are presented in detail.

Chapter 3 Presents the related work in the context of fashion recommendation. Specifically, we emphasize the works that tackle content-based image retrieval

[6] www.chictopia.com

problems. In this chapter, it is also shown a comparison between the methods and contributions of this thesis and the others in the related work .

Chapter 4 Presents the methodology used in this thesis, as well as the problems we propose to tackle and the chosen solutions in detail.

Chapter 5 Presents a characterization and a statistical analysis of our fashion dataset and some interesting conclusions we could obtain through it.

Chapter 6 Presents the experiments and results achieved by this thesis. We present an evaluation of the CNN model, proposed in this thesis, as well as the ranking model. Finally, we present the results of experiments, balancing identity and fashionability, according to the user's preferences, in a world-wide scale.

Chapter 7 Presents the conclusions of this thesis and also the future work we could glimpse for it.

Chapter 2

Background and Concepts

This chapter introduces the key concepts for the better understanding of this thesis. The first section defines CBIR and shows the main current challenges of this area. The second section explains some basic concepts related to our CNN-based approach, besides clarifying the emergence of CNNs, presenting their main applications. The third section explains two main approaches for ranking, that differ according to the choice of using machine learning techniques. The fourth section clarifies the importance of balancing identity and fashionability, considering the visual search of looks.

2.1 Content-Based Image Retrieval

Content-Based Image Retrieval (CBIR) is the field of study concerned with searching and retrieving digital images from a large scale image database, according to users' interests (Sheshasaayee and .C [2014], Marques [2016]). Figure 2.1 shows the typical components of a CBIR system and their interaction.

According to the literature (Wang et al. [2010], Rafiee et al. [2010], Khokher and Talwar [2011], Sheshasaayee and .C [2014], Tunga et al. [2015], Marques [2016]), there are two main research communities that study image retrieval from different perspectives: one being text-based and the other visual based. The first one employs text or keywords to describe the content of the image while visual based uses visual features to describe the content of images, i.e. allows to use an image or a sketch as a query.

Text-based search has the advantage of being naturally quick and intuitive, but there are critical disadvantages like the inherent ambiguity of the language and the dependency on manually annotated labels, which is an expensive, subjective, context-sensitive and incomplete task (Khokher and Talwar [2011]), or tags and meta-data

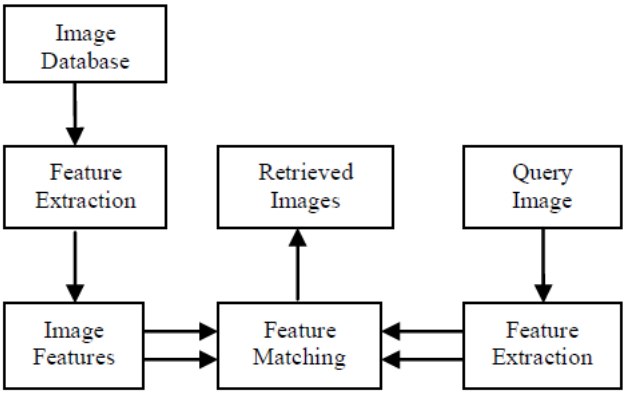

Figure 2.1. Typical components of a CBIR system and their interaction (Khokher and Talwar [2011]).

associated with the file surrounding the image in a website (Marques [2016]). Besides, the content of an image is much richer than what any set of keywords can express (Wang et al. [2010]).

Generally, a visual based search includes low level and domain-specific features. According to Wang et al. [2010], low level features are, typically: color (distribution of color intensity across image), shape (boundaries, or the interiors of objects depicted in the image), texture (homogeneity of visual patterns), spatial relations (the relationship or arrangement of low level features in space) or combination of above features. According to Lew et al. [2006], some examples of domain-specific features or systems are: face recognition, finger prints, handwriting, among others, which form a sort of high level image descriptions or meta-objects.

Regarding to current CBIR low level based techniques, they range from global feature based techniques to region based techniques (Tunga et al. [2015]). The first approach considers an image as a whole, and the main example are color histograms or color descriptors, such as BIC (Stehling et al. [2002]), ACC (Huang et al. [1997]), CCV (Pass et al. [1996]), GCH (Swain and Ballard [1991]), among others. This type of representation often gives disappointing results, because in many cases, images with similar colors do not have similar content. Other approaches consider texture and shape of an image, describing it through global descriptors, e.g. UNSER (Unser [1986]), EOAC (Mahmoudi et al. [2003]), SID (Zegarra et al. [2009]), among others. Some examples of region based representations are the bag-of-words (BoW) models (Wallraven et al. [2003]) and local feature descriptors, such as SIFT (Lowe [2004]) and SURF (Bay et al. [2008]).

Lew et al. [2006] claims that CBIR has been one of the most extensively studied

areas in multimedia community for more than a decade. Nevertheless, there are still open problems in this area, making it possible the emergence of many works improving well-known techniques as well as proposing new ones. Based on the previously mentioned works from the literature, the main challenges related to CBIR can be summarized in:

- Image Representation: CBIR aims to search for images through analyzing their visual contents (Wan et al. [2014]), and thus image representation is the crucial point of CBIR. In this thesis, we choose to represent an image through a feature vector of pieces of clothes.

- Image Similarity Characterization: it is difficult to define a precise measure for similarity, because it is a hard task to interpret the semantic of concepts in different CBIR application areas. The concept of look, in this thesis, is an example of this problem.

- Machine Learning techniques for Image Annotation: in general, the manual annotation task is considered costly, besides it requires, most of the times, a specialist. Thus, there is a demand for new efficient learning algorithms in this context, aiming to assist in these types of tasks. This is specially necessary when dealing with large scale image annotation (He et al. [2015]), whereas most existing methods are devised for small datasets. Fortunately, in this thesis, we can count on a richly annotated fashion dataset, which helps us in this context.

- Query Formulation: query formulation is an essential part of successful information retrieval (Yamin and Ramayah [2011]), and can be a hard task for a user that is not an expert in computers or in the related application area (Lee et al. [2009]). Sometimes it is also hard to describe a concept using only text. In this case, the problem can be mitigated by the use of an image as a query. In this thesis, we choose to use a query image since the concept of look is subjective and not clearly understood, specially for a fashion non-expert user. Using a query image we aim to improve query formulation to achieve better results.

- Query Result Display and Assessment: it is important that the results related to a search can be showed, sorted by their similarity. This is crucial since, in general, a user examines only the first results. In this thesis, we choose a ranking approach to deal with this problem, using the context information, i.e. climate, style and occasion, to estimate the relevance of each image, when compared to the query. Also, we focus on improving the ranking according to the user's needs,

through the balancing between the fashionability level of a look and the user's identity.

- Users' Feedback and Updating: it is important that a system considers users' feedback, modifying its retrieval mechanism in an attempt to return the desirable output (Sheshasaayee and .C [2014]). Few works propose new techniques in this field, and thus it is an area that deserves attention. In this thesis we do not consider user's feedback.

According to Tunga et al. [2015], the first CBIR systems used to focus on analyzing image content via low-level features, such as color, texture and shape. Otherwise, recent systems seek to combine low-level with high-level features that contain perceptual information for humans. Also, Rafiee et al. [2010] and other similar literature review works agree that, nowadays, the main problems to be solved in CBIR are related to image understanding. Specifically, the mapping between image visual features and high-level semantic concepts. Also, Wang et al. [2010] claims that bridging the semantic gap for image retrieval is a very challenging problem yet to be solved. In this context, this thesis aims to reduce this semantic gap, focusing on the approximation of the concept of look through its pieces of clothes, besides using the context information, i.e. climate, style and occasion, to help judging the relevance of a certain look when compared to others.

2.2 Convolutional Neural Networks

Artificial Neural Networks (ANN) are mathematical models that resemble biological neural structures (neurons), which have the computational capacity gained through learning and generalization (Rumelhart et al. [1986], Baldi and Hornik [1989] and Utgoff and Stracuzzi [2002]). An ANN comprises several processing units corresponding to the neurons. These units are interconnected by means of weights, which are numerical values representing the synapses. Synapses are responsible for determining an output that will serve as input to another unit. Figure 2.2 illustrates a basic structure of an ANN.

Many application problems can be solved using ANN, but some issues such as high dimensionality of inputs end up compromising the performance and accuracy of the results, as shown in Keogh and Mueen [2010]. In this context, deep learning has shown its power in learning good representations, specially from a large corpus. According to Bengio et al. [2012], in deep learning, there are a family of machine learning algorithms

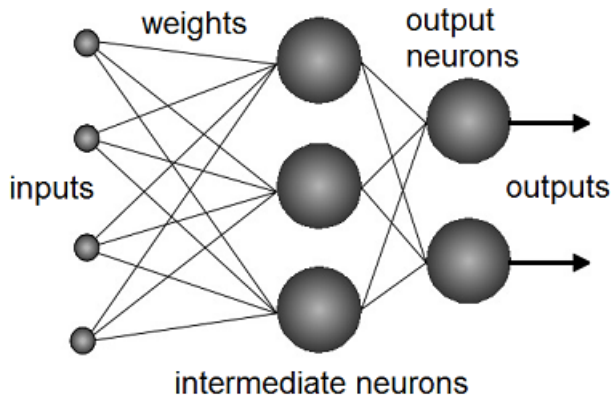

Figure 2.2. Basic structure of an artificial neural network.

that extract high-level abstractions in data by employing deep architectures composed of multiple neural layers.

Regarding deep learning techniques, with the increase of large scale image datasets and the advances in GPU computing, Convolutional Neural Networks (CNN)(Cun et al. [1990]) have received great attention, nowadays. Basically, CNNs are learning models inspired by the functioning of the visual cortex in humans (Zeiler and Fergus [2014]). The main differences between a CNN and a ANN are:

Weight sharing: in ANNs, each neuron of a hidden layer is fully connected to all neurons in the previous layer, and each neuron is completely independent and do not share any connections. CNNs present sparse connectivity and can share weights in a layer.

Scalability: ANNs do not scale well to full images, since they are wastefully connected and the huge number of parameters tend to overfitting.

3D volume of neurons and sub-sampling: unlike a regular neural network, the layers of CNN have neurons arranged in 3 dimensions: width, height, depth. In this case, the neurons in a layer will only be connected to a small region of the previous layer, instead of all of the neurons in a fully-connected manner. Also, a CNN is able to reduce the full image into a single vector of class scores, considering the depth dimension.

According to Zeiler and Fergus [2014], in a CNN, there are many types of layers: convolution layer, sub-sampling layer, normalization layer and fully-connected layers. A CNN is also organized through stages. Each stage is composed of one or more convolution layers in sequence, followed by a sub-sampling layer, which can be followed

by a normalization layer. A CNN can contain several stacked stages after the input layer, which corresponds to the image. After the final stage of the network, one or more fully connected layers are added to the structure, as shown in Figure 2.3.

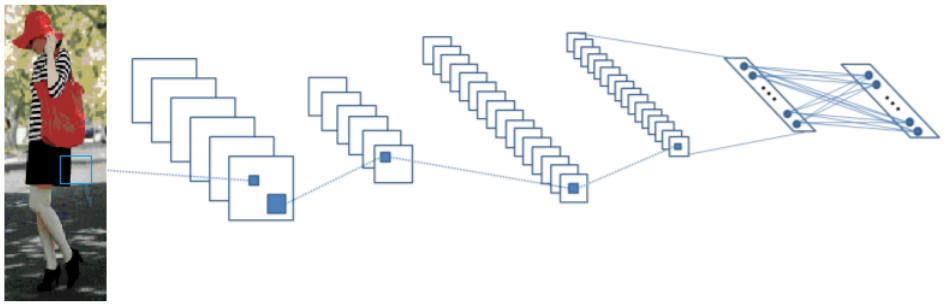

Figure 2.3. CNN structure with multiple layers, adapted from Lecun et al. [1998] .

Regarding the learning process in a CNN, an image is first segmented, then locally analyzed through the learning of filters or feature maps. Formally, a convolution over an image I corresponds to applying the product of Hadamard[1] between the pixel matrix of I and another matrix, called the convolution kernel (Zeiler and Fergus [2014]), i.e. the weight matrix shared by all the units (neurons) in a layer. Supposing an input image with size 30×30, Figure 2.4 (b) shows the result of certain convolution operation in an activation region of a feature map of size 2×2. A sub-sampling operation is also shown in Figure 2.4 (a), applying, in this case, a max pooling operation and feature maps of size 2×2 and a stride size 2. Chapter 4 presents details about the configuration of our CNN model in terms of layers and stages.

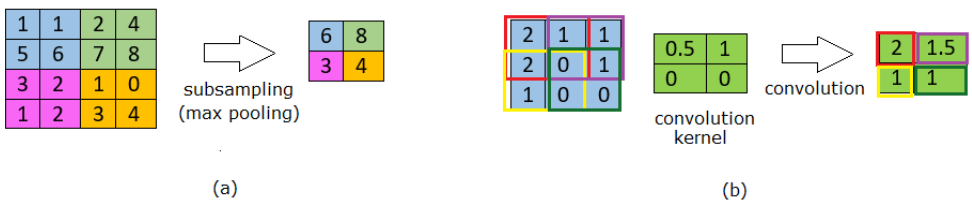

Figure 2.4. Examples of convolution and sub-sampling operations.

Actually, the success of CNNs has shown promising results toward bridging the already mentioned semantic gap. Regarding applications, through convolutional neural networks, it is possible to detect patterns in images in the sense that these patterns

[1]The product of Hadamard of two matrices A_{mxn} and B_{mxn} results in a matrix C_{mxn} so that $c_{ij} = a_{ij}b_{ij}$.

can be exploited for better results. Other applications are handwriting recognition and face detection (Krizhevsky et al. [2012b]). These networks have presented high popularity since 2012, when they achieved superior performance on a 1000-class recognition problem on the ImageNet dataset, containing more than one million images (Krizhevsky et al. [2012a]). CNN-based visual representation has also shown improved performance over handcrafted features on digit recognition, traffic signs (Sermanet and LeCun [2011]) and pedestrian detection (Sermanet et al. [2013]).

2.3 Image Ranking Strategies

In CBIR, a critical aspect of the system is the final ordering of the images, since users consider mostly the topmost results and their perception of the system quality is given according to them (Faria et al. [2010]). As already mentioned, we judge relevance based on how semantically similar are the query and the returned look, considering season, style and occasion.

Based on the use of machine learning techniques, we may present two main approaches for ranking images: if relevance information is available for training, it is possible to produce robust ranking functions using learning to rank strategies, generally obtaining better results (Liu et al. [2007]). Otherwise, it is easily possible to compare images using well-known measures such as Euclidean Distance or Cosine Similarity, making it possible to generate a ranking of images for a given query, sorted from the most to the least relevant, considering these measures. The main difference between the two approaches lies in that learning to rank can automatically learn the parameters of the ranking function using training data, while the traditional ones usually determine the parameters heuristically.

In this thesis, we do not employ relevance information to build our ranking model since this information is not available for training, rather we choose the traditional approach, simply calculating the Cosine Similarity between the compositional feature vectors of the outfits, obtaining good results when compared to the best works in the literature (Simo-Serra and Ishikawa [2016], Moreira et al. [2014]). Chapter 4 presents more details about our methodology, including the ranking model.

2.4 Identity and Fashionability

As already mentioned, fashion social network users certainly have their preferences regarding the composition of looks (Lurie [2000]). These preferences can be interpreted

as their visual identity or personal style. Despite this, these users look for inspiration to choose their own looks, generally because they want to be part of a group. For most of these users it is important to be considered well dressed in a circle of friends (Zanetti and Resende [2013]).

So, we may conclude that, most of times, a user who is looking for clothing inspiration, is not a fashion expert. But, if on the one hand, a user wants to be inspired by images of looks through which she can identify herself, on the other hand, most of times, it is desirable that such images are popular in the fashion social network. So that, it emerges the term fashionability (Simo-Serra et al. [2015]), used to qualify popular fashion looks, that show a good combination of pieces-of-clothes and accessories. Generally, it is possible to find looks with high levels of fashionability in fashion bloggers' profiles of social networks.

In this context, regarding the visual search for looks, we may redefine the concept of similarity, aiming to contemplate this reality related to the fashion domain. In the common scenario, assuming the user's visual identity can be detected through the query image, the returned images must be similar to it. Considering the new restriction, we may analyze the popularity of the returned images in the fashion social network, i.e. their fashionability level.

According to the literature (Hassan Zadeh and Sharda [2014], Schmidt et al. [2016], Ferrara et al. [2014]) it is very common to estimate the popularity of a post through its number of likes, number of comments and other similar metrics. Generally, the number of followers, the size of the post, among others are used as control variables that contribute to the decrease of popularity. In this thesis, after experimenting some combination of metrics, we choose the number of likes and the number of followers as estimates for fashionability.

Considering the fashionability level of the returned images, it is possible to redefine the concept of similarity as a combination of two concepts, not always in compromise, in practice: visual identity and fashionability. In this case, it is important to investigate this relation, aiming to discover if it is possible to build a ranking of images with high levels of fashionability, without compromising the accuracy of the original ranking model, which is based only in the visual identity concept. Also, since the preferences about fashion vary according to the culture of locations (Simo-Serra et al. [2015]), we may analyze the variation of the fashionability aspect in this context, considering fashion trends, ethnicity, culture and lifestyle, in different countries.

Chapter 3

Literature Review

This chapter presents the related work in fashion retrieval and recommendation using different techniques, focusing on CBIR applications. The first section presents a brief description about works, most of them CBIR systems, focusing on image processing techniques. The second section presents some works using deep learning-based techniques. The third section presents a discussion about the contributions of this thesis in relation to the other works in the literature.

3.1 Fashion Recommendation using Image Processing Techniques

In recent years, there has been an increasing interest in fashion related issues, being considered a promising application area for image processing and artificial intelligence approaches. Most works are related to the automation of fashion advice processes and, according to the literature, one of the most important steps in the automation of fashion advice is related to the detection of pieces of clothes. During the literature review, a lot of techniques were investigated, including image recognition, feature extraction, texture segmentation, shape extraction, among others. We have found some works applying image processing techniques to fashion and clothes recognition and they will be discussed in the next paragraphs. Despite this, it is important to emphasize that the focus of this thesis is on the techniques and methods used for learning image features and not on image processing techniques.

Tu and Dong [2010] proposes a model that helps customers to find their most suitable fashion choices in mass fashion information based on multimedia mining and recommendation. The model could be implemented in the context of a fashion on-line

store that analyzes clients' preferences for fashion recommendation. As preferences, the model considers favorite colors, skin tone and style. The model considers only the color of images as a visual feature representation. The mass of data, used as basis for recommendation, consists of images of fashion models in catwalks, which could make the recommendation not reflect the reality of most real buyers. Iwata et al. [2011] also proposes a system that recommends clothes using full-body photos collected from the users' favorite fashion magazines. Specifically, given a photograph of a fashion item (e.g. tops) as a query, the system must recommend photographs of other fashion items (e.g. bottoms) that are appropriate with regard to the query. It considers the popularity rate of websites from which the multimedia data is extracted and most contributions are given in the context of extraction and detection methods.

Lee and Lee [2015] make recommendation of ensemble clothing items, using the concept of paths and meta-paths, in an ensemble clothing dataset. The items, e.g. jacket, coat, t-shirt, among others, their attributes and ensembles are modeled as heterogeneous information that allows semantic analysis. The meta-paths are considered patterns of relationships between items with respect to attributes and ensembles. Relative importance of each meta-path in matching items is learned from an ensemble database, and the coefficient of each meta-path is learned using logistic regression on the feature vector and label pairs. Considering visual features, they also consider only color information and use k-means clustering to group color vectors. Di et al. [2013] present a multi-modal retrieval approach, based on the training of attribute classifiers on fine-grained pieces of clothes (i.e. coats) styles. In this work, given an input query – a text, an image or both – the system returns a ranked list of related items that contains the same visual attributes as the input. They represent features with visual descriptors, focusing on shape and texture, and use SVM to train the system in an attributes vocabulary.

Vogiatzis et al. [2012] describe the recommendation of clothes based on the interaction of users with fashion sites and the similarity between users and models appearing in fashion magazines. The main contribution is an ontology model to map users' profile and learn it through facts defined in a logic programming language, improving recommendation for online fashion stores. Hidayati et al. [2012] present approaches to automatically recognize clothing genre (e.g. formal shirt, t-shirt, among others), with an initial focus on upper-wear clothes. It considers style elements to represent the feature vectors to be learned and provides the genre of clothing according to them.

In one of the first works in fashion recommendation, Shen et al. [2007] propose the recommendation of outfits based on users' descriptions of specific scenarios over a broad range of everyday situations. The approach focuses on learning semantic attributes to

describe clothing, modeling clothing style rules used to predict a dressing style of a person or an event. In this work, the query is a textual description that defines the occasion and how the user wants to look like. Also, the work of Cheng and Liu [2008] defines an approach using a supervised neural network to retrieve images of clothes in a virtual closet, according to a textual input with keywords related to style and occasion.

Some works address the problem of cross-scenario clothing retrieval, many of them applying techniques based on Bag of Words (BoW) (Wallraven et al. [2003], Sivic and Zisserman [2003], Voravuthikunchai et al. [2014]). In Liu et al. [2012], given a photo captured in a general environment (e.g. on street), the problem is to find similar clothing in online shops. It uses human parts detectors and an annotated auxiliary set to learn a similarity transfer matrix to map the set to the online shopping set, deriving clothing similarities. Fu et al. [2013] also address the problem of large scale cross-scenario clothing retrieval using human parts detectors, sparse background reconstruction and the representation of features through bags of visual words. Another approach related to cross-scenario retrieval is proposed by Kalantidis et al. [2013], which focus on methods of pose estimation, clothing segmentation and classification of a query image, followed by suggestions of products from online shopping catalogs. Another very similar work, Yamaguchi et al. [2015], studied the clothing parsing problem using a retrieval based approach. As in Kalantidis et al. [2013], this work also focus on a accurate pose estimation as a prerequisite to the next phases. It combines pre-trained global clothing models, local clothing models learned on the fly from retrieved examples, and transferred parse masks from retrieved examples.

Kiapour et al. [2014] present a game-based approach to get human perception about style. Then, the labeled dataset is trained using a within-class classification of styles. Finally, they explore methods to identify clothing elements that are generally discriminative for a style, and methods for identifying items in a particular outfit that may indicate a style. Although it is an interesting contribution, their dataset is relatively small and the predicted styles are not so common in the fashion area.

Moreira et al. [2014] present a learning to rank (L2R) algorithm for finding similar apparel style given a query image. The proposed algorithm employs an association rule active sampling algorithm to select very small but effective training sets. Further, the algorithm operates on visual and textual elements, in a way that makes it able to expand the query image (for which only visual elements are available) with textual elements, and also to combine multiple elements, using basic economic efficiency concepts. This is a preliminary work in the context of this thesis, which improves upon the state-of-the-art models by 4-8% in terms of mean average precision.

Finally, Jagadeesh et al. [2014] present an automated visual recommendation

system for fashion, where given an image of a fashion item, e.g. a pair of jeans, the goal is to recommend matching fashion items, e.g. tops, that complement the given item. This work shows results in different types of context, including: place, event, season and cultural. It presents a data-driven approach, applying a set of algorithms based on Gaussian models, Markov Chain and Complementary Nearest Neighbor Consensus.

3.2 Fashion Recommendation using Deep Learning Techniques

The recent successes of deep learning techniques applied to CBIR applications made it possible the emergence of diverse types of research works in different application areas. Despite this, according to Wan et al. [2014], "it remains one of the most challenging open problems", "... and the key challenge has been attributed to a semantic gap issue that exists between low-level image pixels captured by machines and high-level semantic concepts perceived by human".

Regarding deep networks of general purpose, there are AlexNet (Krizhevsky et al. [2012a]) and GoogLeNet (Szegedy et al. [2015]) as the main contributions, considered the most important works in classification and detection. Some works in the context of fashion recommendation use these networks and datasets to improve their models before specific tasks in their works. In a general context, the work of Murthy et al. [2014] proposes models for automatic image annotation. They use Convolutional Neural Network (CNN) features extracted from an image and word embedding vectors to represent their associated tags. He et al. [2015] focuses on the issue of large scale image annotation, proposing a novel model based on deep representation learning and tag embedding learning. Specifically, the proposed model learns a unified latent space for image visual features and tag embedding simultaneously. Also, in the similarity context, Okada et al. [2015] proposes a novel Semantic-aware Hashing method (SaH) by discovering knowledge from social media resources to implement approximate similarity search.

With respect to the application of deep neural networks in the fashion context, Huang et al. [2014] presents an attribute-aware fashion-related retrieval system. Using a tree-structure CNN-based approach, they treat the attributes of clothes from the low-level layers of the net in an integrative way, separating them at the high-level layers, according to the semantic. They use the generic precursor AlexNet (Krizhevsky et al. [2012a]) and its dataset as a baseline. In few words, given an image, they use a human-detector that crops and resizes it as the query, which is fed into the CNN

to extract the high-level representation feature of clothes. Then, the extracted feature from the conjunction layer is used to conduct a similarity search to seek for visually and semantically similar clothes from the clothes repository, which is built with images from Amazon and other online stores.

Lin et al. [2015] present a deep search framework to tackle the problem of clothing retrieval in recommendation systems. First, the system also uses the AlexNet (Krizhevsky et al. [2012a]) and its dataset to learn mid-level visual representations. Then, a latent layer is added, making it possible to learn hashes-like representations, fine-tunning it on their clothing dataset, i.e. to learn domain-specific features. Finally, a query image is provided, and similar images are retrieved through a hierarchical search using the learned binary codes and mid-level representations.

Jing et al. [2015] present a content-based image retrieval approach to deploy a commercial visual search system at Pinterest.[1] The system provides applications that, given a chosen item, show the related pins and similar looks in that context. The system extracts local and deep features from the images using a CNN model. It also exploits the rich metadata available at Pinterest, firstly making a prediction of image categories using this data, then applying object detection modules specific to the predicted category.

Simo-Serra et al. [2015] propose a model to learn and predict how fashionable a person looks on a photograph. The concept of fashionability, applied in this thesis, was first mentioned in this work. The model combines four deep networks – each one receiving as input parameters such as: number of followers, age, garments, scene, tags, among others – joined together by a softmax layer and their outputs are used as features for the whole model. The model makes interesting inferences of correlations about fashion and other variables related to the posts and users, such as: age, beauty, location and income class. This work is quite similar to this thesis in some aspects, so that we compare some of our results with theirs.

Iliukovich-Strakovskaia et al. [2016] defend the usage of a fine-grained approach for image classification with pre-trained models to achieve a good predictive quality. The approach mixes the process of raw data (pixels of an image) with the learning of features from deep neural networks models trained on external crafted image datasets.

As an extension of Simo-Serra et al. [2015], Simo-Serra and Ishikawa [2016] present an approach to improve the learning of features related to clothing and fashion in a weakly-labeled dataset. Instead of training networks for classification and using an intermediate-layer representation as a feature vector, they present a method that

[1] http://www.pinterest.com

jointly trains both a feature extraction network and a classification network. In this case, they use a CNN-based model to learn compact (128-dimensional) discriminative features guided by a classifier that learns useful feature maps. This work is one of the baselines of this thesis because it is the current state-of-the-art in our context, besides it applies similar techniques.

Liu et al. [2016] present a new fashion dataset called DeepFashion, containing over 800,000 fashion images, richly annotated in the specific context. Besides, they also present a deep model, FashionNet, which learns clothing features by jointly predicting clothing attributes and landmarks. According to their results, DeepFashion dataset promises more accurate and reliable algorithms in clothes recognition and retrieval, so we decided to include it as a benchmark to our analysis.

Smirnov et al. [2016] propose a fast and accurate fashion item detection model based on deep neural networks. The model improves a general CNN with a system called Kuznech Mobile Recognition system, which can accurately detect all fashion items in a photo, classify each of them and find visually similar items in a large database, and all that in a very short period of time. In this work, they previously trained their approach using GoogLeNet (Szegedy et al. [2015]).

A recent work, Date et al. [2017], proposes a CNN-based method to personalize and generate new custom clothes based on the users' preferences and by learning the users' fashion choices from a limited set of clothes from their closet. According to the authors, by applying this method it is possible to separate the style and content of an arbitrary image and demonstrate how the other image can be stylized using the textures of the prior.

Another recent work, Matzen et al. [2017], applies deep learning methods to learn to extract fashion attributes from images and create a visual embedding of clothing style, used to analyze millions of Instagram photos of people sampled worldwide, in order to study spatio-temporal trends in clothing around the globe. The aim of this work is using temporal and geo-spatial statistics to generate concise visual depictions of what makes clothing unique in each city versus the rest.

Ji et al. [2017] is another recent work which presents a cross-domain approach for fashion image retrieval. It focuses on locating the attention of fashion product items in the query and in database images, considering noisy environments and background. To locate the attention of database images, they exploit the rich tag information available on the e-commerce websites. For query images, they use each candidate image in the database as the context to locate the query attention. They use novel deep convolutional neural networks to learn the attention weights and then extract effective representations of the images.

3.3 Our approach and The State-of-the-Art

After presenting the related work in fashion recommendation using deep networks, BoW, image processing, and other artificial-intelligence-based techniques, this section presents a summary of their contributions and a contextualization of this thesis in this context, as shown in Tables 3.1 and 3.2.

Regarding the aspect visual search (VS in Table 3.1), most works consider to use it, probably because, regardless of the application area, it is difficult to depict some subjective concepts as a look through words, being easier to provide an image as a query to be searched. Other works use, besides an image, textual filters (TF in Table 3.1), aiming to improve results. Regarding to the use of style elements (SE in Table 3.1), e.g. pieces of clothes, universal styles such as: classic, romantic among others, only some works use these semantic concepts as an approach to better understand the meaning of a look. A lot of works tackle, exclusively, image processing techniques to solve CBIR problems. The aspect cross-scenario search (CS in table 3.1) can be considered a challenge for most of works in the literature, since only a few of them take it into account. Ultimately, the use of real-world context (RWC in Table 3.1) in the search of looks is rare in most works, maybe because most of them focus on building CBIR systems in a context independent of users' specific needs. Actually, some works present models based in unreal fashion standards (Tu and Dong [2010], Iwata et al. [2011], Vogiatzis et al. [2012], Kiapour et al. [2014]), such as models from fashion magazines or catwalks.

Considering the features described in Table 3.1, this thesis differs specially in the focus on users' needs (aspects Cross-scenario Search and Real-world Context), which are not considered in many works. Our approach also differs for defining a model using style elements, i.e. pieces of clothes, for the approximation of the meaning of looks, aiming to contribute for improving the aspect of image understanding, reducing the semantic gap in the visual search, considered an important open problem in the CBIR context (Wang et al. [2010]). We also use the context information available in the dataset as a basis to judge the relevance of a candidate image, when compared to the query. On the other hand, most works focuses on building new techniques of image processing, without analyzing context information or trying to better understand the semantic of a query image.

According to the Table 3.2, the use of visual descriptors (VD) is very common in most of the works. Maybe because, as already mentioned, most of them tackle more issues related to the image representation with image processing techniques (IP). Some of these works apply the Bag of Words (BOW) approach to represent different parts

Table 3.1. Comparison of features between this thesis and state-of-the-art. VS = Visual Search, SE = Style Elements, CS = Cross-scenario Search, TF = Textual Filters, RWC = Real-world Context.

	Features				
Works	VS	SE	CS	TF	RWC
Shen et al. [2007]		X		X	X
Cheng and Liu [2008]		X		X	X
Tu and Dong [2010]	X			X	
Iwata et al. [2011]	X				
Hidayati et al. [2012]		X			
Vogiatzis et al. [2012]		X			X
Fu et al. [2013]	X				
Kalantidis et al. [2013]	X		X		
Di et al. [2013]	X				X
Kiapour et al. [2014]		X			
Jagadeesh et al. [2014]	X				X
Yamaguchi et al. [2014]	X		X		
Huang et al. [2014]	X				
Moreira et al. [2014]	X	X	X	X	X
Lee and Lee [2015]		X		X	
Lin et al. [2015]	X				
Jing et al. [2015]	X		X		X
Simo-Serra et al. [2015]	X	X		X	X
Simo-Serra and Ishikawa [2016]	X				
Strakovskaia et al. [2016]	X				
Liu et al. [2016]	X	X	X		
Smirnov et al. [2016]	X		X		
Prutha et al. [2017]		X		X	X
Matzen et al. [2017]		X			X
Ji et al. [2017]	X		X		X
Our approach	X	X	X		X

of the image, achieving good results. From 2011 to 2014, there is a predominance of works that mix the feature engineering techniques using visual descriptors with other statistical and mathematical models (OTH). Recently, the mix with machine learning (ML) classification or prediction models is more common. The most recent works, published in 2016 and 2017, focus on feature learning, specifically using Convolutional Neural Networks (CNN), showing significant improvements in the state-of-the art.

Regarding the use of techniques, this thesis tackles the image representation problem, using a CNN approach to learn the composition of feature vectors of pieces of clothes. Our approach differs from the others by tackling the problem of image similarity characterization, through the definition of a multi-objective function, aiming to

Table 3.2. Comparison of techniques between this thesis and state-of-the-art. VD = Visual Descriptors, BOW = Bag of Words/Features, IP = Image Processing techniques, OTH = Other statistic, mathematical and logical models, ML = Other ML techniques, CNN = Convolutional Neural Networks.

	Techniques					
Works	VD	BOW	IP	OTH	ML	CNN
Shen et al. [2007]				X		
Cheng and Liu [2008]		X		X	X	
Tu and Dong [2010]			X			
Iwata et al. [2011]	X			X	X	
Hidayati et al. [2012]	X		X	X		
Vogiatzis et al. [2012]				X		
Fu et al. [2013]		X	X			
Kalantidis et al. [2013]	X		X	X	X	
Di et al. [2013]	X	X		X		
Kiapour et al. [2014]	X			X	X	
Jagadeesh et al. [2014]	X	X		X		
Yamaguchi et al. [2014]	X		X		X	
Huang et al. [2014]	X		X			X
Moreira et al. [2014]	X			X	X	
Lee and Lee [2015]	X			X	X	
Lin et al. [2015]	X				X	X
Jing et al. [2015]			X		X	X
Simo-Serra et al. [2015]	X	X		X	X	
Simo-Serra and Ishikawa [2016]	X		X	X	X	
Strakovskaia et al. [2016]	X			X	X	X
Liu et al. [2016]				X	X	X
Smirnov et al. [2016]	X		X		X	X
Prutha et al. [2017]		X	X	X		X
Matzen et al. [2017]				X	X	X
Ji et al. [2017]			X			X
Our approach				X	X	X

balance two important concepts for the search of looks: visual identity and fashionability. In this case, we aim to discover a good configuration of values, providing results that match user's identity and present high popularity in the context of fashion. Besides, our experiments are conducted taking into account the cultural differences among countries, which may influence the variation of fashionability.

Chapter 4

Semantic Fashion Retrieval

Our approach for fashion retrieval is divided into two main steps:

- Learning a semantic space in which outfits are effectively represented, and
- Ranking relevant outfits according to a given query.

The approach is shown in Figure 4.1. The CNN model learns compositional feature vectors for the outfits (i.e., fashion images) in the dataset by predicting the probability of occurrence of clothing items. We assume that information of occurrence of clothing items is abundant and available in the form of "weak labels" [Simo-Serra and Ishikawa, 2016]. This information is necessary for learning compositional feature vectors.

Once the compositional features are learned, it is possible to build a semantic space in which images with similar composition of pieces of clothes appear next to each other. In the search, a query image is provided by the user and the CNN Model is able to predict the pieces of clothes in the look and, according to this prediction, the ranking model sorts the group of similar images, generating the preliminary ranking.

4.1 Learning the Semantic Space

Building an effective feature set to represent outfits is of paramount importance for improving fashion retrieval. In particular, we want features to be robust to background changes and to focus entirely on the outfit. Further, features should be meaningful to fashion attributes such as styles, occasions and seasons. Thus, we exploit the composition of outfits, so that outfits are represented by observing how likely are the possible

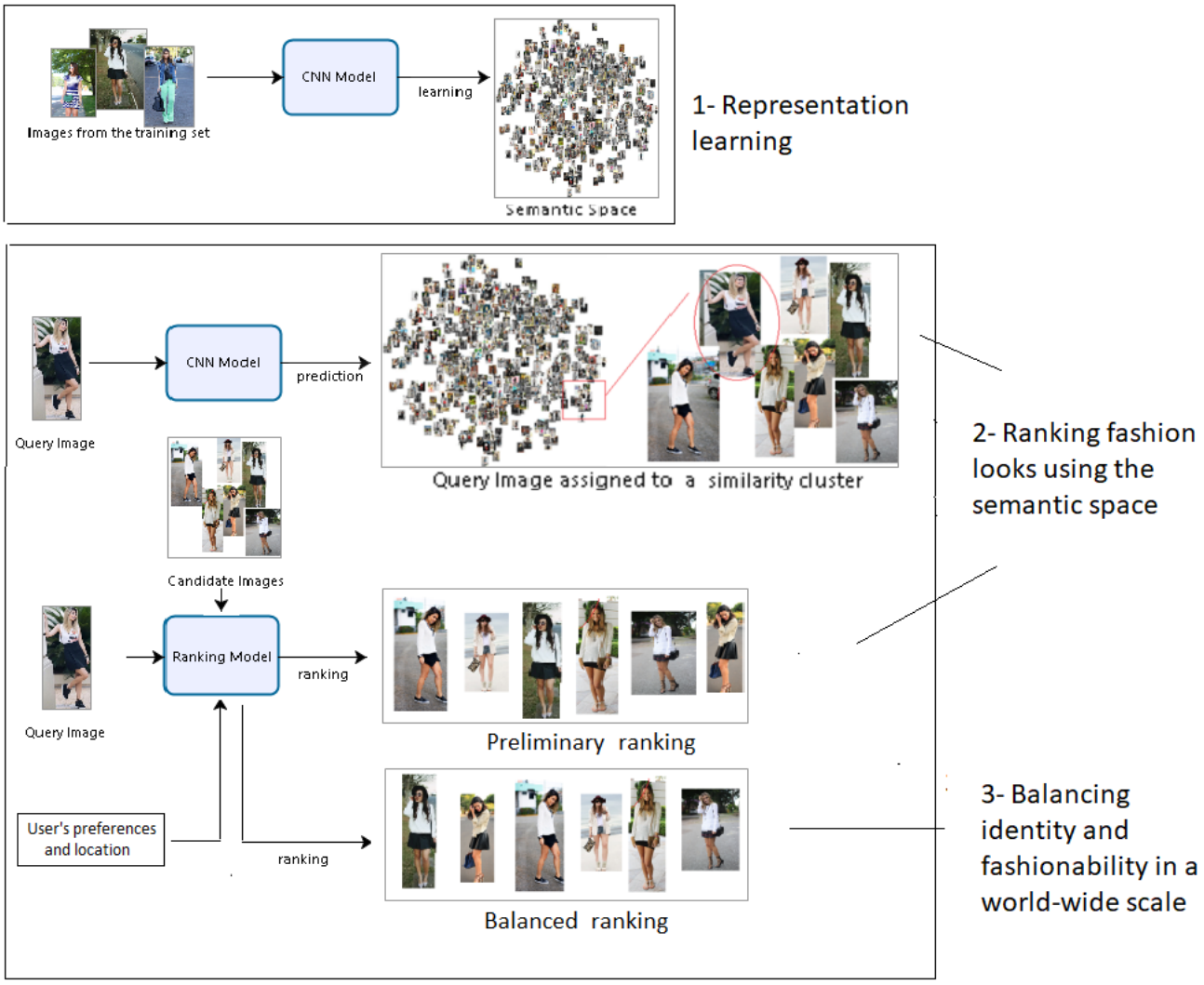

Figure 4.1. An overview of our methodology.

constituents. Convolutional neural networks (CNNs) are renowned for their high recognition performance and are thus one of the must-try algorithms. We used 2×2 kernels for the convolutional filters to keep the number of weights down for the network and allow increasing the number of layers [Simonyan and Zisserman, 2015]. A preliminary analysis showed that dropout[2] in the convolutional layers was not beneficial, and thus dropout is used only in the fully-connected layer to prevent overfitting throughout the architecture. The network output is given as a vector of probabilities associated with k clothing items (i.e., blazer, shirt, skirt, dress etc.), that is, a compositional feature vector. After careful inspection, we decided to fix the CNN output to the $k = 20$ most

[2]Dropout is a regularization technique for reducing overfitting in neural networks by preventing complex co-adaptations on training data. It refers to dropping out units (both hidden and visible) in a neural network (Srivastava et al. [2014]).

popular clothing items and accessories. A full overview of the architecture can be seen in Table 4.1. In terms of complexity, in this thesis, we consider colorful images, using, thus, the RGB format with three color channels, which increases the number of parameters to be adjusted during the training phase.

Further, as shown in Figure 4.2 (a), an image is first submitted to the feature extraction step, in which it is segmented and locally analyzed through the learning of feature maps. In this case, in each layer, each unit of a map performs the same operation – convolution or sub-sampling (pooling) – on the input image, with each unit applying this operation to a specific region of that image – in Figure 4.2 (a), an activated unit (neuron) is highlighted through a small filled square. This process continues until the CNN is reduced to a Multi-Layer Perceptron (MLP) which, finally, estimates the probabilities of each one of the twenty clothing items being in that image.

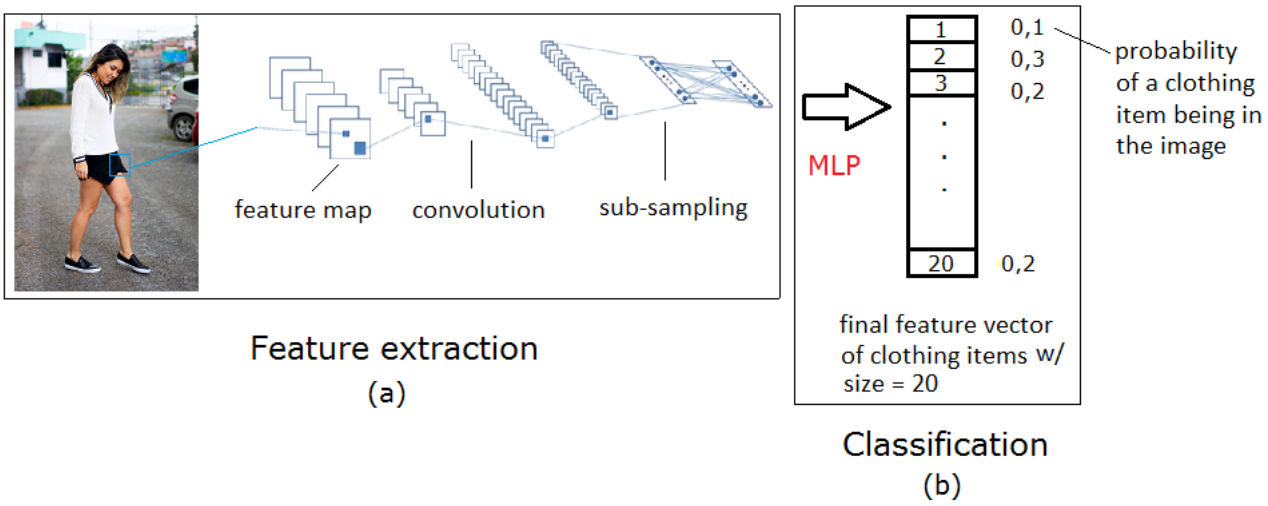

Figure 4.2. CNN learning process.

Training Learning rate was set to 0.01. We used Rectifier Linear Units (Relu) as non linear activations and a dropout probability of 0.2. The mini-batch size is fixed to 16 and training was stopped after 50 epochs with no improvement. We perform a grid search for these hyper-parameters, tuning on the validation set, with early stopping. The best model was chosen according to the smallest loss on the validation set.

4.1.1 Ranking Outfits using the Semantic Space

Once the outfits are properly represented by compositional feature vectors, it is possible to make comparisons between pairs of arbitrary outfits. In this thesis we applied the Cosine Similarity to calculate the distance between two feature vectors $(\vec{q}, \vec{c})$, as shown

type	kernel size	output size	params
convolution	2×2	$31 \times 31 \times 32$	416
convolution	2×2	$30 \times 30 \times 32$	4,128
max pooling	2×2	$15 \times 15 \times 32$	
convolution	2×2	$14 \times 14 \times 64$	8,256
convolution	2×2	$13 \times 13 \times 64$	16,448
max pooling	2×2	$6 \times 6 \times 64$	
convolution	2×2	$5 \times 5 \times 128$	32,896
convolution	2×2	$4 \times 4 \times 128$	65,664
max pooling	2×2	$2 \times 2 \times 128$	
fully-connected		1,024	525,312
dropout (20%)		1,024	
fully-connected		1,024	1,049,600
dropout (20%)		1,024	
fully-connected		20	20,500
Total		20	1,723,200

Table 4.1. Network architecture.

in Equation 4.1. In this case, feature vectors are first normalized to have unitary norm. This process is efficient and does not require additional steps devoted to learn ranking functions.

$$d(\vec{q}, \vec{c}) = 1 - \frac{\vec{q} \cdot \vec{c}}{\|\vec{q}\| \|\vec{c}\|} \tag{4.1}$$

After measuring the distance between a query image and the candidate images from the dataset, these distances are sorted in ascending order and the preliminary ranking is generated.

As a typical CBIR system, the preliminary ranking is built with the meaning of providing similar images considering the content of the query image. On the other hand, the fashion domain brings many issues that can change this scenario, redefining the aim of this particular CBIR system, in which the user is looking for fashion inspiration to create looks. So, besides the representation learning and classification of a look, based on the composition of its pieces of clothes, this thesis also provides a specific measure for the similarity of fashion looks, aiming to better reflect the reality about the searching of outfits. In this case, the search considers not only the user's visual identity, captured through the query image, but also the fashionability of looks – a concept first defined and applied by Simo-Serra (Simo-Serra et al. [2015]) – related to candidate images. This makes it possible for the user to have access to many desirable popular looks, even though they are not similar to his or her query image. We claim this is necessary to be

considered because most of users want to be inspired by popular looks, but, frequently, they are considered lay people in fashion, what reflects the imbalance of the user's visual identity and fashionability, in practice. Also, this approach allows prioritization, according to which the user considers more important, during the search, each time: her or his visual identity or popular looks.

Let α be a constant value that represents a weight given by the user, with the meaning of prioritization of fashionability of candidate images. Equation 4.2 estimates the new similarity index S, considering two images q and c.

$$S(\vec{q},\vec{c}) = \frac{\alpha F(\vec{c}) + (1-\alpha)d(\vec{q},\vec{c})}{2}, 0 \leq \alpha \leq 1 \tag{4.2}$$

Where:
$F(\vec{c})$ estimates fashionability through either the number of likes of a candidate image c or the number of followers of the user who posted candidate image c.
$d(\vec{q},\vec{c})$ represents the distance between two feature vectors of images q and c, i.e. the distance previously estimated using the Cosine Similarity.

We claim the number of likes and the number of followers are good estimates for fashionability, after conducting a set of experiments, aiming to discover estimates in this context. The results of these experiments are shown in Chapter 5. We also claim it is possible to achieve good results in terms of accuracy of the ranking, considering the new measure for the similarity of looks. The results presented in Chapter 5 confirm our hypothesis.

4.2 Ranking Outfits considering User's Location

As already mentioned, the fashionability of a look can vary according to some aspects, including user's location and fashion trends (Simo-Serra et al. [2015], Lurie [2000]). In the case of fashion trends, there is a set of cities, considered the fashion capitals, e.g. São Paulo, Paris, Tokyo, London, New York and Milan, responsible for dictating most fashion trends around the world (Zoe [2008]), which diminishes the impact of this aspect in the choice of looks.

Otherwise, the user's location reflects her or his own culture, preferences and lifestyle, strongly determining the choice of clothes. For instance, in Brazil, a sleeveless casual dress, chosen to go to the church in a Sunday morning in September, would be considered a look with high fashionability, but in the United States, a formal dress and a hat would be a better choice for the same occasion and time. In countries with high extension like Brazil, this difference exists also among states. In São Paulo, it

is common to find women dressing formal shirts and pants everyday, since the city is mostly considered a place for work. Otherwise, in Rio de Janeiro, it is mostly common to find women dressing casual looks, including jeans, short pants, sleeveless dresses or tank tops.

In this context, we decided to make an analysis of identity and fashionability considering different countries, aiming to investigate the differences between rankings built with posts from the same location of the user and without this concern. In this case, we applied Equation 4.2 to estimate the similarity of two images q and c, which are related to posts from the same user's location. Chapter 5 presents details about the results of our analysis, confirming our hypothesis and showing why it is necessary to conduct the search considering posts from the same location.

Chapter 5

Characterization of Data

Fashion is considered a subjective concept. In previous chapters, we already discussed about the difficulty of depicting a look, since the main aspects we could consider to define it are all subjective (Lurie [2000]). In fact, it is also complicated to predict associations of looks and attributes such as season, style and occasion, specially when we analyze the cultural differences between two or more countries, leading to multiple preferences of clothing style around the world.

With this concern, this chapter presents an exploratory analysis of our fashion dataset, including, among others, detected patterns related to clothing in the context of variables such as climate, style and occasion, in different countries.

5.1 The Fashion68k Dataset

Chictopia[1] is a website designed for fashion enthusiasts and bloggers to create profiles, post looks, and socialize with others interested in fashion. The site currently has over 255,000 users. Each post is associated with a look and several tags, indicating the occurrence of certain pieces of clothes and accessories. There are also tags indicating the fashion style of the look, as well as the occasions and seasons for which the outfit is appropriate.

In our experiments we use the Fashion68K dataset (Simo-Serra and Ishikawa [2016]) – a subset of Chictopia dataset – for training the CNN Model. For the evaluation of the model, we extend the Fashion68K dataset with tags related to styles, occasions and seasons, which enable us to compare looks in a semantic level. Regarding the experiments, we select 67,715 images as queries and 1,000 images/looks/posts

[1] www.chictopia.com

are returned in response for each of these queries. Relevance is given as the intersection-union ratio involving styles, occasions and seasons associated with query and returned images. This leads to multiple levels of relevance, varying from 1, when a perfect match occurs, with both looks sharing the style, being appropriate for the same season and occasion, to 0, when a totally irrelevant look does not share any of these attributes with the query.

Finally, as the main hypothesis of this thesis, the principle of compositionality allows us to learn feature vectors for accurately representing looks based solely on the occurrences of clothing items. In this way, after careful inspection, we decided to choose the 20 (twenty) most frequent clothing items to represent a look. The number 20 was chosen for being considered the best configuration during the experiments, using the Fashion68k dataset. So, each look/image/post is represented by a twenty-size feature vector, with the probabilities of each of the 20 clothing items being in that look, after the representation learning phase.

5.2 Fashion Bloggers around the World

This section presents a characterization of users and posts related to the Fashion68k dataset, according to their location.

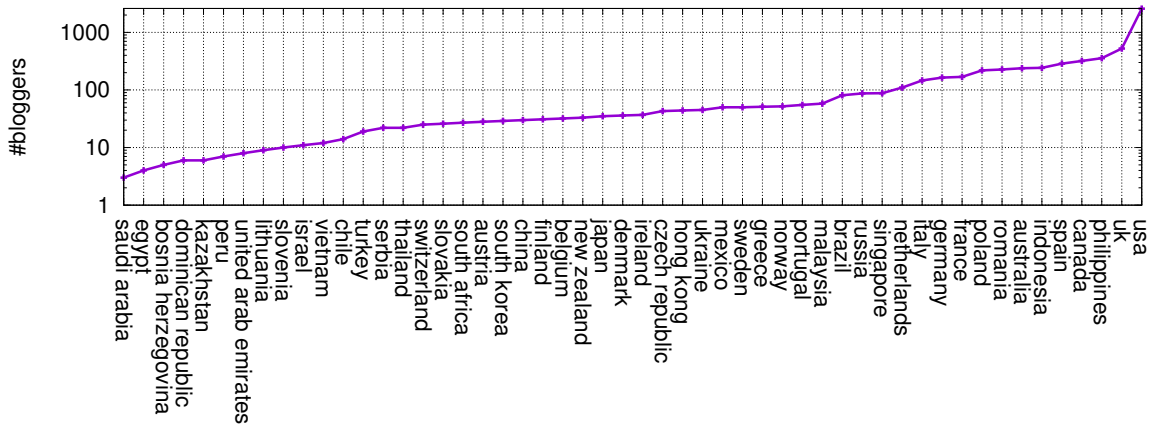

Figure 5.1. Distribution of fashion bloggers around the world.

Firstly, Figure 5.1 shows the distribution of fashion bloggers around the world. The chart shows the United States as the country with more fashion bloggers, which is kind of predictable, since New York is considered one of the main fashion capitals, with the most famous fashion week in the world (Zoe [2008]). Also, the United Kingdom,

in the second place, has London, another fashion capital, considered a great fashion trends dictator.

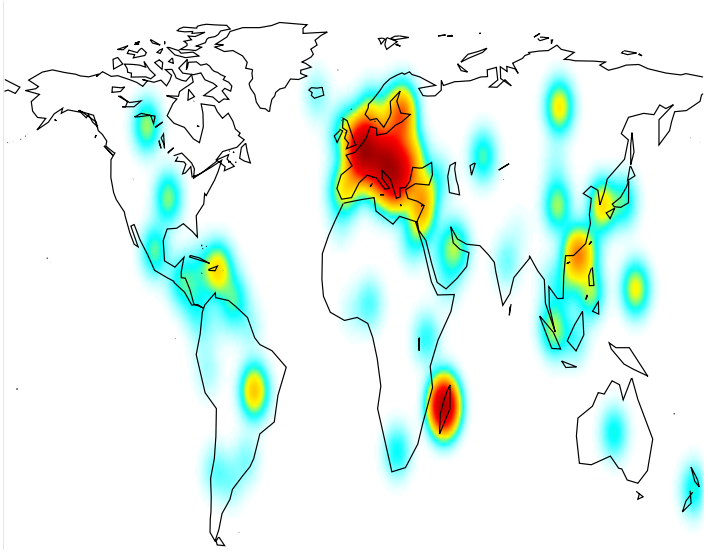

Figure 5.2. Scattering of posts around the world, considering the number of fashion bloggers from each country. In this chart, color red indicates the highest concentration while light blue indicates the lowest.

Figure 5.2 also characterizes fashion bloggers and their posts around the world.

In order to consider differences of population size, it shows the scattering of posts, considering the ratio of posts and population size of the country.

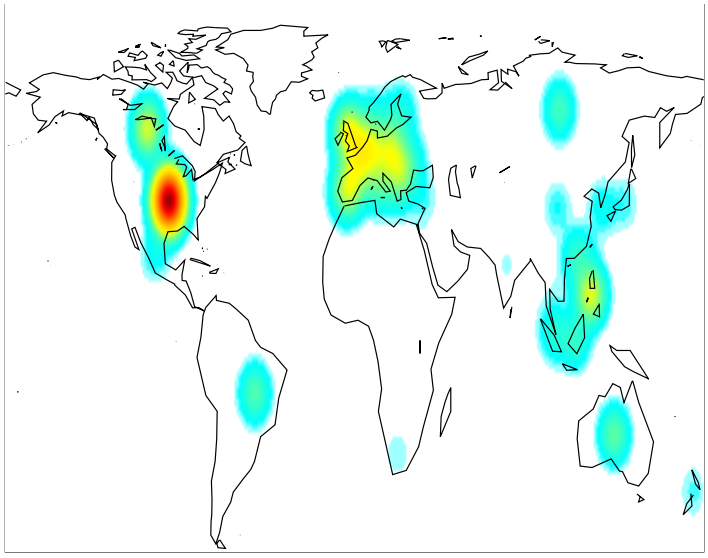

Figure 5.3. Scattering of followers around the world, considering the number of fashion bloggers from each country. In this chart, color red indicates the highest concentration while light blue indicates the lowest.

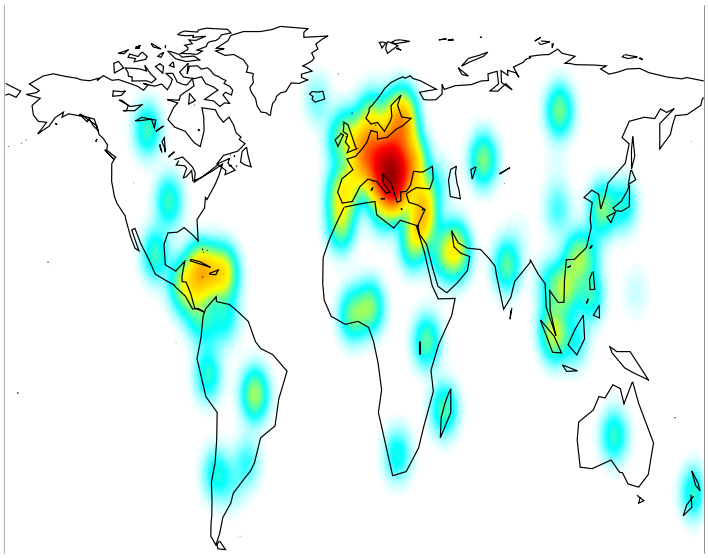

Figure 5.4. Scattering of votes around the world, considering the number of posts from each country. In this chart, color red indicates the highest concentration while light blue indicates the lowest.

Figures 5.3 and 5.4 illustrate the behavior of two aspects related to fashionability in this thesis: followers and votes, i.e. likes. Figure 5.3 illustrates the scattering of

followers of fashion bloggers around the world. In this case, the chart illustrates the average of followers in each country, considering the number of fashion bloggers from it. As seen, the United States is the country with the highest number of followers, as we could expect, considering they present the highest number of fashion bloggers, even though it is presented an average of followers. Figure 5.4 illustrates the scattering of votes of posts around the world. In this case, the chart illustrates the average of votes in each country, considering the number of posts from it. As seen, European users are the ones with the highest number of likes in posts, as expected, considering they are the ones who frequently post in the social network.

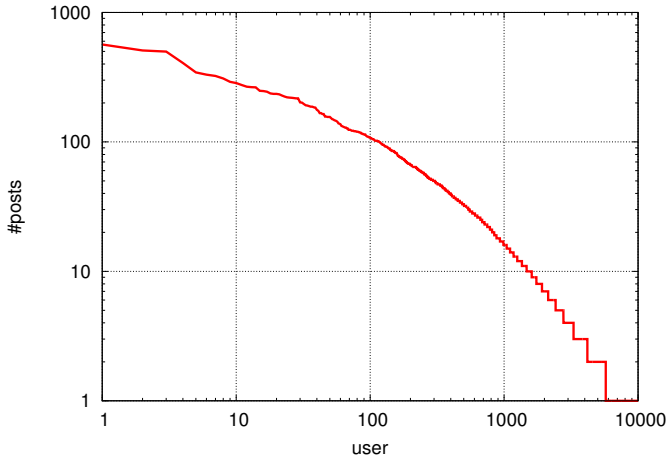

Figure 5.5. Distribution of posts in relation to users.

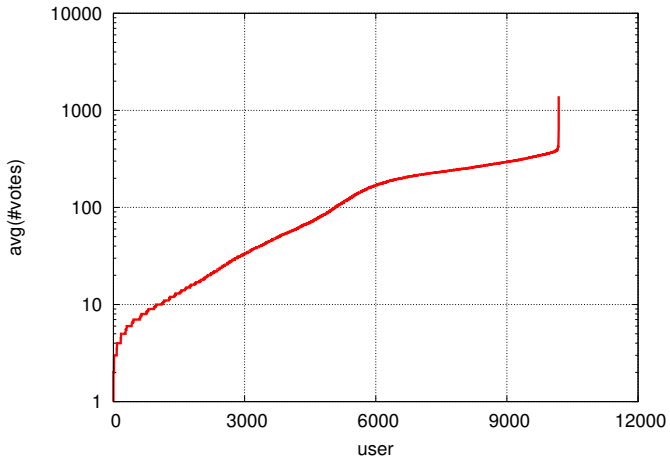

Figure 5.6. Distribution of votes in relation to users.

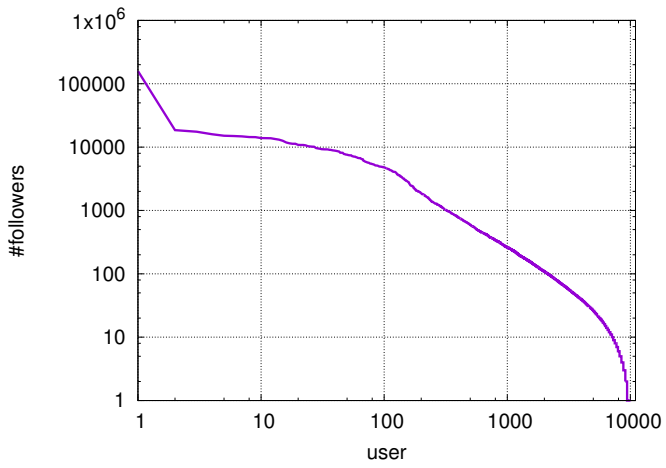

Figure 5.7. Distribution of followers in relation to users.

Regardless of location aspects, Figure 5.5 shows the distribution of posts in relation to users. In this case, it is possible to detect the pattern of few users posting a lot and most of them posting a little. Generally, those are considered the most famous fashion bloggers from the social network, whose popularity uses to grow with posting and interaction. Regarding popularity, Figures 5.6 and 5.7 show, respectively, the distribution of votes and followers in relation to users. Figure 5.6 shows the average distribution of votes, considering the number of posts of a user, illustrating that few users have more than 1000 likes in posts, in average. Figure 5.7 shows that, as expected, very few users have many followers, e.g. the most famous fashion bloggers from the social network, and most users have few followers.

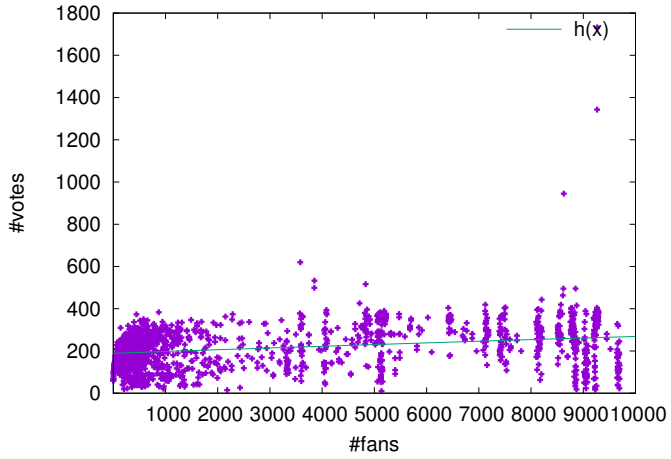

Figure 5.8. Distribution of votes in relation to followers.

Finally, Figure 5.8 illustrates the average and variance of votes in relation to

followers, showing that, on average, the number of votes in posts does not influence the number of followers and vice versa.

5.3 Clothing and Lifestyle Patterns around the World

This section illustrates some detected patterns related to three important concepts used to describe a look, in this thesis: style, occasion and season. Initially, Figures 5.9, 5.11 and 5.13 illustrate the distribution of these three concepts around the world. Similarly, Figures 5.10, 5.12 and 5.14 illustrate a semantic space, which shows the similarity among countries in the world, considering the aspects style, occasion and season.

According to Figure 5.9, the most popular style is chic, which is considered noise in the dataset, since many people choose this word as a style to their look, indiscriminately. Trendy is considered the second most popular style. Since our dataset is related to a fashion network, it is also expected fashion bloggers post looks which reflect the last trends of the current season, using the word trendy to describe their style. The following five most popular styles in the world are: comfortable, vintage, romantic, classic and urban. Regarding very specific styles, vintage looks are very popular nowadays, so we expected it would be on the list. Also, considering nowadays people prioritize comfort, it is easy to understand the popularity of comfortable looks. Styles romantic and classic are known as very popular, being two of the seven universal styles according to the fashion literature (Zoe [2008]). Finally, denim pants, t-shirts and flats are the most common choice of look (Lurie [2000]), being those the key pieces related to the urban style. Of course there are some countries that present specific behaviors, for example, Vietnam shows the 90's style as the most popular, although 90's style can be considered a specialization of urban style.

Regarding similarities among countries, according to Figure 5.10, we can say people in Brazil are dressed similarly in style to people in Argentina. Maybe because Brazilian people are used to visit Argentina a lot, specially during vacation time. This similarity also happens to Australia, Canada and United States; France, United Kingdom and Germany; Turkey and India; South Korean and Japan, among other groups of countries. In some of these cases, the similarities in clothing style may reflect the reality, since these countries are relatively close to each other, which favors the exchange of information and influences. In other cases, similar cultures and the same climate variation throughout the year may influence people's choices of clothes. These similarities can also be seen through the similar histograms in Figure 5.9.

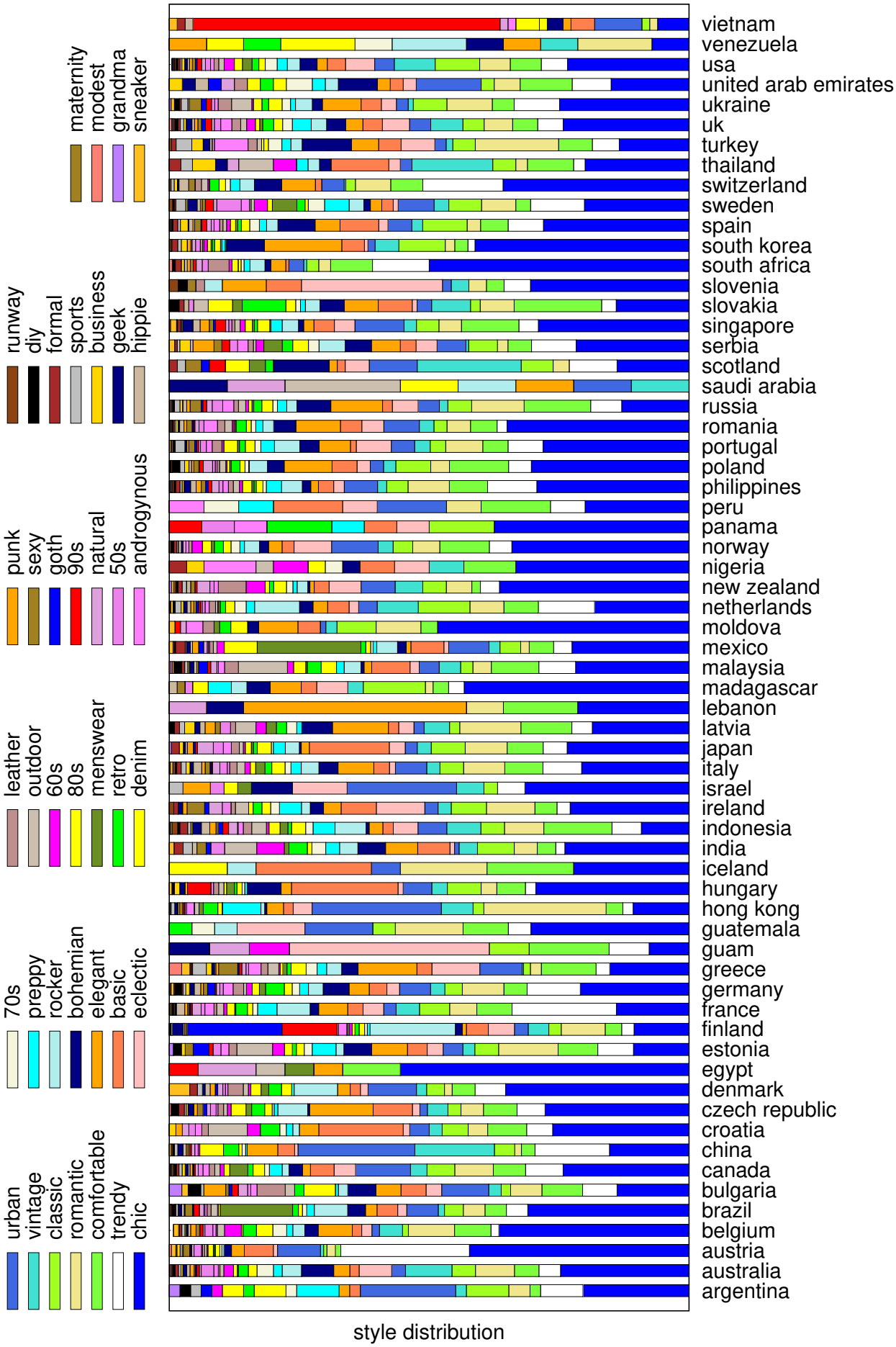

Figure 5.9. Distribution of styles around the world.

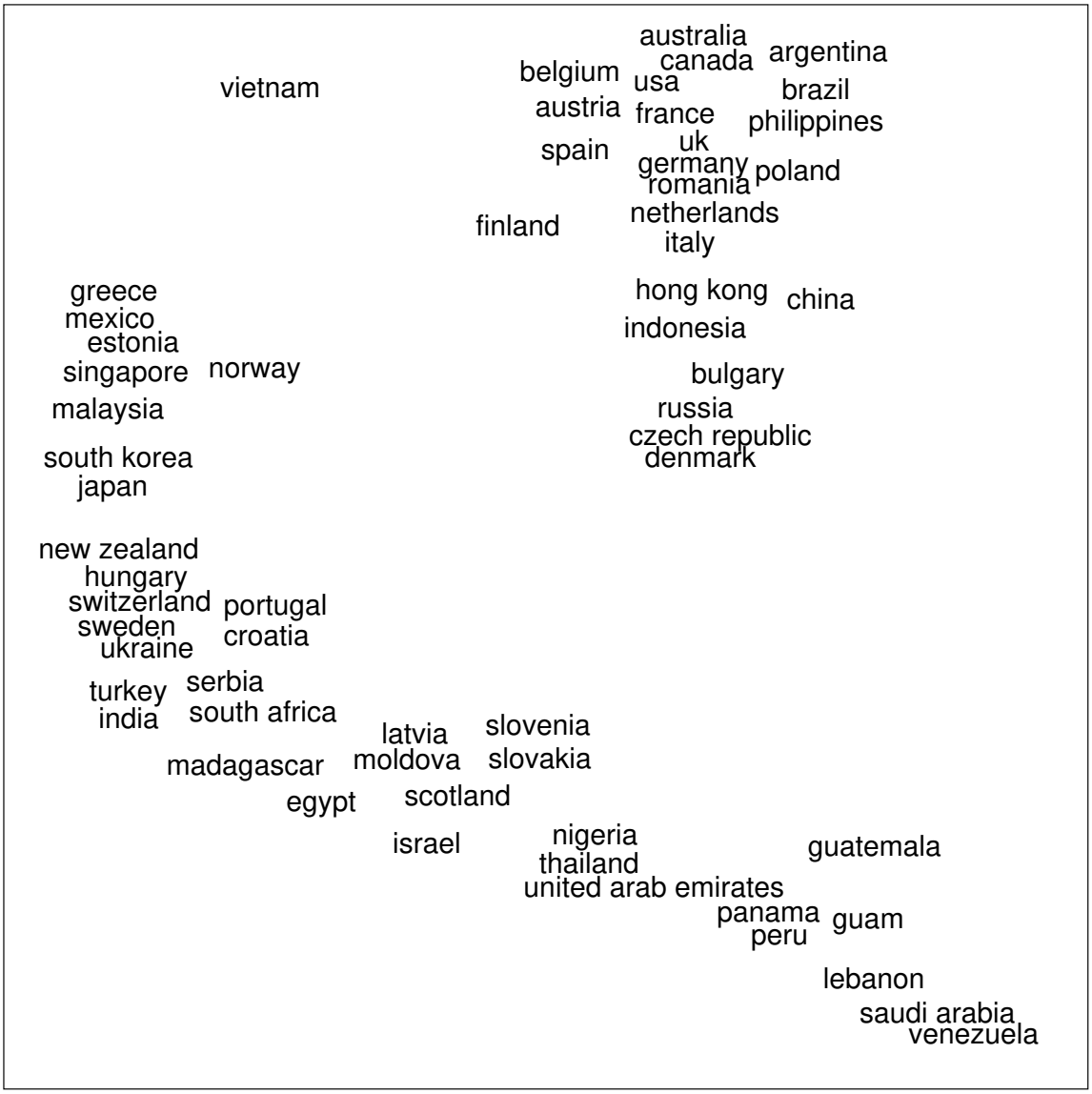

Figure 5.10. Similarity of styles around the world.

According to Figure 5.11, most people do not use or post looks for any occasion in particular (everyday occasion, in chart), which is quite expected since, most of the time, people do not have a specific occasion for which they compose looks, although people generally post photos when they go to parties and other events. This can also mean that people are not sure about labeling occasions, since another common word used to describe a place or event is: other. Otherwise, it is possible to detect some popular patterns, on average: brunch, dinner date, casual party and work, which are considered popular occasions in real-world context.

Considering similarities among countries, related to occasions, in Figure 5.12,

it is possible to see, for example, that Brazil and Australia share similar occasions to where people go and post photos of looks. Maybe because of their similarity in terms of climate, the lifestyle also seems similar. For reasons of cultural and fashion influences or proximity, we can also find similar occasions in: Peru, Guatemala and Panama; Brazil, Canada, United Kingdom and France; Saudi Arab and Qatar; Malaysia and Singapore; Venezuela, Puerto Rico and Costa Rica, among many others. It is interesting to observe that, in this specific aspect, China appears isolated in Figure 5.12, which can indicates either it is a country with a very particular lifestyle when it comes to people who like fashion or people from China are not labeling their posts in a regular manner, when it comes to occasions. It is important to notice that this isolation does not appear in the other distributions, which illustrate the behavior related to style and seasons. The mentioned similarities can also be seen through the histograms in Figure 5.11.

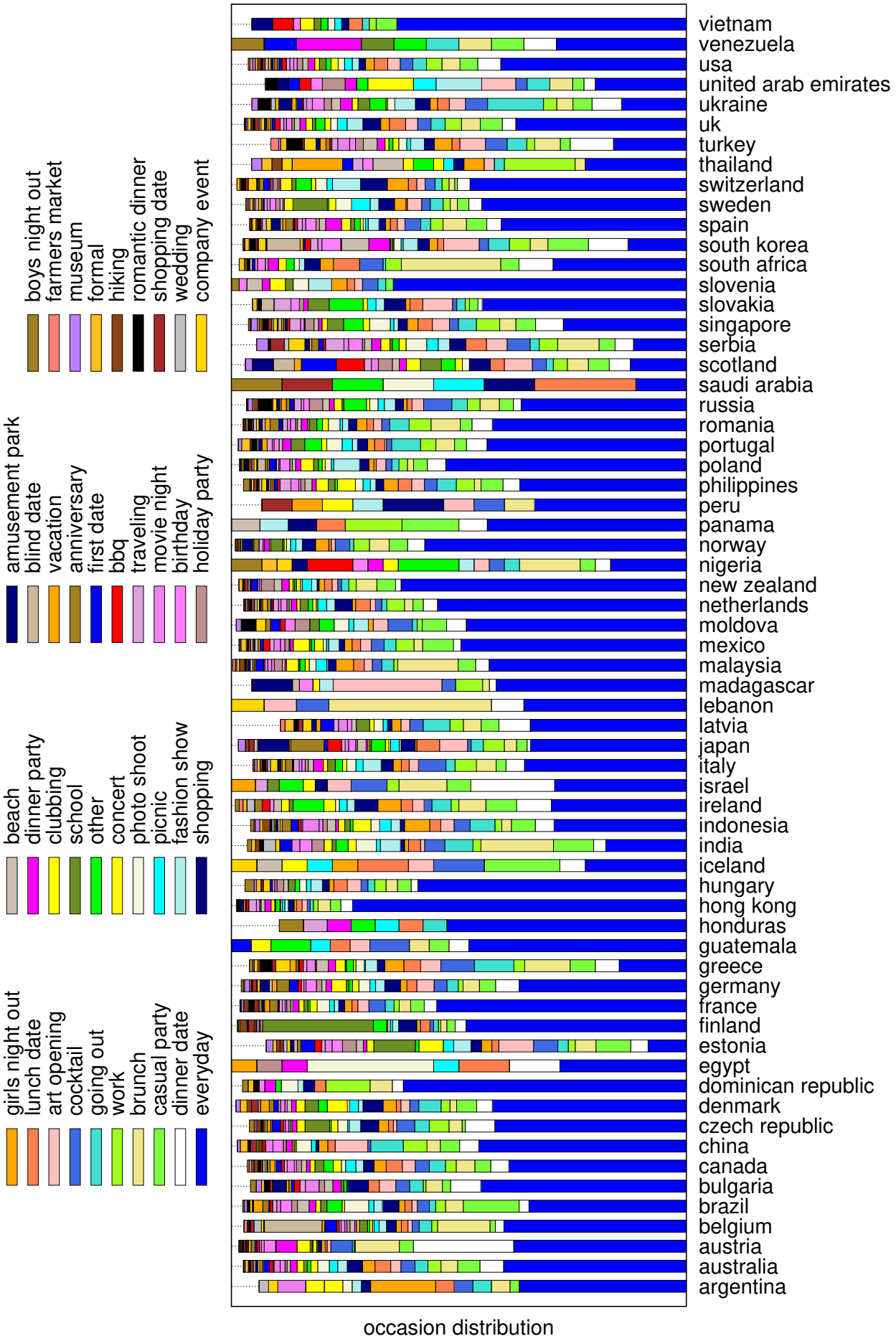

Figure 5.11. Distribution of occasions around the world.

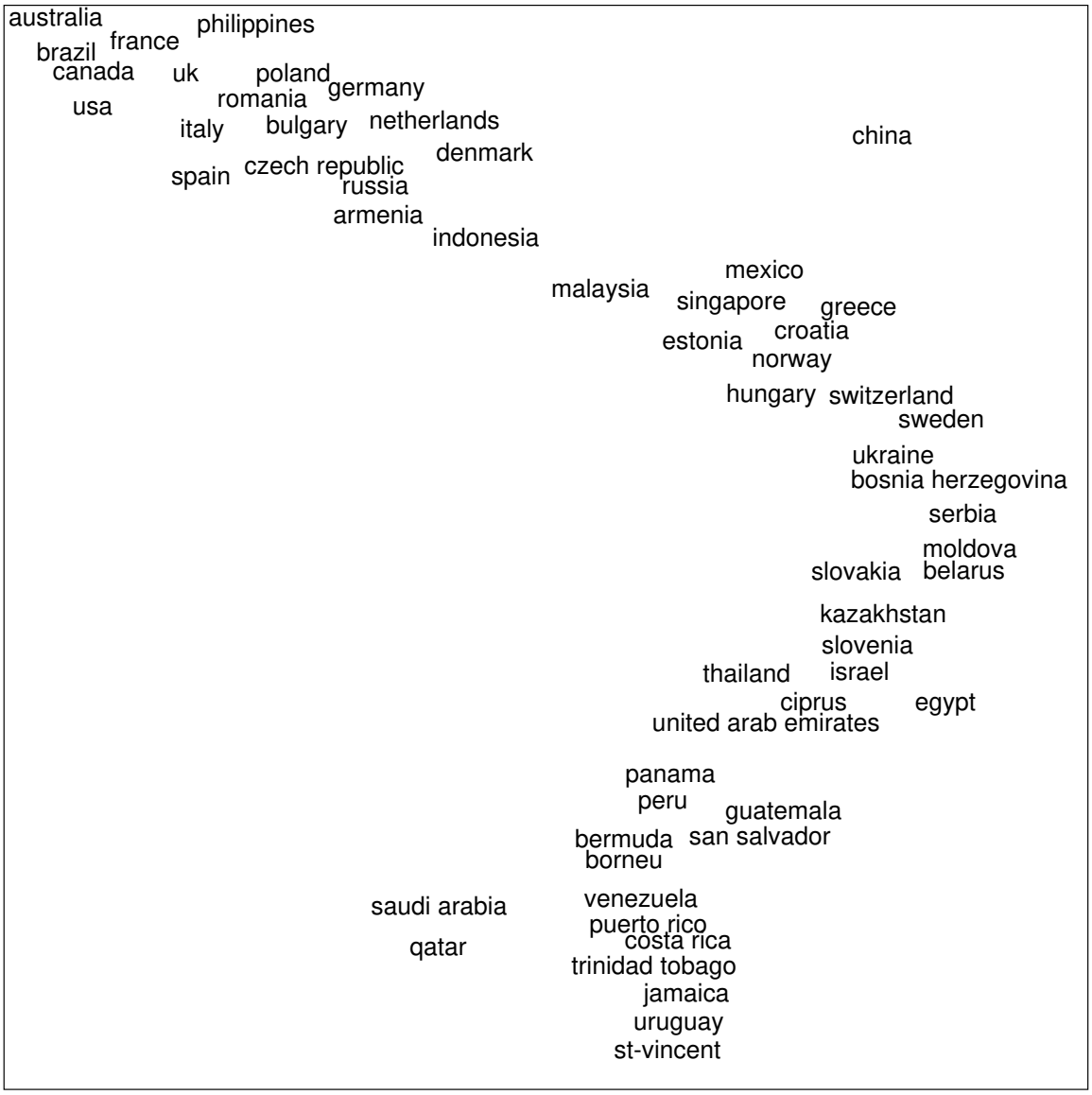

Figure 5.12. Similarity of occasions around the world.

Figure 5.13 shows that, on average, people mostly use and post looks in summer and spring. Maybe because these seasons are characterized by higher temperatures, with more possibilities for the composition of looks, motivating users to share their creativity in the social networks. Although fall and winter are fancier seasons, the associated lifestyle is not very prone to the composition of different looks to be posted.

In relation to similarities among countries considering seasons, in Figure 5.14 it is possible to detect the same behavior of posting looks, appropriated for specific seasons in: Brazil, Australia and Mexico; United States, United Kingdom and Russia, among many others. These similarities may occur according to the pattern of changing of

climate related to each country, when a pattern exists. For example, in Brazil, as well as in Australia, summer predominates throughout the year, and the changes of climate vary a lot. On the other hand, in Russia, United States and United Kingdom, the changes of seasons are more clearly perceived.

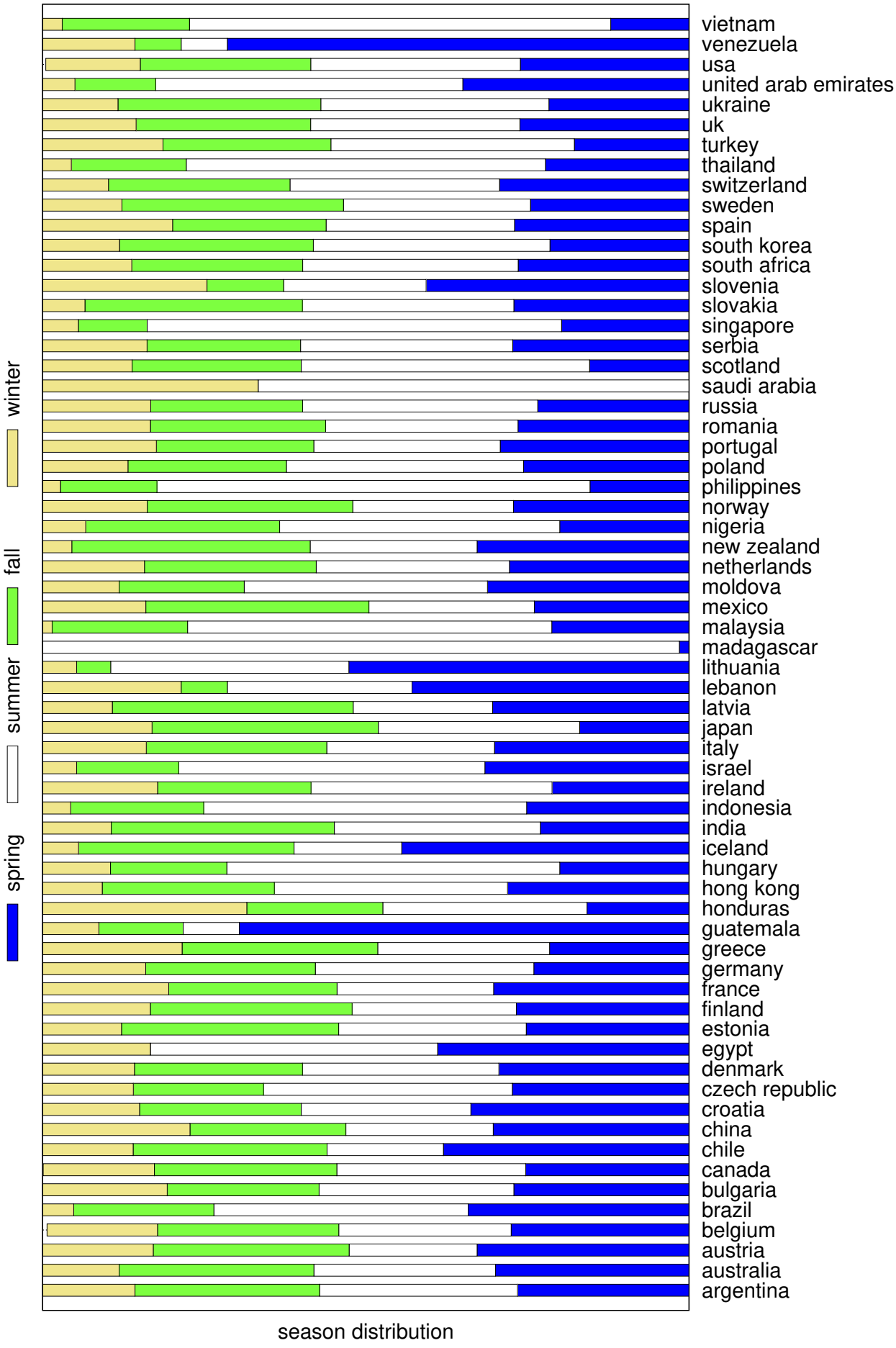

Figure 5.13. Distribution of seasons around the world.

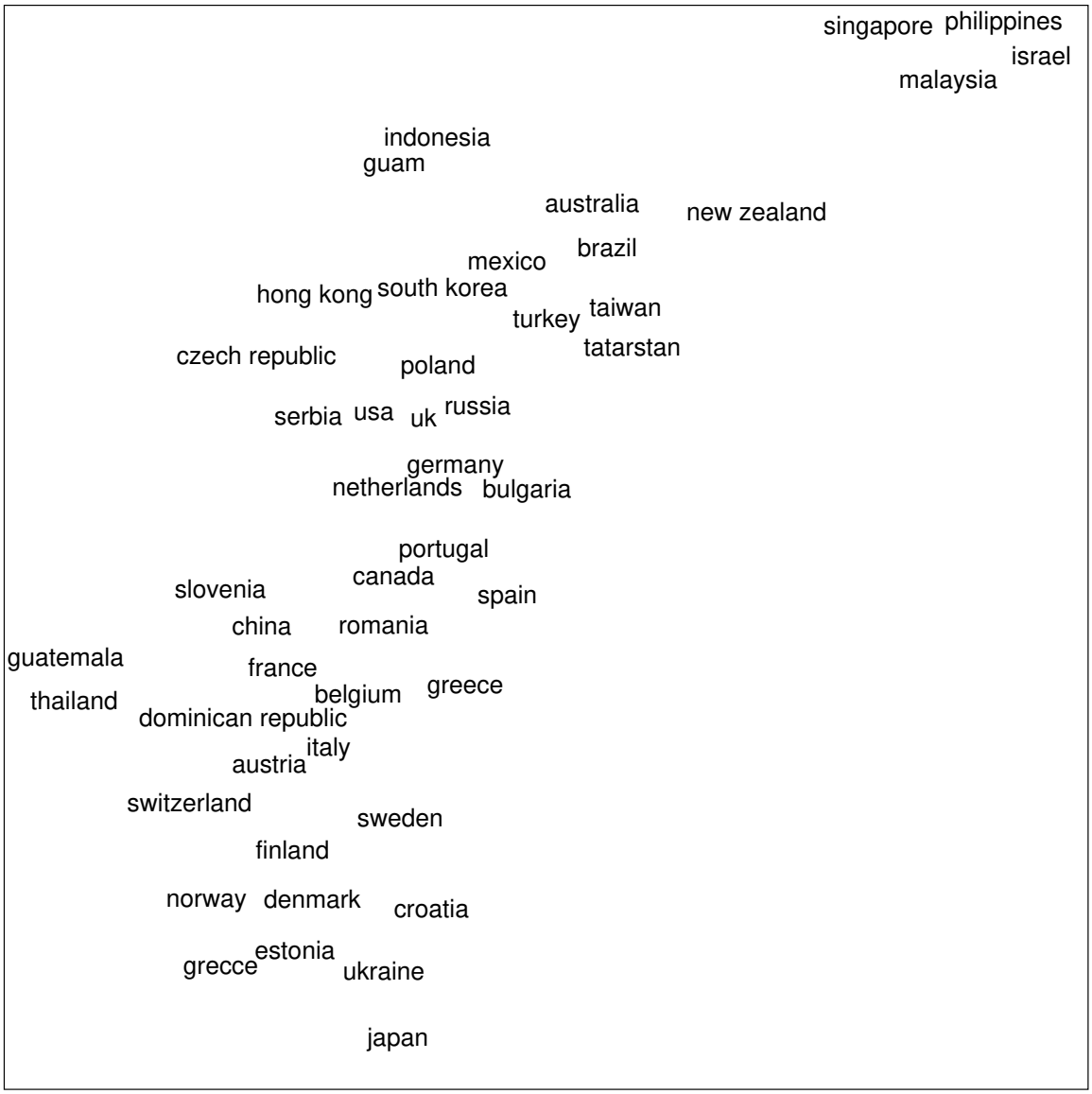

Figure 5.14. Similarity of seasons around the world.

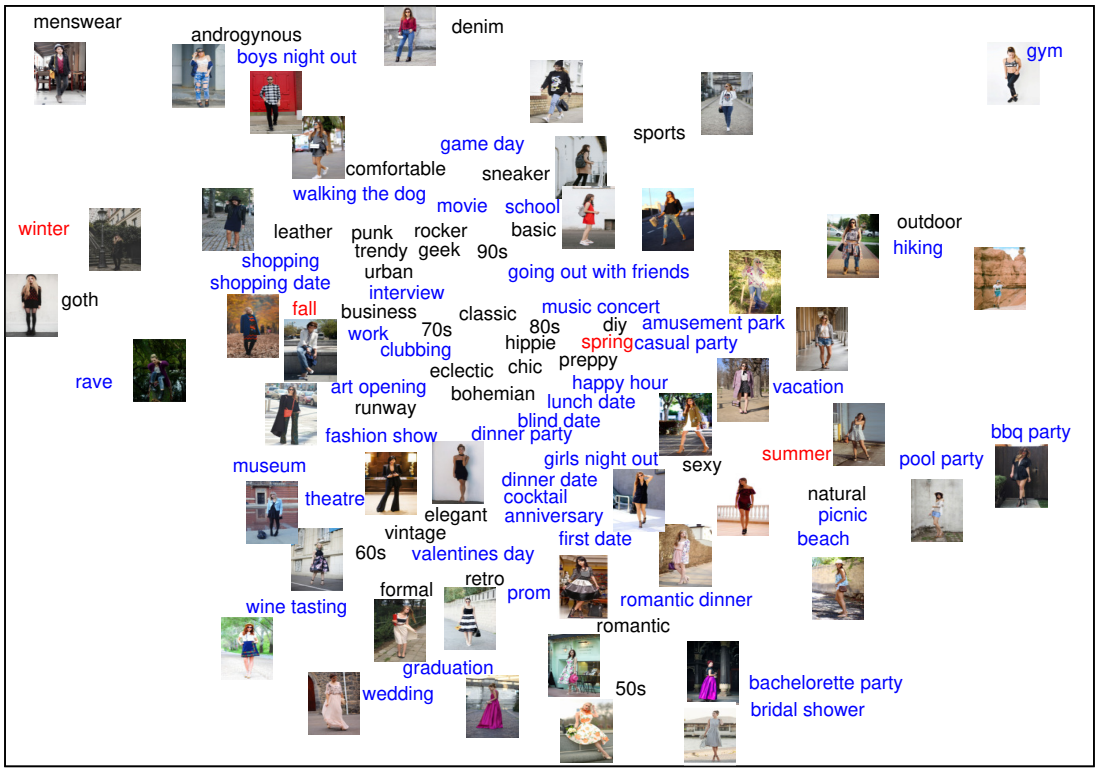

Figure 5.15. Semantic space: the correlation among styles (in black), occasions (in blue) and seasons (in red).

Figure 5.15 shows the semantic space, considering the correlation among aspects style, occasion and season. According to Figure 5.15, it is possible to see many interesting relations, for example, the similarities of looks used in dinner dates, cocktails and anniversaries. Besides, they share the most similar style, elegant. Also, we can see the similarity of looks used in the beach and pool party. In this case, as expected, the style natural is typical, as well as the season, summer. Other examples include: the romantic and 50s styles are close to each other, as are the vintage and 60s styles. The retro style is placed somewhere in between these styles. The same occurs with the occasions museum and theatre, which are close to each other. The sexy style is close to occasions such as girls night out and dating, and runway style is close to the occasion fashion show, and so on. It is also possible to see that looks placed next to winter shows to be darker and composed of more clothing items, while looks that are located next to summer are more colorful. Finally, it is also possible to grasp that our features display a remarkable robustness to background changes and focus mainly on the look.

Figures 5.16, 5.17 and 5.18 illustrate the average and variance of votes in relation

to style, occasion and season. According to Figure 5.16, the five most popular styles, i.e. styles related to looks which receive, on average, the highest number of votes, are: 90's, bohemian, menswear, sporty and sexy. The first two were expected in the list, since they are considered trending styles nowadays. Generally, there are few posts of looks related to menswear and when they appear, the specific public uses to attribute likes, indiscriminately, aiming to support that action, which can explain the high popularity of this style, on average. Similarly, looks related to sexy style use to receive likes indiscriminately, because it gets a lot of attention, besides cultural sexual issues that exist in many countries. The same conclusion may be applied to sporty looks, since sports fans, which compose many groups of people, tend to attribute likes because of the sport itself, and not because of the look. Figure 5.17 shows the five most popular occasions around the world, which are, on average: travelling, romantic dinner, vacation, brunch and fashion show. Considering brunch is among the preferred fashion bloggers' occasion (see Figure 5.11) as well as romantic is among the preferred fashion bloggers' style (see Figure 5.11), it is not a surprise brunch and romantic dinner are on the list. Also, travelling and vacation are very popular occasions in real life, and receive more likes, when compared to regular occasions. Looks used in fashion shows are frequently composed by pieces of clothes considered fashion trends, which tend to increase their popularity. Finally, Figure 5.18 shows there is not a relevant difference related to popularity of looks considering different seasons. On average, summer receive a little more votes than the others.

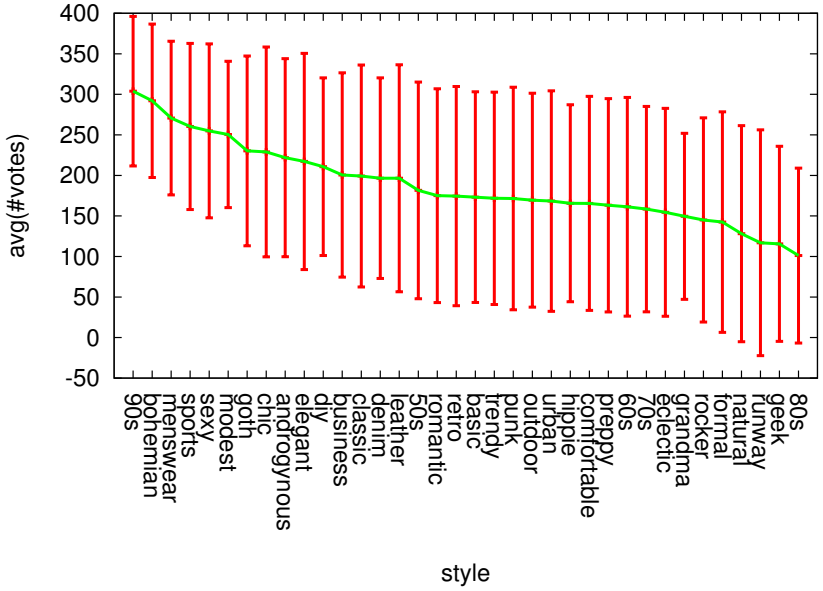

Figure 5.16. Distribution of votes in relation to styles.

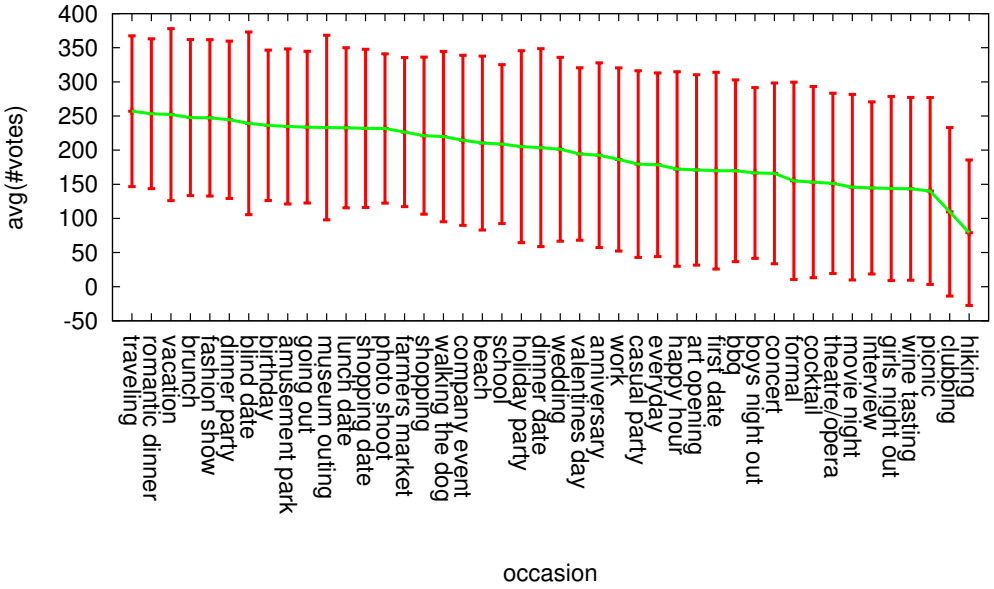

Figure 5.17. Distribution of votes in relation to occasions.

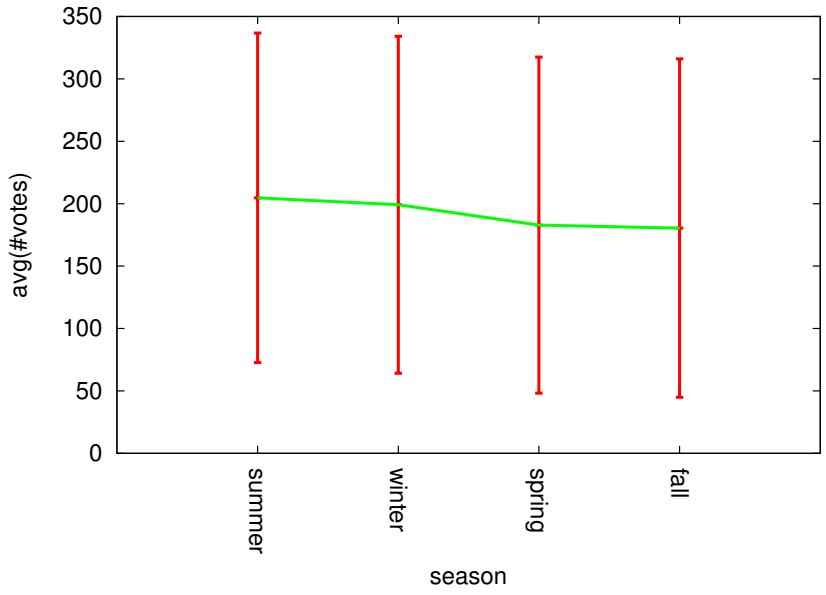

Figure 5.18. Distribution of votes in relation to seasons.

Chapter 6

Experimental Evaluation and Results

This chapter presents the experiments, analysis and the results achieved by this thesis. The first and second sections present details about our experimental evaluation, including our baselines, the evaluation metrics and the evaluation procedures. The third section presents the results of our experimental evaluation considering the ranking built by the CNN model, exclusively based on the user's visual identity, and the balanced ranking, considering visual identity and fashionability. Following, we present the evaluation of the balanced ranking, considering only posts from the same location, which are related to the query and candidate images, showing the impact of this concern in the final ranking.

6.1 Baselines

This section presents the two baselines we used, aiming to compare our results in this thesis. We considered the two following methods, since they present the best results in the literature, considering the context of fashion applications:

- LLDs: each look is represented using a feature set composed of 12 low-level descriptors, including color, texture and shape. Euclidean distance is used to rank relevant looks and a L2R algorithm is also applied (Moreira et al. [2014]).
- StyleNet-1.0: a feature extraction network which minimizes a ranking loss, and a classification network which minimizes the cross-entropy loss (Simo-Serra and Ishikawa [2016]) are trained jointly. The input for the network is composed of "weak labels", which is similar to the input of our model. Otherwise, we train

the model to learn the composition of looks, to be used during the similarity analysis, and they do not consider this composition, exclusively, as basis for the classification task. Besides, we filter the available weak labels, considering for training only the pieces of clothes, and for the evaluation only the context information – occasion, season and style – related to the look, while they use the available weak labels for learning and classification, without a specific criterion.

6.2 Evaluation Procedure and Metrics

This section presents the evaluation procedures we adopted in this thesis, as well as the metrics we used to analyze the quality of our models.

In this thesis, in order to evaluate the preliminary ranking, produced by the CNN (see the Ranking Model in Figure 4.1), we used standard Precision, MAP (Mean Average Precision) and NDCG (Normalized Discount Cumulative Gain) measures (Järvelin and Kekäläinen [2002]), since these are considered standard, when it comes to information retrieval and ranking systems. Regarding evaluation procedures, we conducted five-fold cross-validation, that is, data are arranged into five folds with the same number of queries. At each run, three folds are used as training set, one fold is used as validation set, and the remaining fold is used as test set. The training set is first used by the CNN Model to learn the compositional vectors (see the representation learning phase in Figure 4.1). The test set is used to estimate retrieval performance. The results presented in the next section are the average of the five runs, and are used to measure the overall retrieval performance of the ranking model.

As an important contribution of this thesis, the balanced ranking (see Figure 4.1) is produced through the application of a new score function defined through the balancing of identity and fashionability, considering user's preferences. The balanced ranking is evaluated based on the relation between NDCG, which represents the gain in terms of similarity of user's identity, and fashionability, that represents the gain in terms of popularity of a candidate look, when it comes to fashion. The analysis of the balanced ranking is conducted through the variation of α value (see Equation 4.2, defined in Chapter 4), simulating different preferences for a user. These results are presented in the next section, considering a ranking composed only by posts from the same location of the user and without this concern.

6.3 Results

This section shows the results of experiments conducted in this thesis. The first section presents the results related to the CNN Ranking Model, which produces the preliminary rank. The second section presents an analysis of gain related to our Balanced Ranking Model, which builds a rank, based on the compromise of two important aspects of fashion retrieval: identity and fashionability.

6.3.1 The CNN Ranking Model

In this section, we refer to our CNN Ranking Model as CS–CF (standing for Cosine Similarity with Compositional Features), CS–CO (standing for Cosine Similarity with Contextual features) and CS–IC (standing for Cosine Similarity with Ideal Compositional features). Table 6.1 shows the ranking performance of CS–CF, CS–CO and CS–IC, as well as the ranking performance of the baselines. Low-level descriptors (LLDs) lead to the lowest performance, showing that the performance of low level descriptors as learning to rank features is still very poor. MAP numbers achieved by CS–CF and CS–CO are significantly higher than MAP numbers achieved by StyleNet–1.0. Further, CS–CF and CS–CO perform better than StyleNet–1.0 in the topmost positions, and their performance tend to approximate as the ranking size increases. CS–CO presents the highest numbers for MAP and NDCG superior performance when compared to the others, maybe because the contextual labels used in the training are the same of those used in the evaluation. So, this experiment functions as an upper bound for our approach, since we do not have access, in practice, to labels related to contextual features to train the model. Surprisingly, CS–IC performed poorly. After careful inspection, we suppose it has to do with the feature vectors produced by the CNN, which are not sparse neither binary. This fact may lead to wrong conclusions such as a short skirt is better related to a short than a coat.

Finally, we clarify the retrieval performance of the best performing models by inspecting their performance in each query. Figure 6.1 shows MAP numbers for each query. StyleNet1.0 achieves very high performance for some few queries, but CS–CF achieves better MAP numbers for a larger amount of queries, explaining its overall superiority in terms of MAP. Figure 6.2 shows NDCG@10 numbers for each query. Again, StyleNet1.0 achieves very high performance numbers on few queries, but CS–CF surpasses StyleNet1.0 in most of the queries.

	MAP	NDCG@			Precision@		
		1	5	10	1	5	10
CS–CF	0.472†	0.264†	0.241	0.239	0.495†	0.486	0.485
CS–CO	0.502†	0.291†	0.275	0.272	0.565†	0.551	0.547
CS–IC	0.465	0.254	0.230	0.227	0.482	0.479	0.477
LLDs	0.356	0.188	0.170	0.167	0.374	0.367	0.366
StyleNet–1.0	0.469	0.258	0.238	0.237	0.490	0.484	0.484

Table 6.1. Ranking performance of the different models. Symbol † indicates statistical superiority in relation to StyleNet–1.0, considering Wilcoxon test, with p-value 0.01.

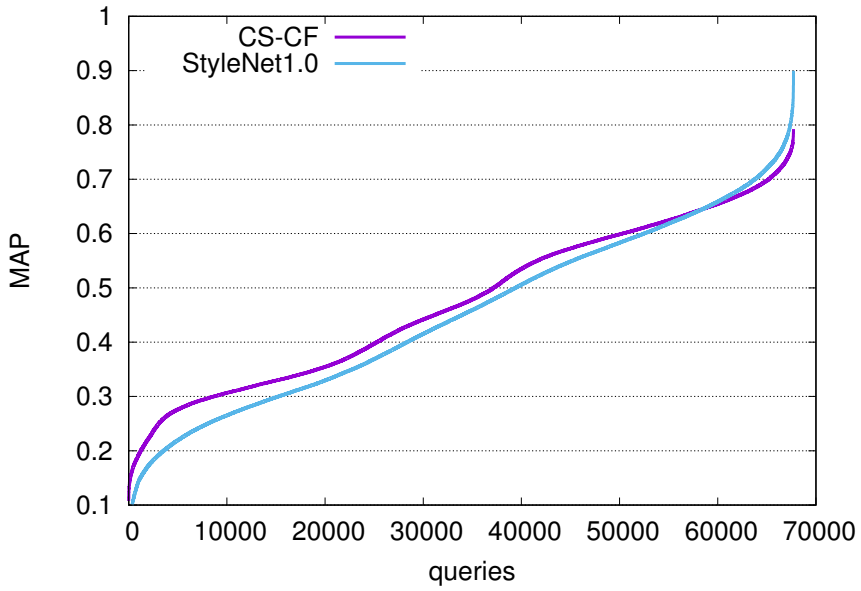

Figure 6.1. CS-CF versus StyleNet1.0 - MAP number for each query.

6.3.2 The Balanced Model

In this thesis, when a user provides an image as a query, it implicitly encodes her or his visual identity, which he or she desires to match. However, this image is not, necessarily, a good reference in terms of fashion. Since the final ranked list should prioritize images with high fashionability and also reflect user's identity (see discussion in Chapter 2), it is necessary a compromise of these two aspects, in order to produce a reasonably balanced ranking. In this context, this thesis aims to show it is possible to build a balanced ranking for the search of looks, satisfying the needs of typical users from fashion social networks. The obtained results show it is possible to build a balanced ranking, considering a loss in terms of NDCG, in most cases.

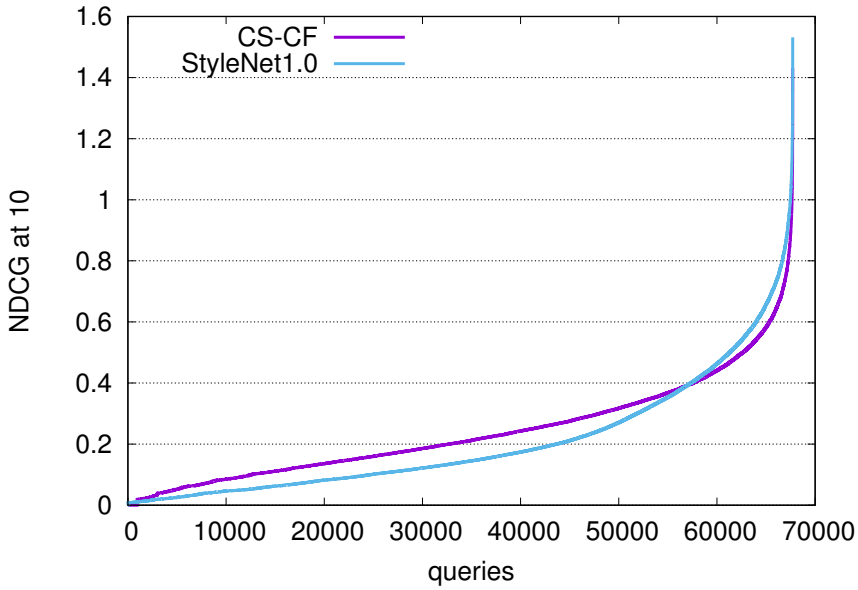

Figure 6.2. CS-CF versus StyleNet1.0 - NDCG@10 numbers for each query.

In order to conduct our analysis we built many balanced rankings, based on the new score function, considering the variation of α value. All curves in the presented results vary according to these values, which are: 0, 0.01, 0.05, 0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8, 0.9. The alpha value represents the user's preference when it comes to fashion popularity, related to the images composing the ranking. In figures, the x axis represents the normalized values for the fashionability measure applied each time, and the y axis represents the NDCG values, considering the new score function for building the balanced ranking. For details related to the score function see Equation 4.2 defined in Chapter 4.

The results, considering posts from the same location of the user and the estimation for fashionability as the number of votes for the candidate image, are presented in Figure 6.3, 6.4 and 6.5 . In general, we can see there is a small loss in NDCG, as soon as it starts the gain related to the fashionability, which represents the expected trend (better observed in Figure 6.5), since these two aspects are, most of times, in non-conformity. China is the country which better illustrates this trend, followed by the United States, which gives confidence to this conclusion, since they are, respectively, one of the great consumers and producers of fashion in the world. Besides, most of the countries show just a little loss in terms of NDCG, considering the gain related to fashionability, which is a good signal. Note that, in all figures, the α value is in each point of the variation curve.

The results considering the estimation for fashionability as the number of fol-

lowers related to the user who posted the candidate image are shown in Figures 6.6, 6.7 and 6.8. Summarily, we can observe, in Figure 6.6), that the values of NDCG decrease considerably when compared to the previous analysis, in which we estimate the fashionability value as the number of votes of the candidate image. Also, in this case, we can observe the same trend, but not so clearly. In Italy, Spain and England, for instance, the NDCG values tend to increase as soon as the fashionability level also increases, which is not a common situation, and could, maybe, indicate that people from these countries are able to make better choices for their looks. Otherwise, in Brazil, Russia, Australia, Greece, Japan and China, we can observe the NDCG values tend to decrease as soon as the fashionability level increases, as expected. In practice, we may say people from these countries tend to be less fashionability-oriented that the others.

Figures 6.9, 6.10 and 6.11 show the results related to the concern about the location of the users. Specifically, Figure 6.9 shows a comparison of results, with and without concerning about the location of the posts, considering the estimation for fashionability as the number of votes for the candidate image. As we can see, the NDCG values decrease considerably when compared to the same experiment, considering only posts from the same location of the users. Although there are specific cases like Italy, for instance, in which we can see the inversion of the expected trend. Also, the loss related to NDCG seems smoother, in some cases.

Summarizing our results, we conclude it is a good choice to make it possible for the user to search for similar looks, considering fashionability since, in general, there is a small loss in terms of NDCG. Indeed, this loss seems to be smoother, as soon as the preference for fashionability increases, which may reflect the actual scenario. Finally, as already mentioned in previous chapters, the use of this parameter helps creating more realistic scenarios of searching (Simo-Serra et al. [2015]). Although, we may assume the limitation of using only NDCG as a metric for the quality of the search, since it ignores issues related to users' satisfaction, which we try to achieve through the use of fashionability parameter.

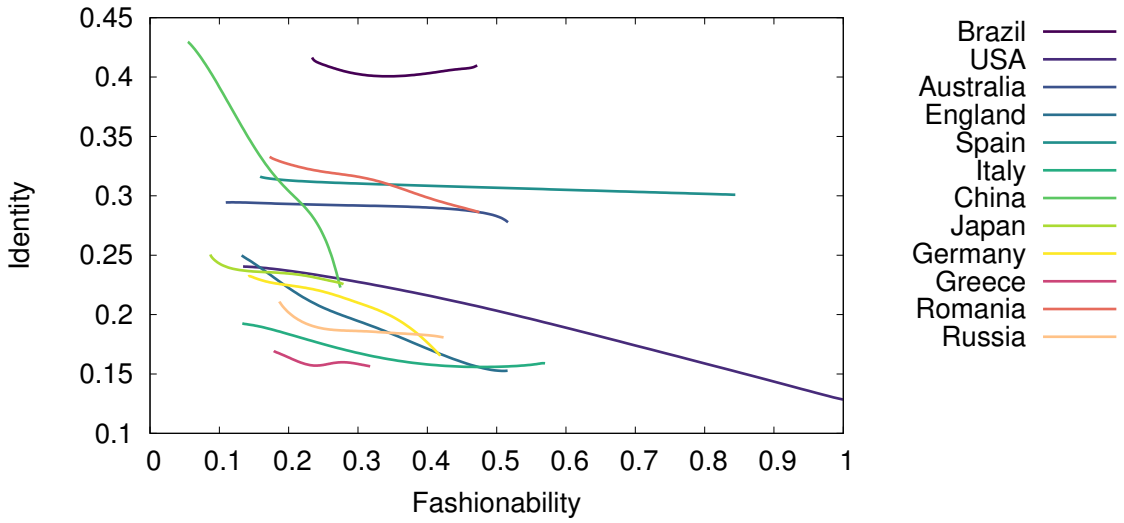

Figure 6.3. Identity versus fashionability - NDCG@1 and the number of votes for the candidate image, considering posts from the same location of the user.

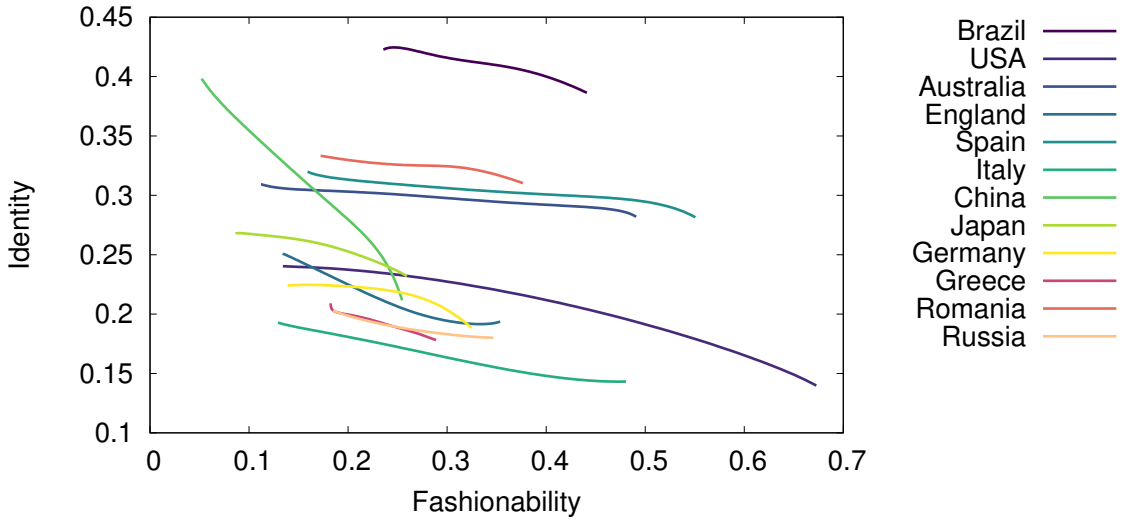

Figure 6.4. Identity versus fashionability - NDCG@5 and the number of votes for the candidate image, considering posts from the same location of the user.

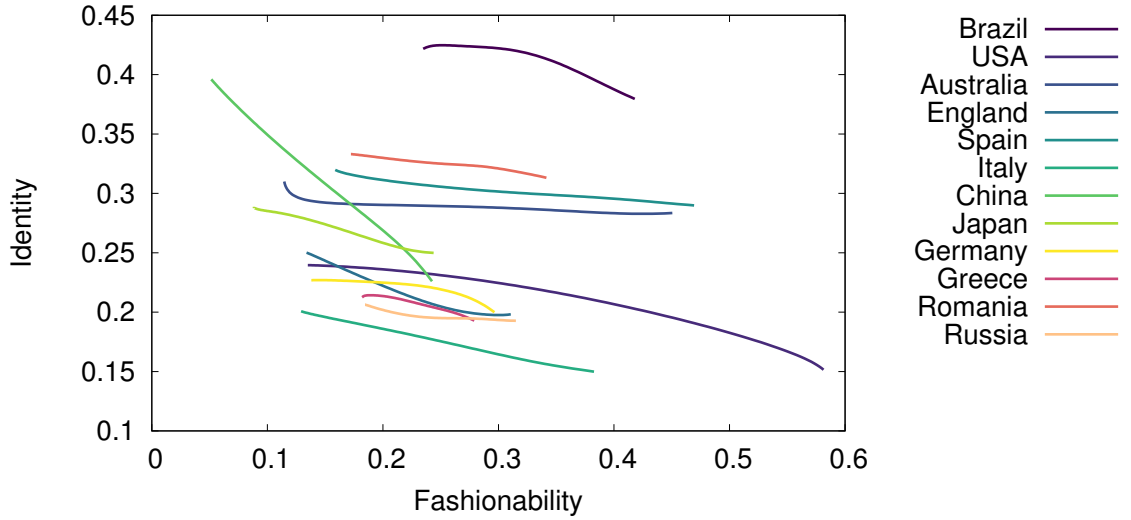

Figure 6.5. Identity versus fashionability - NDCG@10 and the number of votes for the candidate image, considering posts from the same location of the user.

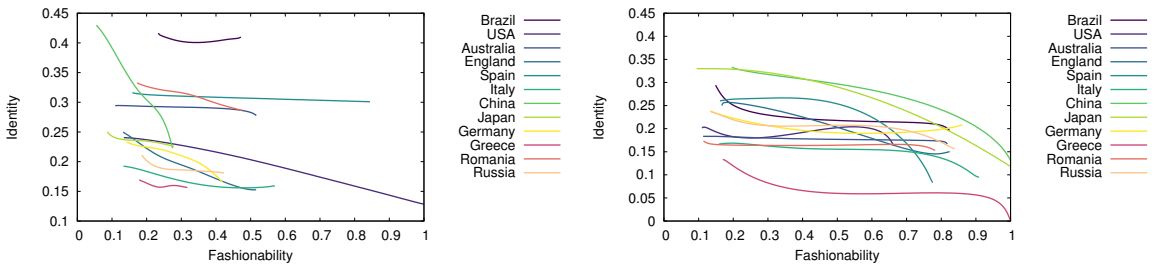

Figure 6.6. The decrease of NDCG. NDCG@1 and the number of votes for the candidate image (Left), NDCG@1 and the number of followers of the user who posted the candidate image (Right), considering posts from the same location of the user.

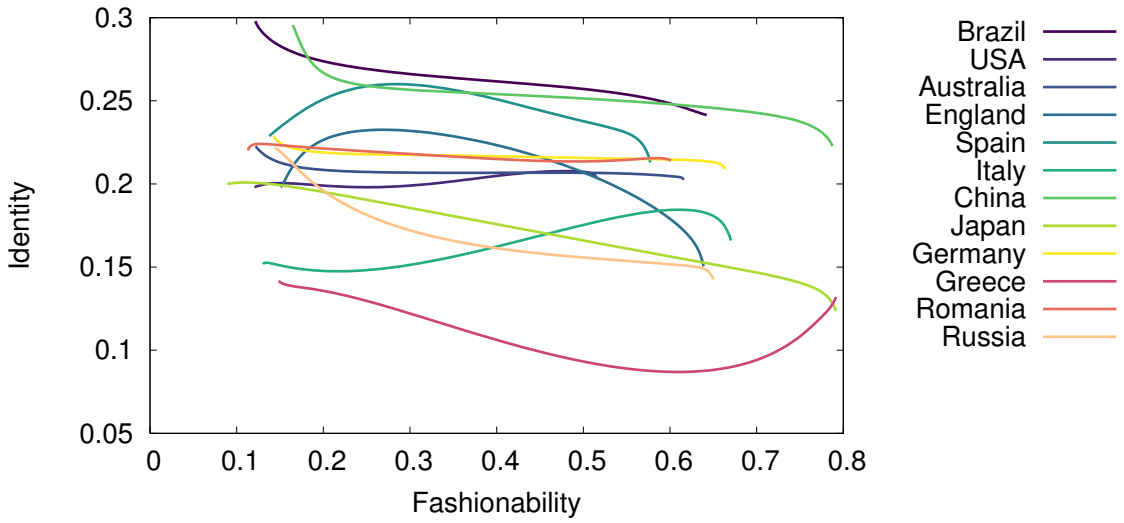

Figure 6.7. Identity versus fashionability - NDCG@5 and the number of followers of the user who posted the candidate image.

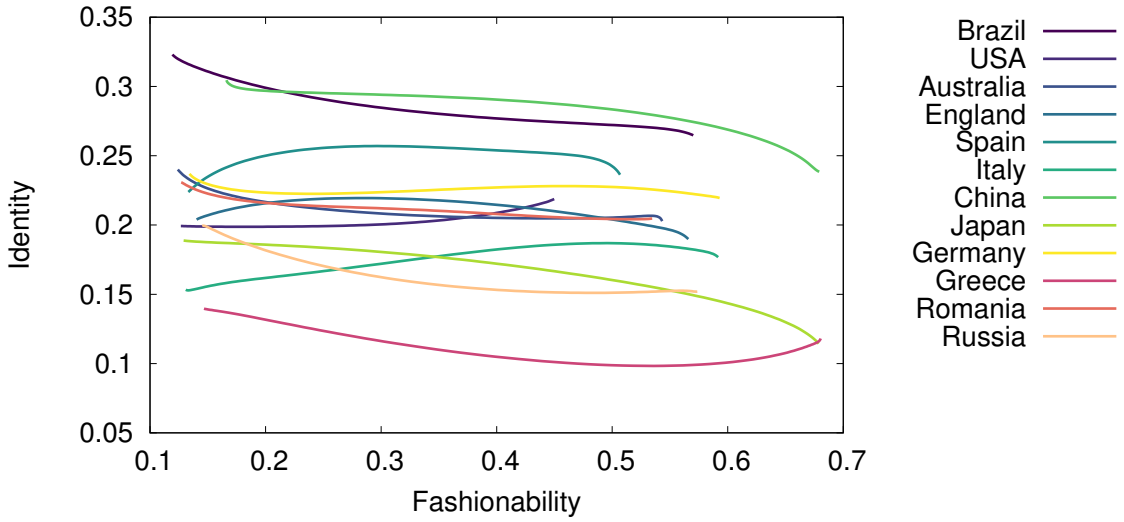

Figure 6.8. Identity versus fashionability - NDCG@10 and the number of followers of the user who posted the candidate image.

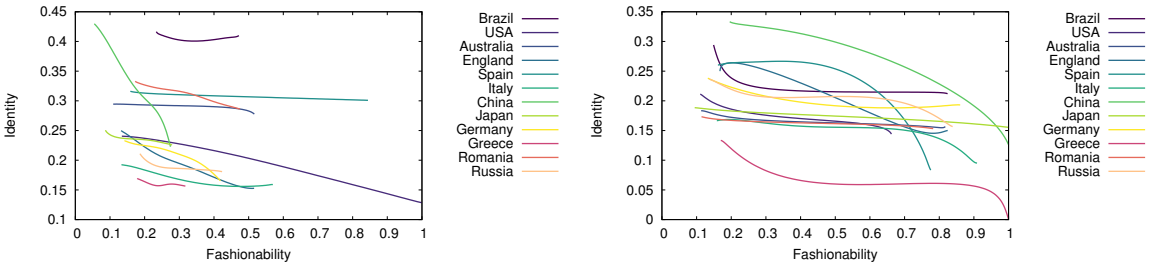

Figure 6.9. Identity versus fashionability. NDCG@1 and the number of votes for the candidate image, considering posts from the same location of the user (Left) and without this concern (Right)

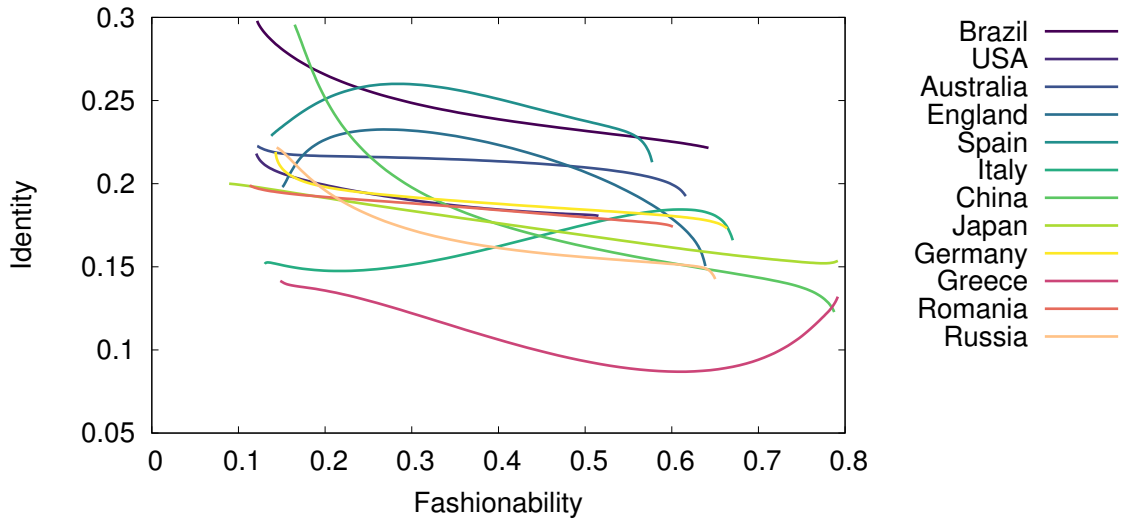

Figure 6.10. Identity versus fashionability - NDCG@5 and the number of votes for the candidate image.

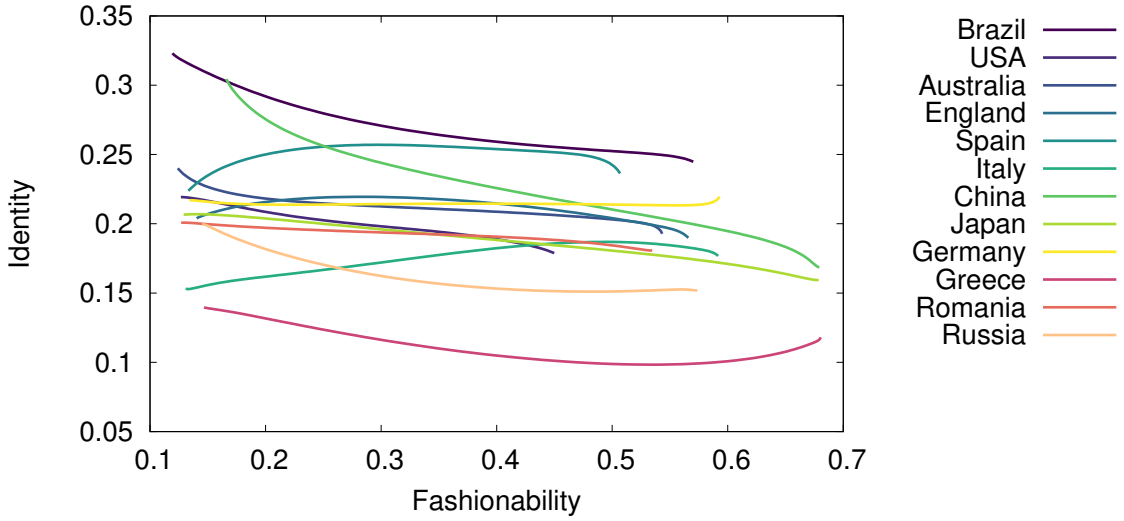

Figure 6.11. Identity versus fashionability - NDCG@10 and the number of votes for the candidate image.

Chapter 7

Conclusion and Future Work

This chapter presents our conclusions and the future directions we could glimpse for this thesis. The first section presents a review of our statement and contributions, aiming to contextualize our results. The second section presents the next steps one may conduct in order to improve or extend the results of this thesis.

7.1 Conclusions of this Thesis

The aim of this thesis is to build a visual search model which works by comparing outfits in a semantic space. Regarding the CBIR area, Rafiee et al. [2010], Tunga et al. [2015] and other similar works from the literature agree that the main open problems are related to image understanding, and bridging the semantic gap is considered a challenging problem yet to be solved (Wang et al. [2010]). This is specially important in the fashion context, regarding the subjectivity related to the concept of outfits. In this context, this thesis presents advances, trying to reduce the semantic gap, approximating the concept of an outfit through its constituent pieces of clothes, applying the principle of compositionality. This principle allows us to learn feature vectors for accurately representing outfits based solely on the occurrences of clothing items.

Our results confirm our main hypothesis that the the principle of compositionality is a determining factor for improving representation learning, and consequently, the retrieval effectiveness. We showe that the lowest ranking performance is attributed to the low-level descriptors approach, used along with a learning to rank algorithm [Moreira et al., 2014], which may indicate the representation of outfits, considering only features like color, shape and texture, is not a good choice, since our ranking approach is reasonably simple. Regarding the retrieval performance, we also show that MAP numbers achieved by our CS−CF model are significantly higher than MAP

numbers achieved by StyleNet–1.0 (Simo-Serra and Ishikawa [2016]), the representative state-of-the-art in fashion retrieval. Further, CS–CF performs better than StyleNet–1.0 in the topmost positions. In this way, we can affirm our CNN Model is considered a good approach for learning the representation of outfits, with a good impact in retrieval performance.

Frequently, users from fashion blogs want to be inspired by popular outfits, but most times these looks do not match his or her identity. With this concern, this thesis also presents another contribution in order to improve the effectiveness related to the search of potentially inspiring outfits. We formulate the search procedure as a multi-objective problem in which outfits are ranked based on a proper balance, conducted by the user, considering two important fashion-related concepts: visual identity and fashionability. We claim this balance is advantageous and should be taken into consideration during the search, so that the user's actual needs can be reflected in the final ranking. Analyzing the results related to the estimates for fashionability, Figures 6.5 and 6.8 for instance, we may conclude fashionability is better estimated through the number of likes in the post related to a look than the number of followers related to the user who posted the look, considering the NDCG. Also, our results show it can be a good choice to apply the concept of fashionability in the search. Actually, through our experiments we confirm it is possible to bring many popular fashion looks to the top rank positions, also matching, reasonably, the user's identity.

Finally, this thesis presents a world-scale analysis of identity and fashionability, with the hypothesis that the search for outfits should be conducted considering user's living place, because culture and lifestyle vary among countries and may impact the choice of outfits. Our results show that each country presents its own pattern, many times, differing significantly from each other. Also, our results show a considerably gain in terms of NDCG, considering the experiments conducted with posts from the same location of the user when compared to the other experiments, without this concern.

7.2 Future Work

We could glimpse the following directions for future work:

- The development of a prototype application that enables a user to send a query image referring to an outfit and defines her or his preferences related to visual identity and fashionability, getting a ranking of similar images, as result. With this app, it is possible to view, analyze and evaluate the results obtained in this thesis more intuitively.

- Our relevance judgment may be considered too strict in the sense that we do not take into account any possible relationship between different styles, occasions and seasons. Thus, an outfit must be associated with only one style, and there is only one occasion and only one season for which it is appropriate. As a result, relevance vanishes if there is not an exact match between the semantics of the query and the semantics of the returned outfit. As future work we plan to take into account the relationship between different styles, occasions and seasons. This means that an outfit that is suitable for a wedding may also be suitable (to some extent) for a graduation party. We may employ the SkipGram algorithm (Mikolov et al. [2013]) in order to measure the extent to which different styles, occasions and seasons are related to each other, and relevance may be assessed by considering the relationship between them.

- Our model does not take into account possible co-occurrence patterns between different clothing items. This information is valuable since we may exploit co-occurrence patterns in different ways in order to improve the CNN representations. Firstly, if the network gets confused between skirts and dresses, then the occurrence of tops can be used to increase the odds of skirts. To model these co-occurrence patterns, we intend to create a normalization layer which would update the probabilities of clothing items by taking into account the co-occurrence information between them. Further, in order to help the network to better distinguish between mutually exclusive clothing items, we intend to devise a loss function which puts a higher cost when mutually exclusive items are missclassified.

7.3 Limitations of this Thesis

This section presents the limitations regarding this thesis. According to our approach and experiments, we may assume that:

- NDCG is a non-realistic measure, when it comes to reflect users' satisfaction. In order to support our results, it is important to conduct experiments with real users, aiming to discover their preferences when searching for fashion looks.

- Our experiments were conducted using only one architecture model and one dataset. In order to give more credibility to the evaluation of our compositional approach, it is important to apply it in different datasets and implement it using other architecture models.

- We try to approximate the representation of fashionability considering the number of likes and followers, but it should be analyzed more deeply, according to our results, since different patterns were found and conclusions are still subjective.
- Our CNN model considers the clothing items in isolation. It could be improved by taking into account possible co-occurrence and relationship between them. The same occurs in our ranking model, which could be less strict, considering the relationship among styles, occasions and seasons.

Bibliography

Baldi, P. and Hornik, K. (1989). Neural networks and principal component analysis: Learning from examples without local minima. *Neural Netw.*, 2(1):53--58. ISSN 0893-6080.

Bay, H., Ess, A., Tuytelaars, T., and Van Gool, L. (2008). Speeded-up robust features (surf). *Comput. Vis. Image Underst.*, 110(3):346--359. ISSN 1077-3142.

Bengio, Y., Courville, A. C., and Vincent, P. (2012). Unsupervised feature learning and deep learning: A review and new perspectives. *CoRR*, abs/1206.5538.

Callan, G. O. (2007). *Enciclopedia da Moda*. Companhia das Letras. ISBN 8535909567.

Cheng, C.-I. and Liu, D. S.-M. (2008). An intelligent clothes search system based on fashion styles. In *2008 International Conference on Machine Learning and Cybernetics*, volume 3, pages 1592–1597. ISSN 2160-133X.

Cun, Y. L., Boser, B., Denker, J. S., Howard, R. E., Habbard, W., Jackel, L. D., and Henderson, D. (1990). Advances in neural information processing systems 2. chapter Handwritten Digit Recognition with a Back-propagation Network, pages 396--404. Morgan Kaufmann Publishers Inc., San Francisco, CA, USA.

Date, P., Ganesan, A., and Oates, T. (2017). Fashioning with networks: Neural style transfer to design clothes. *CoRR*, abs/1707.09899.

Di, W., Wah, C., Bhardwaj, A., Piramuthu, R., and Sundaresan, N. (2013). Style finder: Fine-grained clothing style detection and retrieval. In *2013 IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pages 8–13. ISSN 2160-7508.

Eytan, D. (2016). Are fashion bloggers able to convert followers into buyers?

Faria, F. F., Veloso, A., Almeida, H. M., Valle, E., Torres, R. d. S., Gonçalves, M. A., and Meira, Jr., W. (2010). Learning to rank for content-based image retrieval. In

Proceedings of the International Conference on Multimedia Information Retrieval, MIR '10, pages 285--294, New York, NY, USA. ACM.

Ferrara, E., Interdonato, R., and Tagarelli, A. (2014). Online popularity and topical interests through the lens of instagram. In *Proceedings of the 25th ACM Conference on Hypertext and Social Media*, HT '14, pages 24--34, New York, NY, USA. ACM.

Fu, J., Wang, J., Li, Z., Xu, M., and Lu, H. (2013). *Efficient Clothing Retrieval with Semantic-Preserving Visual Phrases*, pages 420--431. Springer Berlin Heidelberg, Berlin, Heidelberg.

Hassan Zadeh, A. and Sharda, R. (2014). Modeling brand post popularity dynamics in online social networks. *Decis. Support Syst.*, 65(C):59--68. ISSN 0167-9236.

He, Y., Wang, J., Kang, C., Xiang, S., and Pan, C. (2015). Large scale image annotation via deep representation learning and tag embedding learning. In *Proceedings of the 5th ACM on International Conference on Multimedia Retrieval*, ICMR '15, pages 523--526, New York, NY, USA. ACM.

Hidayati, S. C., Cheng, W.-H., and Hua, K.-L. (2012). Clothing genre classification by exploiting the style elements. In *Proceedings of the 20th ACM International Conference on Multimedia*, MM '12, pages 1137--1140, New York, NY, USA. ACM.

Huang, J., Kumar, S. R., Mitra, M., Zhu, W.-J., and Zabih, R. (1997). Image indexing using color correlograms. In *Proceedings of the 1997 Conference on Computer Vision and Pattern Recognition (CVPR '97)*, CVPR '97, pages 762--, Washington, DC, USA. IEEE Computer Society.

Huang, J., Xia, W., and Yan, S. (2014). Deep search with attribute-aware deep network. In *Proceedings of the 22Nd ACM International Conference on Multimedia*, MM '14, pages 731--732, New York, NY, USA. ACM.

Iliukovich-Strakovskaia, A., Dral, A., and Dral, E. (2016). Using pre-trained models for fine-grained image classification in fashion field. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '16, New York, NY, USA. ACM.

Iwata, T., Watanabe, S., and Sawada, H. (2011). Fashion coordinates recommender system using photographs from fashion magazines. In *Proceedings of the Twenty-Second International Joint Conference on Artificial Intelligence - Volume Volume Three*, IJCAI'11, pages 2262--2267. AAAI Press.

Jagadeesh, V., Piramuthu, R., Bhardwaj, A., Di, W., and Sundaresan, N. (2014). Large scale visual recommendations from street fashion images. In *Proceedings of the 20th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '14, pages 1925--1934, New York, NY, USA. ACM.

Järvelin, K. and Kekäläinen, J. (2002). Cumulated gain-based evaluation of ir techniques. *ACM Trans. Inf. Syst.*, 20(4):422--446. ISSN 1046-8188.

Ji, X., Wang, W., Zhang, M., and Yang, Y. (2017). Cross-domain image retrieval with attention modeling. *CoRR*, abs/1709.01784.

Jing, Y., Liu, D., Kislyuk, D., Zhai, A., Xu, J., Donahue, J., and Tavel, S. (2015). Visual search at pinterest. In *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '15, pages 1889--1898, New York, NY, USA. ACM.

Kalantidis, Y., Kennedy, L., and Li, L.-J. (2013). Getting the look: Clothing recognition and segmentation for automatic product suggestions in everyday photos. In *Proceedings of the 3rd ACM Conference on International Conference on Multimedia Retrieval*, ICMR '13, pages 105--112, New York, NY, USA. ACM.

Keogh, E. and Mueen, A. (2010). *Curse of Dimensionality*, pages 257--258. Springer US, Boston, MA.

Khokher, A. and Talwar, R. (2011). Content-based image retrieval: State-of-the-art and challenges. *International Journal of Engineering Trends and Technology (IJETT)*, 9(2):207–211. ISSN 2230-7818.

Kiapour, M. H., Yamaguchi, K., Berg, A. C., and Berg, T. L. (2014). *Hipster Wars: Discovering Elements of Fashion Styles*, pages 472--488. Springer International Publishing, Cham.

Krizhevsky, A., Sutskever, I., and Hinton, G. E. (2012a). Imagenet classification with deep convolutional neural networks. In *Proceedings of the 25th International Conference on Neural Information Processing Systems*, NIPS'12, pages 1097--1105, USA. Curran Associates Inc.

Krizhevsky, A., Sutskever, I., and Hinton, G. E. (2012b). Imagenet classification with deep convolutional neural networks. In Pereira, F., Burges, C. J. C., Bottou, L., and Weinberger, K. Q., editors, *Advances in Neural Information Processing Systems 25*, pages 1097--1105. Curran Associates, Inc.

Lecun, Y., Bottou, L., Bengio, Y., and Haffner, P. (1998). Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324. ISSN 0018-9219.

Lee, C.-J., Lin, Y.-C., Chen, R.-C., and Cheng, P.-J. (2009). Selecting effective terms for query formulation. In *Proceedings of the 5th Asia Information Retrieval Symposium on Information Retrieval Technology*, AIRS '09, pages 168--180, Berlin, Heidelberg. Springer-Verlag.

Lee, H. and Lee, S. (2015). Style recommendation for fashion items using heterogeneous information network. In *Poster Proceedings of the 9th ACM Conference on Recommender Systems, RecSys 2015, Vienna, Austria, September 16, 2015.*

Lew, M. S., Sebe, N., Djeraba, C., and Jain, R. (2006). Content-based multimedia information retrieval: State of the art and challenges. *ACM Trans. Multimedia Comput. Commun. Appl.*, 2(1):1--19. ISSN 1551-6857.

Lin, K., Yang, H.-F., Liu, K.-H., Hsiao, J.-H., and Chen, C.-S. (2015). Rapid clothing retrieval via deep learning of binary codes and hierarchical search. In *Proceedings of the 5th ACM on International Conference on Multimedia Retrieval*, ICMR '15, pages 499--502, New York, NY, USA. ACM.

Liu, S., Song, Z., Wang, M., Xu, C., Lu, H., and Yan, S. (2012). Street-to-shop: Cross-scenario clothing retrieval via parts alignment and auxiliary set. In *Proceedings of the 20th ACM International Conference on Multimedia*, MM '12, pages 1335--1336, New York, NY, USA. ACM.

Liu, Y., Xu, J., Qin, T., Xiong, W., and Li, H. (2007). LETOR: Benchmark dataset for research on learning to rank for information retrieval. In *Learning to Rank Workshop in conjuntion with SIGIR.*

Liu, Z., Luo, P., Qiu, S., Wang, X., and Tang, X. (2016). Deepfashion: Powering robust clothes recognition and retrieval with rich annotations. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR).*

Lowe, D. G. (2004). Distinctive image features from scale-invariant keypoints. *Int. J. Comput. Vision*, 60(2):91--110. ISSN 0920-5691.

Lurie, A. (2000). *The Language of Clothes.* Holt Paperbacks. ISBN 978-0805062441.

Mahmoudi, F., Shanbehzadeh, J., Eftekhari-Moghadam, A.-M., and Soltanian-Zadeh, H. (2003). Image retrieval based on shape similarity by edge orientation autocorrelogram. *Pattern Recognition*, 36(8):1725 – 1736. ISSN 0031-3203.

Marques, O. (2016). Visual information retrieval: The state of the art. *IT Professional*, 18(4):7–9. ISSN 1520-9202.

Matzen, K., Bala, K., and Snavely, N. (2017). Streetstyle: Exploring world-wide clothing styles from millions of photos. *CoRR*, abs/1706.01869.

Mikolov, T., Sutskever, I., Chen, K., Corrado, G., and Dean, J. (2013). Distributed representations of words and phrases and their compositionality. In *Advances in Neural Information Processing Systems 26: 27th Annual Conference on Neural Information Processing Systems 2013*, pages 3111--3119.

Moreira, M., dos Santos, J. A., and Veloso, A. (2014). Learning to rank similar apparel styles with economically-efficient rule-based active learning. In *Proceedings of International Conference on Multimedia Retrieval*, ICMR '14, pages 361:361--361:368, New York, NY, USA. ACM.

Murthy, V. N., Can, E. F., and Manmatha, R. (2014). A hybrid model for automatic image annotation. In *Proceedings of International Conference on Multimedia Retrieval*, ICMR '14, pages 369:369--369:376, New York, NY, USA. ACM.

Okada, C. Y., Pedronette, D. C. G. a., and da S. Torres, R. (2015). Unsupervised distance learning by rank correlation measures for image retrieval. In *Proceedings of the 5th ACM on International Conference on Multimedia Retrieval*, ICMR '15, pages 331--338, New York, NY, USA. ACM.

Pass, G., Zabih, R., and Miller, J. (1996). Comparing images using color coherence vectors. In *Proceedings of the Fourth ACM International Conference on Multimedia*, MULTIMEDIA '96, pages 65--73, New York, NY, USA. ACM.

Rafiee, G., Dlay, S. S., and Woo, W. L. (2010). A review of content-based image retrieval. In *2010 7th International Symposium on Communication Systems, Networks Digital Signal Processing (CSNDSP 2010)*, pages 775–779.

Rumelhart, D. E., McClelland, J. L., and PDP Research Group, C., editors (1986). *Parallel Distributed Processing: Explorations in the Microstructure of Cognition, Vol. 1: Foundations*. MIT Press, Cambridge, MA, USA. ISBN 0-262-68053-X.

Schmidt, R., Möhring, M., Härting, R.-C., Reichstein, C., and Keller, B. (2016). *Influencing Factors Increasing Popularity on Facebook – Empirical Insights from European Users*, pages 383--394. Springer International Publishing, Cham.

Sedeke, K. (2012). Effective fashion blogs and their impact on the current fashion industry. Master's thesis.

Sermanet, P., Kavukcuoglu, K., Chintala, S., and Lecun, Y. (2013). Pedestrian detection with unsupervised multi-stage feature learning. In *Proceedings of the 2013 IEEE Conference on Computer Vision and Pattern Recognition*, CVPR '13, pages 3626--3633, Washington, DC, USA. IEEE Computer Society.

Sermanet, P. and LeCun, Y. (2011). Traffic sign recognition with multi-scale convolutional networks. In *The 2011 International Joint Conference on Neural Networks*, pages 2809–2813. ISSN 2161-4393.

Shen, E., Lieberman, H., and Lam, F. (2007). What am i gonna wear?: Scenario-oriented recommendation. In *Proceedings of the 12th International Conference on Intelligent User Interfaces*, IUI '07, pages 365--368, New York, NY, USA. ACM.

Sheshasaayee, A. and .C, J. (2014). Relevance feedback techniques implemented in cbir: Current trends and issues. *International Journal of Engineering Trends and Technology (IJETT)*, 10(4):166–175. ISSN 2231-5381.

Simo-Serra, E., Fidler, S., Moreno-Noguer, F., and Urtasun, R. (2015). Neuroaesthetics in fashion: Modeling the perception of fashionability. In *CVPR*, pages 869–877. IEEE Computer Society.

Simo-Serra, E. and Ishikawa, H. (2016). Fashion style in 128 floats: Joint ranking and classification using weak data for feature extraction. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 298–307.

Simo-Serra, E. and Ishikawa, H. (2016). Fashion style in 128 floats: Joint ranking and classification using weak data for feature extraction. In *IEEE CVPR Conference on Computer Vision and Pattern Recognition*, pages 298--307.

Simonyan, K. and Zisserman, A. (2015). Very deep convolutional networks for large-scale image recognition. In *ICLR International Conference on Learning Representations*.

Sivic, J. and Zisserman, A. (2003). Video google: A text retrieval approach to object matching in videos. In *Proceedings of the Ninth IEEE International Conference on*

Computer Vision - Volume 2, ICCV '03, pages 1470--, Washington, DC, USA. IEEE Computer Society.

Smirnov, E., Kulinkin, A., and K. Ivanova, M. P. (2016). Deep learning for fast and accurate fashion item detection. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '16, New York, NY, USA. ACM.

Srivastava, N., Hinton, G., Krizhevsky, A., Sutskever, I., and Salakhutdinov, R. (2014). Dropout: A simple way to prevent neural networks from overfitting. *J. Mach. Learn. Res.*, 15(1):1929--1958. ISSN 1532-4435.

Stehling, R. O., Nascimento, M. A., and Falcão, A. X. (2002). A compact and efficient image retrieval approach based on border/interior pixel classification. In *Proceedings of the Eleventh International Conference on Information and Knowledge Management*, CIKM '02, pages 102--109, New York, NY, USA. ACM.

Swain, M. J. and Ballard, D. H. (1991). Color indexing. *Int. J. Comput. Vision*, 7(1):11--32. ISSN 0920-5691.

Szegedy, C., Liu, W., Jia, Y., Sermanet, P., Reed, S. E., Anguelov, D., Erhan, D., Vanhoucke, V., and Rabinovich, A. (2015). Going deeper with convolutions. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2015, Boston, MA, USA, June 7-12, 2015*, pages 1--9.

Tu, Q. and Dong, L. (2010). An intelligent personalized fashion recommendation system. In *2010 International Conference on Communications, Circuits and Systems (ICCCAS)*, pages 479–485.

Tunga, S., D, J., and Gururaj, C. (2015). A comparative study of content based image retrieval trends and approaches. *International Journal of Image Processing (IJIP)*, 9(3):127–155. ISSN 1985-2304.

Unser, M. (1986). Sum and difference histograms for texture classification. *IEEE Trans. Pattern Anal. Mach. Intell.*, 8(1):118--125. ISSN 0162-8828.

Utgoff, P. and Stracuzzi, D. (2002). Many-layered learning. *Neural Computation*, 14:2497--2539.

Vogiatzis, D., Pierrakos, D., Paliouras, G., Jenkyn-Jones, S., and Possen, B. J. H. H. A. (2012). Expert and community based style advice. *Expert Syst. Appl.*, 39(12):10647--10655. ISSN 0957-4174.

Voravuthikunchai, W., Crémilleux, B., and Jurie, F. (2014). Image re-ranking based on statistics of frequent patterns. In *Proceedings of International Conference on Multimedia Retrieval*, ICMR '14, pages 129:129--129:136, New York, NY, USA. ACM.

Wallraven, C., Caputo, B., and Graf, A. (2003). Recognition with local features: the kernel recipe. In *Proceedings Ninth IEEE International Conference on Computer Vision*, pages 257–264 vol.1.

Wan, J., Wang, D., Hoi, S. C. H., Wu, P., Zhu, J., Zhang, Y., and Li, J. (2014). Deep learning for content-based image retrieval: A comprehensive study. In *Proceedings of the 22Nd ACM International Conference on Multimedia*, MM '14, pages 157--166, New York, NY, USA. ACM.

Wang, H. H., Mohamad, D., and Ismail, N. A. (2010). Approaches, challenges and future direction of image retrieval. *CoRR*, abs/1006.4568.

Yamaguchi, K., Kiapour, M. H., Ortiz, L. E., and Berg, T. L. (2015). Retrieving similar styles to parse clothing. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 37(5):1028–1040. ISSN 0162-8828.

Yamin, F. M. and Ramayah, T. (2011). User web search behavior on query formulation. In *2011 International Conference on Semantic Technology and Information Retrieval*, pages 182–188. ISSN 2166-0697.

Zanetti, C. and Resende, F. (2013). *Vista Quem Você é - Descubra e Aperfeiçoe seu Estilo Pessoal*. LEYA, Casa da Palavra. ISBN 8577343650.

Zegarra, J. A. M., Leite, N. J., and da Silva Torres, R. (2009). Wavelet-based fingerprint image retrieval. *Journal of Computational and Applied Mathematics*, 227(2):294 – 307. ISSN 0377-0427. Special Issue on Emergent Applications of Fractals and Wavelets in Biology and Biomedicine.

Zeiler, M. D. and Fergus, R. (2014). *Visualizing and Understanding Convolutional Networks*, pages 818--833. Springer International Publishing, Cham.

Zoe, R. (2008). *Style A to Zoe: The Art of Fashion, Beauty & Everything Glamour*. Grand Central Publishing. ISBN 0446535869.