Ranking Images Based on Aesthetic Qualities

Aarushi Gaur

Submitted for the Degree of
Doctor of Philosophy
from the
University of Surrey

Centre for Vision, Speech and Signal Processing
Faculty of Engineering and Physical Sciences
University of Surrey
Guildford, Surrey GU2 7XH, U.K.

April 2015

© Aarushi Gaur 2015
Summary

The qualitative assessment of image content and aesthetic impression is affected by various image attributes and relations between the attributes. Modelling of such assessments in the form of objective rankings and learning image representations based on them is not a straightforward problem. The criteria can be varied with different levels of complexity for various applications. A highly-complex problem could involve a large number of interrelated attributes and features alongside varied rules. An example of such an application is fashion-interpretation. In this case one can use attribute recognition to label different parts such as clothing and body shape automatically. Thus, the presence or absence of objects in the image is not ambiguous and a similarity measure can be established between images. It is however not clear how to establish such measure between the aesthetic impressions the images make.

As a first contribution an approach for ranking images by pooling from the knowledge and experience of crowdsourced annotators is presented. Specifically, the highly subjective and complex problem of fashion interpretation and assessment of aesthetic qualities of images is addressed. To utilize the visual judgements, a novel dataset complete with labellings of various attributes of clothing and body shapes is introduced. Large scale pairwise comparisons of the order of tens of thousands are performed by annotators who follow fashion. Various consistency measures are then applied to verify agreement and correlation between the annotators to rule out inconsistencies amongst them. Based on the annotations, reliable rankings to automatically compare images according to fashion rules are established.

Then, Bag of Visual Words object recognition is used to perform classification of the attributes. By incorporating annotator rankings from the first stage and these classification estimates in a lookup model for automatic assessment of images, pairwise comparisons can be automatically performed. Each visual attribute of clothing and body shape is represented within the rankings.

Next, rankings obtained from the crowdsourcing procedure are included within several matching approaches to achieve a matching-based ranking. Nearest neighbour matches can be found for a pair of test images which can be compared with their rankings from the annotators. This can be utilized to establish which configuration is ranked better in an image pair. In particular, two prominent approaches of Bag of Visual Words and Local Descriptor Matching are employed to facilitate an evaluation. Several random splits of the dataset proposed in the first stage are used to form the training and test sets. Matches obtained are incorporated within an approach introduced to generate a global ranking. Evaluation from this stage is used as a comparative basis for the approach proposed next in which a learning procedure based on graphical modelling captures the annotator rankings.

Finally, a novel approach for learning image representation based on qualitative assessments of visual aesthetics is proposed. It relies on a multi-node multi-state model that represents image attributes and their relations. The model is learnt from pairwise image preferences provided by annotators. To demonstrate the effectiveness the approach is applied to fashion image rating, i.e., comparative assessment of aesthetic qualities. The attributes and their relations are assigned learnt potentials which are used to rate the images. Evaluation of the representation model has demonstrated a high performance rate in ranking fashion images.
Key words: Visual assessment, Aesthetics, Attributes, Annotation, Ranking, Matching, Graphical modelling, Image representation.

Email: a.gaur@surrey.ac.uk
Acknowledgements

I would like to thank my supervisor Dr. Krystian Mikolajczyk for his guidance in this thesis. In addition I would like to thank University of Surrey, CVSSP and its staff for providing me with support for this research.
Declaration of Originality

This thesis and the work to which it refers are the results of my own efforts. Any ideas, data, images or text resulting from the work of others (whether published or unpublished) are fully identified as such within the work and attributed to their originator in the text, bibliography or in footnotes. This thesis has not been submitted in whole or in part for any other academic degree or professional qualification. I agree that the University has the right to submit my work to the plagiarism detection service TurnitinUK for originality checks. Whether or not drafts have been so-assessed, the University reserves the right to require an electronic version of the final document (as submitted) for assessment as above.

Author Signature: Aarushi Gaur

Date: 9/4/15
# Contents

1 Introduction ........................................... 1
   1.1 Applications ........................................ 4
   1.2 Challenges ........................................... 5
   1.3 Objectives ........................................... 7
   1.4 Overview ............................................ 8

2 Related work ........................................ 9
   2.1 Introduction ........................................... 9
   2.2 Crowdsourcing ........................................ 10
      2.2.1 Image rating ..................................... 11
      2.2.2 Annotation noise ................................ 12
      2.2.3 Ranking from preferences ...................... 13
      2.2.4 Applications .................................... 15
   2.3 Image matching ....................................... 16
      2.3.1 Existing approaches ............................. 16
      2.3.2 Recent developments ............................ 20
   2.4 Visual assessment for aesthetics .................. 21
      2.4.1 Quality assessment .............................. 21
      2.4.2 Aesthetics ....................................... 22
      2.4.3 Datasets ......................................... 23
   2.5 Summary ............................................. 25
<table>
<thead>
<tr>
<th>Section</th>
<th>Content</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>3 The Dataset</strong></td>
<td>29</td>
</tr>
<tr>
<td>3.1 Dataset</td>
<td>29</td>
</tr>
<tr>
<td>3.1.1 Dataset and labelling</td>
<td>30</td>
</tr>
<tr>
<td>3.2 Aesthetic comparisons</td>
<td>31</td>
</tr>
<tr>
<td>3.2.1 Control sets</td>
<td>32</td>
</tr>
<tr>
<td>3.3 Crowdsourcing for qualitative assessments</td>
<td>33</td>
</tr>
<tr>
<td>3.3.1 Selection criteria</td>
<td>33</td>
</tr>
<tr>
<td>3.3.2 Annotation task</td>
<td>35</td>
</tr>
<tr>
<td>3.4 Evaluation of annotations</td>
<td>35</td>
</tr>
<tr>
<td>3.4.1 Agreement amongst annotators</td>
<td>36</td>
</tr>
<tr>
<td>3.4.2 Annotator validation</td>
<td>37</td>
</tr>
<tr>
<td>3.5 Rankings based on pairwise comparisons</td>
<td>37</td>
</tr>
<tr>
<td>3.5.1 Kemeny-Young method</td>
<td>38</td>
</tr>
<tr>
<td>3.5.2 Rankings from annotators</td>
<td>39</td>
</tr>
<tr>
<td>3.6 Experimental results</td>
<td>40</td>
</tr>
<tr>
<td>3.6.1 Criteria for agreement</td>
<td>40</td>
</tr>
<tr>
<td>3.6.2 Ranking using Kemeny-Young method</td>
<td>41</td>
</tr>
<tr>
<td>3.6.3 Annotator vs. expert agreement</td>
<td>43</td>
</tr>
<tr>
<td>3.6.4 Inter-annotator agreement</td>
<td>45</td>
</tr>
<tr>
<td>3.6.5 Rankings from annotators</td>
<td>48</td>
</tr>
<tr>
<td>3.7 Conclusions</td>
<td>50</td>
</tr>
<tr>
<td><strong>4 Detecting attributes automatically</strong></td>
<td>51</td>
</tr>
<tr>
<td>4.1 Attribute recognition</td>
<td>51</td>
</tr>
<tr>
<td>4.1.1 Performance evaluation</td>
<td>52</td>
</tr>
<tr>
<td>4.2 Rankings based on attribute recognition</td>
<td>53</td>
</tr>
<tr>
<td>4.2.1 Lookup model</td>
<td>54</td>
</tr>
<tr>
<td>4.2.2 Accuracy measure</td>
<td>54</td>
</tr>
<tr>
<td>4.3 Experimental results</td>
<td>54</td>
</tr>
<tr>
<td>4.3.1 Attribute recognition</td>
<td>55</td>
</tr>
<tr>
<td>4.3.2 Rankings with attribute recognition</td>
<td>57</td>
</tr>
<tr>
<td>4.4 Conclusions</td>
<td>59</td>
</tr>
</tbody>
</table>
5 Ranking images using matching

5.1 Background .............................................. 61
  5.1.1 Matching with Bag of Visual Words (BoV) .......... 62
  5.1.2 Matching with Local Descriptor Matching (LDM) .... 64
  5.1.3 Global ranking based on matching ............... 65

5.2 Evaluation design for ranking .................. 66
  5.2.1 Dataset ........................................ 66
  5.2.2 Measures ...................................... 67
  5.2.3 Setup ......................................... 69

5.3 Experimental results ....................... 70
  5.3.1 BoV with inverted file ....................... 71
  5.3.2 BoV with min-Hash ......................... 74
  5.3.3 Local Descriptor Matching (LDM) ............ 78
  5.3.4 Global ranking using LDM ................. 79
  5.3.5 Summary ..................................... 80

5.4 Conclusions ........................................ 81

6 Ranking images by applying graphical modelling

6.1 Image model: learning and recognition .... 83
  6.1.1 Graph based model ......................... 84
  6.1.2 Learning image ranking ................... 85
  6.1.3 Learning attributes potentials ............ 85
  6.1.4 Fashion aesthetics ......................... 86

6.2 Graph based model example .................. 87
  6.2.1 Node potentials ......................... 89
  6.2.2 Edge potentials .......................... 89
  6.2.3 Final product potential .................. 91
  6.2.4 Correlation Coefficient .................. 92

6.3 Ranking based on automatic labelling .... 93

6.4 Experimental results ....................... 94
  6.4.1 Individual annotator ranking correlation .... 95
  6.4.2 Rankings from learnt attribute potentials ..... 96
6.4.3 Attribute recognition ........................................... 97
6.4.4 Rankings with random predictive estimates of attributes .... 98
6.4.5 Rankings with attribute recognition ............................... 99
6.4.6 Rankings correlations ........................................... 100
6.4.7 Attributes potentials ............................................ 100
6.4.8 Ranking based on learnt attribute potentials ................. 103
6.4.9 Image ranking with attribute recognition ....................... 104

6.5 Comparison with baselines ....................................... 105
6.6 Conclusions ..................................................... 105

7 Discussion ........................................................ 107

7.1 Summary ......................................................... 107
  7.1.1 Crowdsourced judgements .................................. 108
  7.1.2 Attribute recognition ....................................... 109
  7.1.3 Matching based evaluation .................................. 110
  7.1.4 Graphical modelling ........................................ 111
  7.2 Future work .................................................... 112

A Dataset: Aesthetics Based on Fashion Images .................... 115

A.1 Images ......................................................... 115
A.2 Aesthetic comparisons .......................................... 116

Bibliography ...................................................... 119
List of Figures

1.1 Problem of ranking n image configurations based on aesthetics into an ordered list. .............................................. 1
1.2 Application of the approach to fashion interpretation. ......................................................... 2
1.3 Some applications of the proposed approaches. ................................................................. 4
1.4 Challenges associated with attributes. ................................................................. 6
1.5 Challenges due to the requirement of meeting a vast range of fashion rules. .................. 6
1.6 An outline of the chapters of the thesis. ................................................................. 8

2.1 Some example images from the released datasets. .................................................. 25
2.2 Some example images from the datasets that have not been released yet. .......... 26

3.1 Clothing and body shape attributes in the dataset. ............................................. 30
3.2 Some example images from the dataset. ................................................................. 31
3.3 Inclusion of control sets within 10 annotation lists. .............................................. 32
3.4 Selection criteria and information provided to the participants for performing the crowdsourcing task. ................................................................. 34
3.5 The tool that was used to perform annotations. .................................................. 35
3.6 Accuracy, Cohen’s κ and Kendall’s τ when comparing decision results. .................. 41
3.7 Accuracy, Cohen’s κ and Kendall’s τ between expert and 10 annotators. ........ 44
3.8 Images that illustrate the ordering over image space. ........................................... 48
3.9 Images that illustrate the ordering over attribute space. ........................................... 49

4.1 Images showing correct recognition and false categorisations for top clothing attributes. ................................................................. 55
4.2 Images showing correct recognition and false categorisations for bottom clothing attributes. ................................................................. 57
4.3 Correctly ranked pairs when using ranking with the recognition based error estimates. ................................................................. 58
<table>
<thead>
<tr>
<th>Figure</th>
<th>Description</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>5.1</td>
<td>General process of achieving image matching.</td>
<td>62</td>
</tr>
<tr>
<td>5.2</td>
<td>Steps for creating BoV.</td>
<td>62</td>
</tr>
<tr>
<td>5.3</td>
<td>Process of generating a voting file that is input to the Kemeny-Young method.</td>
<td>66</td>
</tr>
<tr>
<td>5.4</td>
<td>Testing using the absolute accuracy measure.</td>
<td>68</td>
</tr>
<tr>
<td>6.1</td>
<td>Object representation model for modelling ranked lists.</td>
<td>84</td>
</tr>
<tr>
<td>6.2</td>
<td>Example for learning node and edge potential from a ranking.</td>
<td>87</td>
</tr>
<tr>
<td>6.3</td>
<td>Performance of attribute recognition using recall and precision.</td>
<td>97</td>
</tr>
<tr>
<td>6.4</td>
<td>Kendall’s $\tau$ correlation between strong, weak and random annotators.</td>
<td>100</td>
</tr>
<tr>
<td>6.5</td>
<td>Attributes potentials learnt from rankings.</td>
<td>101</td>
</tr>
<tr>
<td>6.6</td>
<td>Attribute relations potentials learnt from rankings.</td>
<td>102</td>
</tr>
</tbody>
</table>
## List of Tables

3.1 Preferences honoured from 700 total preferences when using the Kemeny-Young method. .................................................. 42
3.2 Kendall’s $\tau$ (with the associated p-values) when $N_{try}$ is 10 and 160. .... 42
3.3 Inter-annotator agreement for the random set of annotations. ................. 46
3.4 Inter-annotator agreement for the real annotations. .......................... 46
3.5 Inter-annotator correlation for the random set of annotations. ............... 47
3.6 Inter-annotator correlation for the real annotations. ........................ 47
4.1 Recall and precision for the clothing and body shape attributes. .......... 56
5.1 Parameters used in the setup when evaluating matching approaches of $BoV$ and LDM. ............................................................ 69
5.2 Image and configuration-based absolute accuracy with vocabulary of size 1K for inverted file. .............................................. 72
5.3 Image and configuration-based absolute accuracy with vocabulary of size 100K for inverted file. .......................................... 72
5.4 Image and configuration-based absolute accuracy with vocabulary of size 1M for inverted file. .............................................. 73
5.5 Image and configuration-based accuracy after assignment with vocabulary of size 100K for inverted file. ....................................... 74
5.6 Image and configuration-based accuracy after assignment with vocabulary of size 1M for inverted file. ....................................... 74
5.7 Image and configuration-based absolute accuracy with vocabulary of size 1K for min-Hash. .................................................... 75
5.8 Image and configuration-based absolute accuracy with vocabulary of size 100K for min-Hash. ................................................. 76
5.9 Image and configuration-based absolute accuracy with vocabulary of size 1M for min-Hash. .................................................... 76
5.10 Image and configuration-based accuracy after assignment with vocabulary of size 100K for min-Hash. .................................................. 78
5.11 Image and configuration-based accuracy after assignment with vocabulary of size 1M for min-Hash. .................................................. 78
5.12 Image and configuration-based accuracy for LDM. ............................... 79
5.13 Kendall’s $\tau$ and accuracy for LDM. .................................................. 80

6.1 Initial product potential for 3-node graphical model. ............................... 88
6.2 Final product potential for 3-node graphical model. ............................... 93
6.3 Kendall’s $\tau$ for representation model with $N_{try} = 10$ and 160. .......... 95
6.4 Kendall’s $\tau$ for representation model with larger values of $N_{try}$. ......... 96
6.5 Accuracy of representation model assuming node states are known. ............ 97
6.6 Accuracy of representation model for attribute recognition with 20%, 17% and 25% error and increasing % error. ................................. 98
6.7 Accuracy of representation model for attribute recognition with 10% error and increasing % error. ................................................. 99
6.8 Accuracy of representation model assuming node states are known. ............ 104
6.9 Accuracy of representation model for attribute recognition with 10% error and increasing % error. ................................................. 104

A.1 States of an image configuration. ...................................................... 117
Chapter 1

An introduction to ranking images based on aesthetic qualities

In this work the problem of comparing images based on qualitative assessment of image content and aesthetic impression is addressed. This is in contrast to ranking based on relevance to well defined image content. In this case the presence or absence of objects in the image is not ambiguous and a similarity measure can be established between images. It is not obvious on how to determine such a measure between the aesthetic impressions of the images. Given a collection of images, an ordering for these images is established based on their aesthetic qualities. Based on such an ordering it can be determined which image is ranked higher. A configuration of an image is considered which represents the states of the attributes within the image. A total number of $n$ configurations are ranked to obtain a ranking with $n$ entries as shown in Figure 1.1.

Figure 1.1: Problem of ranking $n$ image configurations based on aesthetics into an ordered list.
Such an approach can be applied to various computer vision areas, in this work the methods are employed for fashion interpretation. This is because this field of studying has received far less attention than is warranted. Studies related to the aesthetically oriented fashion and clothing industry are quite limited in the computer vision community, but have started to attract some interest in recent times. From the literature review, research focuses on either recognition or retrieval of clothing with some emphasis on fashion aesthetics [127, 79, 91]. According to figures published by the British Fashion Council, the British fashion industry is currently worth £26 billion [4]. In addition, the UK beauty industry is worth £17 billion to the country’s economy [2]. Such a significant financial potential alongside the challenging technical requirements in terms of computer vision motivate this work with the goal to promote relevant research in a domain which offers much scope for an interesting study.

This chapter highlights motivation behind the work and introduces the problems being addressed. The applicability of the work to the research community and some applications related to industry are also outlined. In addition, challenges associated with the problems are given and an overview of the thesis structure is provided.

Figure 1.2: Application of the approach to fashion interpretation where a large range of factors generally need to be considered. Some of these include different attributes, such as, clothing and body shape, various rules and additional factors like texture, colour and pattern. In this particular example, an apple body shape with recommended outfit is shown.

An example of ranking images according to certain criteria is illustrated in Figure 1.2. In various applications these criteria can be very complex. In fashion there are some general rules for determining the suitability of an outfit. These include a range of categories which are inherently complex. As an example, a fitted skirt can be worn with a specific top by a certain body type where the pattern, texture, fabric and colour aspects may also need to be considered.
to make reliable decisions. Due to the inter-related nature of the various attributes and rules it becomes difficult to annotate images for training an automatic approach. That is why a different approach that can capture the influence and relations between many components from a ranked list of images is adopted.

The goal of this work is to compare two or more images and rank them according to certain criteria based on their aesthetic qualities. As described previously due to a range of inter-related attributes and rules adopting an automatic approach becomes difficult here. Therefore a different approach that learns the attributes and their relations from a ranked list of images is utilized. The main contributions of the thesis are highlighted below:

- As a first contribution a novel dataset of 1064 images that are fully labelled with attributes of body shape, top and bottom clothing is introduced. And this dataset is annotated by aesthetic pairwise comparisons performed by an expert and annotators who have knowledge of fashion, i.e. mid-level annotators.

- A novel approach is proposed to obtain an objective ranking using visual assessments in form of a large number of pairwise comparisons. This is based on applying various consistency measures so that annotations can be combined into one ranking.

- An implementation based on bag of visual words approach which is utilized for recognition of clothing and body shapes is evaluated to determine prediction estimates.

- Nearest neighbour matching approaches of bag of visual words and local descriptor matching are evaluated to assess ranking of a pair of test images over a range of parameters. Based on this, an approach for global ranking is introduced to test the overall ordering of the configurations.

- Given the data which is in the form of an objective ranking an approach in which a graphical model is trained is proposed. Such a representation model captures all the attributes and relations between them.
1.1 Applications

To facilitate research in this work and the community a novel dataset has been collected. Images have been fully labelled with attributes of body shape, top and bottom clothing along with aesthetic pairwise assessments. This dataset puts an emphasis on visual aesthetics and coordinating attributes for specific configurations. Due to the presence of a number of controlled attributes, the data can be incorporated in a number of clothing related applications such as retrieval and recognition problems (cf. Figure 1.3(a)) [127, 79]. Besides applicability to the general scenarios it can also be utilized for performing visual assessment and annotation related studies [52, 78, 91]. Particularly in this instance the applications may benefit from the annotated pairwise comparisons performed.

The approaches proposed can be used within other fashion-related work [91] when performing comparisons. Besides, they can be employed in various other applications, such as assessing facial beauty [52] and for performing qualitative assessments [125, 131, 129]. An example could make use of the annotation methods proposed in Chapter 3. In addition, representations introduced can be utilized within retrieval and recognition systems [127, 79]. For example, image representation from the model in Chapter 6 could be used to address relevancy of the query images based on certain criteria in retrieval.

![Figure 1.3](image)

Figure 1.3: Some applications of the proposed approaches; (a) clothing recognition, (b) for fashion industry and fashion conscious people, (c) use within mobile phone applications and (d) fashion montage.

Automatically assessing paired-images also allows for a vast number of general applications.
One such area is its use as an aid for blind and partially sighted people to make fashion and style accessible. Although a huge amount of work still needs to be done in the area, work proposed can help bridge the gap between visual impairment and fashion. This is very relevant because most would agree that this can make one feel confident about themselves and help one’s desire for feeling attractive and achieving a unique style [1].

Some of the other applications include usage within the fashion industry, examples for which are demonstrated in Figures 1.3(b), 1.3(c) and 1.3(d). Consider the range of factors that influence a person when they go looking for clothing. Budget is always a strong factor but the psychological state of a person has a controlling influence. The latter is a combination of a sense of style which is based on self perception, trends and social peers. External influences involve advertisements and sales persons, these can be biased, motivated purely by sales. The level of consciousness when considering these factors will vary from person to person and day to day. Automatic assessment can be employed for personal use by a shopper during their decision making process when considering potential purchases. This can be beneficial to both the shopper and retailer. The shopper gets feedback based on appropriate selections given their personal criteria. While the return rates for the retailer could be reduced. Moreover, the approach could also be specialized to more specific scenarios such as to learn specific/individual preferences, or fashion rules in various cultural groups.

Given widespread use of smartphones, it can also serve as an on-the-go companion by inclusion as an app. In fact, one could even share appropriate dressing choices via social media and benefit from more advice and feedback. Furthermore, it can be utilized for fashion montages as simply a leisure time activity when experimenting with different looks. In this instance, one would receive guidance on an outfit without feeling a pressure of biased judgement from another person.

**1.2 Challenges**

There are a number of interesting applications for ranking images based on aesthetic qualities. However, a range of challenges need to be addressed so that an effective image representation can be obtained. A set of challenges stem from the problems associated with the attributes of
clothing and body shapes. A range of clothing and body shape categories need to be considered in order to allow a practical and varied choice of configurations. These configurations should also be reasonably universal in appeal. Then, attributes of clothing are required to be recognized (cf. Figure 1.4). The recognition system needs to distinguish between different clothing and account for variability due to shape and other factors such as colour, fabric, pattern, illumination etc. In addition, problems associated with shape proportions and body measurements need to be addressed for recognition of body shapes. This requires an appropriate dataset with multiple examples of the chosen attributes for recognition.

Figure 1.4: Challenges in ranking images based on aesthetic qualities associated with attributes; (a) recognition of different clothing attributes worn on upper half and lower half of the body with additional variation caused as a result of texture, pattern and colour and (b) recognition of the body shape attributes.

Figure 1.5: Challenges in ranking images based on aesthetic qualities due to the requirement of meeting a vast range of fashion rules, such as, if fitted trousers work well with hourglass shape, what blouse coordinates with fitted trousers and what colour of blouse coordinates with the trousers.
Another set of challenges occur due to the inter-related nature of clothing and body shape attributes and fashion rules as demonstrated in Figure 1.5. For a given number of body shapes $s$, top clothing $t$ and bottom clothing $b$, there are $s \times t \times b$ total configurations possible. The number of these configurations may increase dramatically when considering additional factors such as accessories, patterns, colour etc. In order to obtain a global ranking of images the more appropriate configurations need to be determined by capturing relations between the attributes and following fashion rules.

### 1.3 Objectives

The focus is to explore different objective rankings and image representation. An automatic assessment of sets of paired-images is performed using distinguishable configurations. This approach should demonstrate proven performance to discern between configurations as well as provide an optimal configuration based on an evaluation protocol. Then performance using the protocols is evaluated against baseline methods. The methods used must compare favourably or better than the baseline methods to indicate their effectiveness. Therefore, objective evaluation methods have also been proposed in this work. The objectives for various aspects of the work undertaken here are outlined below:

- Review the literature related to crowdsourced annotations, nearest neighbour search and visual assessment for judging aesthetic qualities.

- From the review conducted establish any relevant baseline methods, potentially suitable datasets and evaluation measures.

- If required generate a dataset with a number of images for every configuration of clothing and body shape attribute and annotate it with attribute labellings.

- Propose and implement an evaluation protocol to test the performance of approaches and to allow a comparative study.

- Design and implement a user interface for data annotation by multiple annotators for performing a large number of annotations and collect annotations based on aesthetics.

- Utilizing these implement a direct approach for automatic assessment which serves as a reference for evaluation.
• Investigate the application of optimisation techniques for performing similarity based searching on image configurations.

• Propose a representation that captures all the essential aspects in aesthetic assessment.

• Based on the investigation establish the most effective approach for automatic assessment.

1.4 Overview

After establishing the problem and applicability in the first chapter, the following chapter discusses the relevant literature. This covers various aspects of crowdsourcing, matching and aesthetics including the existing methods as well as state-of-the-art. An outline of the thesis for Chapters 3, 4 and 5 and 6 based on the main contributions made in them is shown in Figure 1.6.

The dataset together with the attributes alongside the crowdsourcing and ranking procedure are described in Chapter 3. Next, attribute recognition and an implementation of a baseline method for automatic assessment are evaluated in Chapter 4. Then, steps needed to achieve objective rankings based on the approach of matching are presented in Chapter 5. This chapter also provides an evaluation of the matching techniques over a range of parameters. Image representation from the learning approach based on global rankings is described in Chapter 6. The experimental results for the evaluations are presented and compared with the previously tested methods. Finally, conclusions from the work and associated future work is outlined, including improvements that can be made in some of the areas in Chapter 7.
Chapter 2

Related work

In this chapter work related to the problems addressed is given. Literature relevant to crowdsourcing is first presented in Section 2.2. Then, existing and recent developments for matching are discussed in Section 2.3. Finally, work in the context of image representations related to aesthetics is described in Section 2.4.

2.1 Introduction

To establish ranking of configurations based on aesthetic qualities of images, judgements from annotators are required. These are obtained using crowdsourcing and in order to address this related work is reviewed. The various methods in which images can be rated using annotators and problems arising as a result of this process are studied in Section 2.2.1 and Section 2.2.2. After an appropriate method for image rating has been determined, preferences obtained need to be ranked. Methods for this are investigated in Section 2.2.3 and applications that currently use crowdsourcing to achieve their goals are also studied in Section 2.2.4.

An image configuration can be matched to its nearest neighbour using a matching approach. A pair of test images with their associated matches can be compared with their scoring in a ranking obtained using crowdsourcing. To facilitate such a study, established approaches alongside more recent developments into nearest neighbour matching are reviewed in Section 2.3.1 and Section 2.3.2.
Assessing photo quality in the context of aesthetics has been gaining interest due to the growing number of digital images. As one of the prominent types of media in daily life its used within various applications. These include retrieving higher quality images, managing large collections of photos and photo albums. To establish a background knowledge of the field and determine any potential approaches, methods used in literature for quality assessment of images are reviewed in Section 2.4.1. In addition a more specific study into aesthetics is performed in Section 2.4.2 to establish suitable methods and evaluation measures for the purpose of this work. To conduct this work a dataset for studying aesthetics is required and for this an investigation into availability of existing datasets in literature is performed in Section 2.4.3.

### 2.2 Crowdsourcing

Using crowdsourcing is becoming increasingly popular in the field of machine vision. This introduces different challenges which need to be addressed therefore it is a highly active area of research. Crowdsourcing uses human intelligence and common sense for tasks that are relatively straightforward for humans, but difficult to complete using computers alone [56, 99]. Standard platforms such as Amazon Mechanical Turk, oDesk and Microsoft’s Universal Human Relevance System are becoming prominent. Requesters post tasks along with how much they are willing to pay for it. The workers can then choose the tasks they wish to complete. Although there may be a cost factor associated with crowdsourcing, it is generally hugely beneficial in many areas. For example, when there is a need for large scale annotated data.

The quality of annotated data is generally vital so there is a need to produce high quality results. This maybe quite challenging since some tasks are more appropriate for certain workers. Moreover, some workers are willing to work harder which would generally lead to better results. Due to these factors the areas of tackling annotation noise, matching tasks to workers, bonus payments for high quality results has gained considerable attention [95, 51, 56, 113]. Crowdsourcing tasks that have been completed to high standard offer many advantages. Not only can they provide annotated data, but provide useful visual assessments [91, 129]. These can be used to gain perspectives of individuals to construct objective measures by taking advantage of the underlying pattern and correlation.
2.2. Crowdsourcing

2.2.1 Image rating

There are several approaches to obtain ratings from annotators in this context. One of the most prominent one is that of absolute rating. A user is presented with a single image and is asked to assign a score usually between 1 and 10. The work for this tends to be mostly based on a method of absolute ratings given on a Likert scale [76]. This is a psychometric scale where the user is questioned about how much they agree with a statement. For such an approach multiple users are required to rate each image to obtain a distribution of ratings. True score can then be calculated using an average. This method may be prohibitive because different users from various backgrounds usually employ different criteria when rating images [52, 91]. Mean Opinion Score (MOS) is a Likert-style questionnaire [98]. A user is asked to provide rating from a discrete 1-5 scale which ranges from Bad to Excellent to grade the quality. Repeat tests are performed to obtain the overall rating by averaging scores.

Another approach for rating involves the use of sorting a collection of images which may produce reliable ratings [96]. A user is asked to sort images based on some criteria. It can be challenging for the user to sort a large dataset because all the images have to be considered together. A recent method involves presenting the user with a pair of images. They are asked which one is more appropriate (binary decision). Such pairwise comparisons reduce the level of challenge and ambiguity associated with the ratings. This is the chosen method to label facial beauty in [52]; assumption is that people in general have a consistent opinion on facial attractiveness. The absolute scores are obtained by minimizing a cost function. This penalizes images that are not in agreement with one of the pairwise preferences. A more recent method of $k$-wise comparison ($k = 10$) is proposed in [91]. The number of pairwise preferences attained from a $k$-wise rating is $\binom{k}{2}$. To perform each $k$-wise rating 10 random instances are shown to the users. They are required to rank each instance from most attractive to least attractive.

Crowdsourcing allows Quality of Experience (QoE) experiments to be conducted with ease because they do not have to be limited to academic laboratories. These experiments enable the estimation of how satisfied a user is in the subjective sense. For example, when browsing the web or shopping online. When QoE is based on subjective assessment, it can be based on absolute judgements or preference judgements. In [26] a framework for QoE using crowdsourcing for multimedia content is developed. Full sets of paired comparisons are utilized to
derive interval-scale scores by consistency checking, probabilistic choice modelling and web-based implementation. However, the problem of down sampling pairs was not considered. The authors in [109] describe crowdMOS which is a crowdsourced measure for testing subjective audio quality. The MOS test is performed for QoE assessment. To measure the uncertainty in subjective assessments they use a two-way random effects model and perform a post screening procedure. Consequently bonuses are awarded to workers to facilitate the framework. A method that combines preference and absolute judgements is proposed in [129]. Both kinds of judgements are combined directly through a unified probabilistic model to estimate the score of objects. In addition using a batch-mode active learning method they construct a set of queries for both the test sets to obtain cost-effective experiments.

### 2.2.2 Annotation noise

With the use of crowdsourcing large quantity of labels can be collected with ease, but these labels may be very noisy. The level of difference between expert annotations was investigated in [95]. They verified if the annotations differ much and whether they need to be repeated. They also studied whether non-expert annotations can be reliably utilized for ground-truth annotations in a benchmark scenario. An approach that uses pre-labelled gold standard dataset is proposed in [9]. In this method if an annotator generates contradicting labels from the standard dataset, then all the labels generated by that annotator may be rejected (unreliable). This approach can be advantageous because it can evaluate the annotators online. However, it requires a pre-labelled dataset that can serve as a gold standard which may be prohibitive. Another approach works by collecting multiple labels for each sample [120, 42]. These labels are later utilized through majority voting or model to generate the most likely label of the sample.

A new algorithm based on support vector machine (SVM) is proposed in [40]. This identifies low-quality and malicious annotators in order to decrease their influence on the supervised learning process. The assumption is that each annotator is either good or bad. Subsequently they introduced a method for pruning low-quality annotators in [41]. This uses a model trained from the complete labelled dataset from all the annotators as ground truth. Based on the assumption that good annotators behave in a similar manner, an approach for identifying good annotators by spectral clustering in the worker space is proposed in [27]. Matching of tasks
to workers from a mechanism design perspective is done in [51]. Using the bipartite graph between workers and task, they propose a uniform mechanism for the allocation. While in [56] a method to dynamically adjust payment for crowdsourcing tasks based on the quality of the work done is introduced. A multi-round principal-agent model is considered where a worker provides the effort level at each round. However, the requester is not directly aware of this.

An incremental relabelling mechanism is proposed in [134] which uses active learning to choose data for relabelling that has already been labelled. This is done for those labels where there is not enough confidence. An empirical Bayesian algorithm is introduced in [108] that eliminates annotators who assign labels randomly in an iterative manner. Such annotators make the task expensive along with reducing the quality of final labels. Another way to control the level of noise is to use crowd-teaching systems as proposed in [113]. This is based on a Bayesian model that interactively teaches the workers in an adaptive method. The problem of feature multi-selection is addressed in [110]. From amongst a large set of candidate features they determine the features that need to be addressed and the frequency of judging each one. They develop algorithms for regression with averaging of judgements. A method to improve the quality of ratings for unlabelled training sets with the use of re-rating is explored in [116]. This works by detecting instances and acquiring additional ratings for these after the rating process finishes.

Other approaches for modelling labellers’ quality have also been proposed in [124, 123]. They make use of labels that have been determined by the annotators to draw inferences using a Bayesian model. A parameter vector representing the annotator’s expertise, another one encoding response to the image and a linear classifier on parameter vector of each image are inferred. In contrast the Bayesian model proposed in [81] can also be used on unlabelled sample for better quality labels.

### 2.2.3 Ranking from preferences

A line of work related to crowdsourcing focuses principally on obtaining a ranking from preference aggregation models. For a consensus ranking multiple preferences over objects are combined. The preferences over objects can range from arbitrary comparisons, partial rankings to full rankings. Some popular areas where multiple preferences need to be combined
include information retrieval, collaborative filtering and social choice [121].

Most work in the area can be divided into two categories of permutation based and score based models. The permutation based methods are based in the permutation space and are computationally intensive. One of the most commonly explored methods is the Mallows model [85]. A distribution over the permutation is defined, the parameters for this include a central permutation and a dispersion parameter. The inference then involves finding this central permutation such that likelihood of the observed rankings is maximised. In most cases finding the central permutation is very difficult and intractable [84]. There are other generalizations of the Mallows model, one such is the CPS model [106]. In this items are drawn without replacement for a permutation. However, it still suffers from the drawback that possible permutations for all items need to be considered to find an optimal (central permutation).

The score based models overcome some of the limitations of the permutation based models by using real valued scores. The goals for these models is to learn a set of scores so there is one score per item to sort the items. One of the earlier score based method includes Borda Count [13]. It finds scores of the items by counting the number of points that are assigned to each one. A well known model for pairwise score is the Thurstone model [118]. This model is based on the assumption that an option’s quality is a Gaussian random variable. So for the basic case when there are two options, then let two Gaussian random variables represent the quality of these options. According to Thurstone’s model, when a person judges the better option from two options, they draw a realization from the quality distribution of both the variables. They then choose the higher quality option. In contrast the Bradley-Terry model uses Gumbel random variables instead of Gaussian variables [19]. There have been several extensions proposed of these models such as the Elo Chess rating system for rating chess players [48] and RankNet which is used in learning to rank [23].

A score based approach has been developed in [50] to model the joint pairwise matrix. In this the preferences expressed are first combined into a single preference matrix, the scores from the factorization of which are used to rank the items. A limitation of this method may arise due to combining all the preferences into a single matrix. As a result of which the individual information gets lost so outliers can influence the single preference matrix and affect the final scores. A flexible probabilistic score based model called the Multinomial Preference Model
2.2. Crowdsourcing

(MPM) over pairwise comparisons is developed in [121]. This expresses preference over objects using real valued scores. The convex optimization problem during inference from which can be solved using standard gradient techniques. An active learning method based on Thurstone model for pairwise rank aggregation is developed in [102]. This method works by asking the user to compare a pair of objects that are adaptively chosen at each iteration. This is done to make the process of performing paired comparisons cost effective as compared to random sampling. A limitation of this arises due to the model needing retraining on obtaining assessment for a new pair. This can add up the cost in a crowdsourced setting. Related work was performed in [28] where the active learning model is based on the Bradley-Terry Model for rank aggregation. This does not need to be retrained when an observation for a new pair is obtained because it employs an online Bayesian updating scheme. The associated inefficiencies arise because workers may want to make more than one judgement in each session. Generally in a crowdsourcing setting workers may work on multiple tests in a single session. To overcome this a batchmode active learning method is developed in [129].

2.2.4 Applications

The accurate annotation of video from keyframes by utilizing crowdsourced marketplaces is addressed in [122]. Labour pools are used to perform labelling of entities in the form of micro-tasks. Pixel-based features extracted from these are used for interpolation between keyframes to maintain superior performance at a given budget. In [37] an online apparel retailer is studied which uses both machine and human processing for recommendation. Machine processing is used for the structured data such as age, size etc. to generate the relevancy score. This score is passed onto a crowdsourcing stylist to incorporate unstructured information like images of merchandise or customer. To answers queries on data the use of crowd sourcing service (CrSS) alongside standard relational database operators is proposed in [99]. They use CrSS as a database where computations are performed by human processors and present hQuery which is a query model. A method for fine-grained recognition involving human-in-the-loop is introduced in [43]. This uses an online game to reveal important features which are then incorporated within an algorithm to learn classifiers for the categories. In a crowdsourcing scenario, [86] conduct experiments on Amazon Mechanical Turk to understand different voting rules. They make use of two voting problems involving sliding puzzles and picture of dots for
Chapter 2. Related work

2.3 Image matching

Similar images in the context of matching refer to images of the same object or scene taken under different conditions like lighting changes, viewpoint variations, occlusions etc. A naive solution would be to search the database exhaustively and sort the results using a similarity metric. While this approach guarantees a solution, it becomes prohibitive in real-world scenarios. Moreover, distance functions in the visual domain tend to require high-dimensional computations. In order to make matching practical, the domain of approximate nearest neighbour search (aNN) becomes important. This allows faster search times as a trade-off for a marginal loss in accuracy [11, 32, 34, 49, 71]. Many of the existing matching techniques are based on the Bag of Visual Words (BoV) model [31, 93, 104, 114].

2.3.1 Existing approaches

Various works address different stages of the image matching problem. Methods are addressed in four stages including feature matching, vocabulary construction, matching and post-processing.

Feature matching: Keypoints within an input image are detected and get assigned to feature descriptors which describe the content around the keypoints. Popular methods for key-point detection include Difference of Gaussian (DoG), maximally stable extremal regions (MSER) and multi-scale Hessian interest points [82, 88]. It is common to use a 128-dimensional Scale-invariant feature transform (SIFT) descriptor due to its proven superior performance [82, 89, 114]. This uses 4x4 array of 8 gradient orientation histograms around the key-points making it robust to geometric and photometric changes. Moreover, SIFT is also invariant to a small shift of few pixels within a region so it is more robust to localisation error. An image is typically represented by 500-5000 local descriptors [133]. Other notable SIFT-like methods include PCA-SIFT and the Gradient Location and Orientation Histogram (GLOH) descriptor [65, 89]. PCA-SIFT reduces the dimensionality of the commonly used 128 dimensional SIFT descriptor to 36 dimensions. However, this makes the descriptor less discriminative. Although, GLOH
is more distinctive than the SIFT descriptor, it is more expensive to compute. This is because it uses extended spatial regions when computing the histograms. An evaluation of various descriptors was carried out in [89]. It demonstrated the robustness of SIFT and SIFT-like descriptors compared to other descriptors. SURF which is related to SIFT was later proposed, where a speed-up of the order of 3-7 can be achieved over SIFT without consequential loss of quality [14].

Other proposed descriptors include BRIEF and BRISK which output a binary string [24, 74]. So matching can be performed using Hamming distance to allow faster computation. Another method for extracting a binary string using trace transform that is robust to common distortions is proposed in [21, 22]. Method works by pre-processing a scaled version of the original image using a Gaussian kernel. Estimates of two circus functions are made after the trace transform.

Work into image retrieval also includes methods using attribute features and Fisher vectors [44, 101]. In [44], three different image descriptors are used for retrieval; Fisher vectors, attribute features and text features. Fisher vectors, derived from the parameters of Gaussian mixture model (GMM) are computed using local descriptors (e.g. SIFT). They aggregate local descriptors to global descriptors. Attribute descriptors are constructed using image classifiers for every word in a vocabulary. A text descriptor in this context is built using image annotations. A histogram is generated for each image using the annotations. A vector of locally aggregated descriptors (VLAD) which is a method to efficiently aggregate local image descriptors is proposed in [60]. A compact representation which is derived from both BoV and Fisher kernel representation is produced.

**Vocabulary construction:** The descriptors are then vector quantised into clusters and get assigned to visual words. A visual vocabulary is constructed using k-means clustering or some other method [114]. Samples from the database or a random set can be used as training data for clustering. If we have a vocabulary $V$ encoded from $1, ..., |V|$, an image is then represented by a vector of size $|V|$. Each element of this vector will indicate the number of features represented by that visual word in the BoV approach. Visual words within this model are weighted using term frequency-inverse document frequency ($tf-idf$) [114]. The $tf$ part uses information related to the number of features assigned to a visual word. It represents the consistency within an image. The $idf$ part down-weights visual words that appear frequently within a database.
This is done to avoid bias caused due to commonly occurring words.

It has been shown that an extension of BoV model to high dimensions of the order of one million components performs well for large scale searches [59, 103]. The use of k-means algorithm now becomes prohibitive and as an alternative the approximate k-means (AKM) can be used for clustering [103]. Hard assignment involves the assignment of feature descriptors to a single cluster or visual word. This means that two features that may be close to each other in feature space will be considered totally different if they get assigned to different clusters. The actual separation between features may not be well represented, thus, resulting in problems with quantisation. Improvements to the quantisation error introduced due to hard assignment can be done using soft assignment [104]. Soft assignment based in the descriptor space can be carried out by assigning a descriptor to several visual words. Alternatively, Hamming embedding (HE) maybe used to reduce quantisation error [61]. Alongside the visual word for a descriptor, a binary signature is also used.

**Matching**: Indexing systems tend to be based on methods that require the use of the inverted list technique, hashing, or a tree-structure. In the inverted list a visual word from the vocabulary is followed by the image indices of images which contain that word [114]. Other notable indexing techniques include locality sensitive hashing and kd-trees [16, 49]. In a broad sense, hashing requires the use of generated hash functions to map data to codes [11, 49]. So similar data points have a high probability of collision or have the same hash code. Hash functions and codes are commonly referred to as projections and buckets. A bucket holds the points that are mapped to it. Methods can be broadly classified into two categories, data-independent and data-dependent. The former requires no use of prior information from the database to construct the hash functions. While the latter makes use of this information generally using a learning-based method. LSH, part of the first category, uses a vector representation and is unarguably one of the most popular hashing methods. A LSH family $H(c, cr, P_1, P_2)$ has two properties for $x, y \in \mathbb{R}^d$:

\[
\begin{align*}
\text{If } & ||x - y|| \leq r \text{ then } P_H[h(y) = h(x)] \geq P_1 \\
\text{If } & ||x - y|| \geq cr \text{ then } P_H[h(y) = h(x)] \leq P_2
\end{align*}
\]

(2.1) (2.2)

It should be noted that for $H$ to be useful $P_1 > P_2$ [11]. An example that satisfies the above two conditions is the Hamming space family. Hamming space is the set of all $2^N$ binary strings
of length $N$. Hamming distance between two binary strings of the same size is given by the number of bits at which they differ. A $k$-bit LSH function is defined as:

$$g(p) = [h_1(p), h_2(p), ..., h_k(p)]$$ (2.3)

Each $h_i$ is chosen randomly from the hash family $\mathcal{H}$ and results in a single bit value. A total of $l$ hash functions are chosen. Next $l$ hash tables are constructed using the $l$ hash functions such that $j$th hash table contains the points using $g_j$. Every training example is inserted into each of the $l$ tables. Querying involves iterating through all the hash tables to return the bin indexed by the query under consideration $g(q)$. A search through the returned bins can then be performed.

To ensure similar points get projected to the same code and achieve reasonable recall, a number of bits and tables are required [55].

Another method is the min-Hash which is a LSH for sets [34] where min-Hashes are grouped into $s$-tuples called sketches which must agree for retrieval. Geometric min-Hashing incorporates the spatial extent as well as geometric information of visual features to construct hash keys [32]. Therefore it allows superior recall and lower false positives. Repeatable hash keys are constructed using visual words and semi-local geometric information.

A limitation of kd-trees occurs due to the need to back track through the tree to find its nearest neighbour. Therefore, suboptimal methods that impose limitations on back tracking have been implemented to gain speed over certainty [15]. This is done by restricting distance from query point or set number of back tracking node transversal. However, similarity searching using kd-trees has shown to work well for SIFT descriptors and is still widely adopted [10, 82, 103].

**Post-processing:** After an estimate of candidate images is made for query, further steps can be carried out for refining the matches. Some of the commonly used methods include fast matching based on RANSAC and query expansion [33, 61, 103]. Typically, post processing methods make use of spatial information to reduce false alarms. This may be achieved by estimating an affine homography between a query region and target image. Spatial verification is usually applied to a limited number of top results. After each stage of verification, a decision is made whether the next result needs to be verified or not. In [33], a threshold of 20 images was used to determine when to stop the verification.

Spatially verified results can be used for query expansion. This can either be carried out by using a single new query or multiple queries (the ranked results need to be combined in this
instance). Query expansion is used for improving retrieval performance by re-issuing highly ranked results from the query as a new query [33]. Most query expansion methods have the same basis. A number of estimated candidate images are first spatially verified against query features. If the number of inliers found exceeds a certain threshold, the image is back-projected using the estimated transformation to the query region. Using this a new query is generated which is issued. Amongst the various approaches, average query expansion is commonly used [58, 100].

### 2.3.2 Recent developments

Three contributions are made in [12] to improve object retrieval. These include RootSIFT, novel query expansion method and improved image augmentation method. Usage of RootSIFT and discriminative query expansion (DQE) is recommended due to the additional computation cost being negligible. Improvement at the visual word weighting level is proposed in [135], where a novel IDF expression is introduced. With the use of $L_p$-norm pooling, they combine term frequency, document frequency, document length and codebook information in the IDF weighting scheme. In [119] existing techniques are combined in order to achieve the best possible retrieval quality. Based on BoV, HE and VLAD, a common model is proposed using an aggregation procedure with a selective match kernel.

The problem of finding approximate nearest neighbours in Hamming space is popularly solved by LSH for binary codes. In [94] they propose the first practical data-structure of its kind by building multiple hash tables for solving the exact K-NN in Hamming space. A novel method for fast computation of min-Hash signatures is proposed in [30]. It exploits the inverted file structure to obtain signatures. A need for fast min-Hash arises because a high performance of recall and precision requires the employment of a large number of sketches. A novel algorithm Complementary Projection Hashing (CPH) is proposed in [63] which is based on explicitly satisfying requirements for a good hashing method. The aim is to find hashing functions that cross the sparse region of the data. Alongside ensuring that the data points are evenly distributed in the hypercubes generated by these functions.
2.4 Visual assessment for aesthetics

Assessing image quality based on visual perspective has gained momentum in recent years [68, 83, 87, 92]. Web based image retrieval is starting to reach maturity where a user not only desires to retrieve images, but specify higher quality as a priority. Rank aggregation in recent research is often associated with content-based search systems [54, 107], with specific applications for web image searching [80, 77]. Although studies related to clothing are attracting interest, the related research in this area is still limited [112, 57, 78].

2.4.1 Quality assessment

Earlier works related to image quality assessment assume the presence of a reference image. The main objective is to evaluate the difference between a target image and reference image [38, 117]. However, the human vision system does not require a reference image to determine the visual quality of an image. So the domain of assessment of image quality without any reference becomes important [75]. This is called blind image quality assessment (BIQA) and in recent years has been attracting attention [90, 115, 130, 126].

The state-of-the-art methods for BIQA generally use learning to assess image quality from subjective human judgements. Training datasets are utilized by regression. They can be classified into two main categories of distortion specific methods [25, 111] and distortion independent methods [90, 115, 130, 126]. The first category is more useful when there is a specific kind of distortion. It works by estimating the quality of an image by quantifying the artefacts due to the distortion process. While the second category is more general purpose and therefore more challenging. There is no information of the distortion available which makes it the focus of various studies. A method that uses a SVM to determine the type of distortion is proposed in [90]. It uses support vector regression (SVR) model specific to the distortion for quality assessment. In [115] three regression models are trained for three sets of features. These include statistics of natural images, distortion textures and blur (noise). An estimate of the image quality is obtained after weighting the combination of the three. A summary of features and regression algorithms employed for recent BIQA methods is presented in [130]. Most methods described require a large amount of images to be scored by humans which maybe prohibitive. To address
Chapter 2. Related work

This a novel approach for BIQA without the use of human scored images for learning is introduced in [126]. A quality aware clustering algorithm is used to learn centroids which are used as codebooks. These infer the quality of an image patch to determine the quality of an entire image. A slightly different approach which involves the use of generic image descriptors is proposed in [87]. They use two generic descriptors including BoV and Fisher vectors to assess the aesthetics of images. The problem of evaluating the colour harmony of photos in the context of aesthetic quality classification is explored in [92]. An improvement in the classification is seen when their colour harmony feature is combined with the aesthetic based features (blur, edges and saliency features).

2.4.2 Aesthetics

Most work in literature tends to focus on identification of attractive facial characteristics. Methods for this can be considered as either being geometric or landmark feature based. An automatic facial beauty scoring system is developed in [6] based on 8 landmark ratios. A beauty assignment function is learned from a training dataset rated by annotators. The approach is tested on a dataset of 80 images on a scale of 1-4. An ensemble of features including landmark distances and ratios, an indicator of facial symmetry, skin smoothness, hair colour and the coefficients of an eigenface decomposition are assembled in [47]. Their method is evaluated on two datasets with 92 images in each. Later they improved their approach in [64] by making improvisations to the feature selection method. Most of these methods are limited due to small datasets that rely on a small subset of population (uniform ethnicity, age etc). Moreover, these methods are not fully automatic relying on manual localization of landmark features. Some of these limitations are addressed in [52]. A larger number of 2056 images are collected for a method that predicts female beauty. It uses an automatically learned appearance model without landmarks.

The problem of automatic makeup application is addressed in [53]. An example is used to transfer a certain style of makeup to a face. A study of female attractiveness based on modalities such as face, dressing and voice is performed in [91]. They study the impact of individual and joint modalities on human sense of beauty. Approaches to obtain image rankings and learn attribute based models are discussed.
2.4. Visual assessment for aesthetics

Attractiveness related to clothing is also studied, but is more focussed in sociology as in [73]. This investigates whether people perceive others differentially as a function of clothing attractiveness. Six different models in business attire were used as a stimuli. Three of them are wearing attractive clothing and the other three are wearing unattractive clothing. As expected models in attractive clothing were perceived better.

A method for parsing clothing in fashion photographs is presented in [127] where a novel dataset for garment items is also introduced. They present results on using information about clothing estimates to improve pose identification. Cross-scenario clothing retrieval is addressed in [79]. Using a human photo taken from the street they find similar clothing from online shops. Key components proposed in this include human/clothing parts alignment and an auxiliary daily photo dataset.

There are some existing works in literature that target the problem of clothing recommendation. One of the pioneering methods for giving product suggestions based on real-world user scenarios is the recommender in [112]. This is based on common sense reasoning to map the requirements of user with appropriate products (rather than vision based learning). Some works in particular target colour harmonization such as [29, 36]. In [36] a colour harmonization technique is presented. After quantifying the colour harmony of an image, it modifies the hue channel to shift the colours for a more harmonious feeling. Similar idea is explored in [29] in which a clothing colour combination system that helps users select appropriate colour combinations is presented. Pairing of clothing parts is addressed in [57, 78]. The graph model proposed in [57] is used such that given a cloth part it proposes another one. Similar idea is exploited in [78] but introduces occasion oriented components. It uses a latent SVM based model for matching rules, attributes and occasion components.

2.4.3 Datasets

There are several datasets for assessing facial beauty but very few that are related to fashion. In the facial dataset of [52] there are a total number of 2056 facial images which have been collected from a website HOTorNot. Some of the clothing related datasets relevant to the work being performed are now described.

A large clothing dataset was introduced in [128] which includes 937 people and 25441 cloth
instances. They utilize this dataset to address the problem of clothing recognition for surveillance quality videos. The data consists of 8 categories of manually labelled clothing such as shirt, jeans, short skirt, long skirt etc. Different shapes of clothing attributes are not considered, e.g., they use short/long skirt but not fitted/flare/straight skirts. Moreover, the images are collected from surveillance videos so the clothing worn by people is not very clear. This makes it unsuitable for studying aesthetic qualities in images.

A new dataset of 8035 images from PASCAL VOC 2010 and H3D using the person’s category was collected in [18]. Images in this dataset are centred at the full body of a person. The annotations for the attributes are provided by Amazon Mechanical Turk. A limitation of this dataset comes from the use of only 5 clothing related attributes which include t-shirt, shorts, jeans, long sleeves and long pants. Example images from this dataset are shown in Figure 2.1 (a).

In the data from [127] they collect a total of 158235 photographs from a website Chictopia.com which is a social platform for fashion conscious users. They also select a subset of 685 images for annotation of body parts and clothing attributes. However, this dataset does not address specific attribute configurations i.e. if a top is being worn with a specific trouser. Also they do not consider specific shapes of clothing, e.g., they address attributes like blouse and cardigan but not fitted top or flared trousers. Some examples of images drawn at random from this dataset are shown in Figure 2.1 (b).

A multimodal dataset including faces, dressing images and voices was presented in [91]; some dressing images are shown in Figure 2.1 (c). In this dataset You-tube videos of various talk shows, reality shows etc. are used to address Western and Eastern countries for comparison (from programs such as Chinese New Year Event, Eurovision, Britain Next Top Model, American Idol). They select one facial image, a full body one and a voice snippet for each female instance and collect a total of 620 instances (Westerners and Easterners). A person is sometimes holding a non-fashion related object (e.g. microphone) that obscures the attributes in this dataset. Furthermore, they do not address coordinating attributes of different clothing (upper and lower body).

An occasion-oriented clothing dataset that includes clothing annotation was introduced in [78] (cf. Figure 2.2 (a)). This dataset has been collected from shopping websites and photosharing
Figure 2.1: Some example images from the released datasets. (a) Attributes of people dataset from [18] used for recognizing attributes. (b) Fashionista dataset of [127] in which images are used to study clothing parsing. (c) Dressing images from M2B Dataset of [91] which includes additional modalities of face and voice to study human beauty; images from both the Western and Eastern sets are shown.

websites. It consists of a large number of 9469 full body images and further images of upper and lower body. They use Amazon Mechanical Turk for annotating the clothing and occasion attributes. Although this dataset addresses a large number of attributes, it does not consider coordinating the clothing attributes based on their shape. For example, jeans and shorts are considered but flared and fitted trousers/skirts are not. Moreover, it is not publicly available yet and no fashion experts were involved in the annotation.

A data to facilitate cross-scenario clothing retrieval based on online shopping and daily photos is collected in [79]. This dataset consists of a large number of images for upper and lower body. These images have been collected from shopping websites and Flickr.com; some examples of these are shown in Figure 2.2 (b). Images have been annotated with various clothing attributes using Amazon Mechanical Turk. However, the dataset has been incorporated for retrieval so no annotations based on aesthetic qualities have been performed. It is also not publicly available yet so the coordinating of attributes may not be addressed.

2.5 Summary

The area of crowdsourcing is an active area of research and work focuses on obtaining ratings, reducing annotation noise and generating rankings. Some of the more prominent methods
Figure 2.2: Some example images from the datasets that have not been released yet. (a) WoW dataset in [78] which contains images of the full body, upper body and lower body. It also includes annotations for occasion categories, examples for which are also shown. (b) Cross scenario retrieval dataset from [79] which includes online shopping (top) and daily photo datasets (bottom). Online shopping images contain less variance and occlusions than the daily photo images.

for rating include absolute rating, sorting a collection of images, pairwise comparisons and k-wise rating [76, 96, 52, 91]. As part of this work a subjective rating is required so that an objective ranking can be obtained. Providing an absolute ranking score within a certain range of values by an annotator is less reliable than comparing few images at a time [52, 91]. Pairwise preference score is efficient to carry out by an annotator and more reliable in terms of following the objective fashion rules. The annotator only has to indicate which of the two presented images is more aesthetically pleasing. Therefore, this is the chosen method for obtaining the ratings. From the review conducted pairwise comparisons have only been utilized in studying facial attractiveness [52].

Several approaches have been adopted in literature to reduce annotation noise. These include the use of pre-labelled gold standard dataset, collecting multiple data labels for majority voting and modelling the labeller quality [9, 120, 124, 81]. To validate annotations performed in this work, measures of agreement and correlation will be utilized [86, 95]. These can be computed from the annotated lists efficiently to check consistency. Ratings obtained from the annotators can be aggregated to obtain rankings using active learning model based on Thurstone
2.5. Summary

model or Bradley-Terry Model for pairwise rank aggregation [102, 28]. An intuitive method for obtaining a ranking would be to consider its long established use in preferential voting systems [66, 132]. The Kemeny-Young method which was developed for voting systems will be utilized in this work.

There is a considerable amount of work related to the problem of image matching. Besides the commonly employed SIFT features, studies into other features such as BRIEF, BRISK, Fisher vectors and VLAD have been performed [82, 24, 74, 44, 60]. During vocabulary construction soft assignment and HE can be utilized to reduce quantisation error [104, 61]. Matching usually utilizes data structures based on an inverted file, hashing or tree based [114, 16, 49]. For post processing spatially verified results are incorporated for query expansion [33, 58]. The more recent work focuses on combining the strength of existing methods to achieve the best possible results such as incorporating BoV, HE and VLAD [119]. Against this backdrop of extensive literature, BoV representation with SIFT features will be employed to investigate matching. For comparison an implementation that utilizes the full set of feature descriptors using Local Descriptor Matching will also be used [72].

For the study of attractiveness most work focuses on facial attractiveness and to a lesser extent on clothing [6, 52, 91]. Some work in literature also addresses clothing recommendation. This includes use of colour harmony, recommendation of cloth parts based on another one and occasion components [29, 57, 78]. It was established that the existing datasets from these relevant works were either not suitable or publicly unavailable [127, 91, 78, 79]. Therefore, a new relevant dataset will be collected to perform automatic assessment of paired-images based on aesthetic qualities. The images in this dataset will not include any restrictions in terms of the person’s facial beauty, their ethnicity or age.

Based on the literature review conducted, existing approaches utilized and novel approaches proposed in this work are highlighted below:

- The objective of this work is to form an image representation directly from a ranked list of images and to rate outfits to reflect recommendations of fashion experts. This goal forms an important novel contribution of the work.

- Pairwise comparisons will be utilized for rating images in this work. This method has so far been used for studying facial attractiveness in work related to image aesthetics [52].
• A novel dataset with images that are fully labelled with attributes of body shape, top and bottom clothing as well as aesthetic pairwise comparisons will be introduced as part of this work. Such a dataset is required due to limitations associated with the existing datasets or their public unavailability [127, 91, 78, 79].

• Large scale comparisons of the order of tens of thousands will be performed by mid-level annotators and a subset will also be annotated by an expert. Comparisons to this extent and by an expert have not been previously performed for such a scenario and are required to capture the user’s perspective of underlying aesthetics in images [52, 91, 78].

• An evaluation protocol and baseline will be proposed and implemented to test the approaches. Measures for this process will be based on existing measures of agreement and correlation in literature but will be incorporated to test the problems being considered [86, 95].

• A ranking will be obtained using Kemeny-Young method which originates from the domain of electoral voting [66, 132]. From the literature review, such a method has not been incorporated in establishing aesthetic scores within images.

• Utilizing the pairwise comparisons performed by annotators and expert, an approach to generate a ranking will be proposed.

• Existing nearest neighbour approaches based on BoV and Local Descriptor Matching over a range of parameters will be evaluated for the problem of scoring images based on aesthetics [114, 16, 49, 72]. An approach for creating a global ranking based on matches obtained will also been proposed.

• Based on an objective ranking, an approach which learns a graphical model and captures attributes and their relations will be proposed.
Chapter 3

The Dataset

This chapter presents the dataset that is utilized in this work. Images from this dataset are used to perform pairwise comparisons by annotators and an approach that obtains an objective ranking using these visual assessments is also presented. Firstly, the dataset, its structure and usage within the annotation procedure is presented in Section 3.1 and Section 3.2. Then, process of acquiring annotations with the design of tool, guidelines and image pairs is explained in Section 3.3. Next, measures for evaluating the annotations and criteria for their validation is described in Section 3.4. Then, method for obtaining objective rankings using the annotations is presented in Section 3.5. Finally, performance of the annotations and ranking approach is presented in Section 3.6.

3.1 Dataset

In this section the dataset with the relevant attributes and labellings is presented. It describes the process of achieving data along with the attributes and how these are coordinated to form complete images. It also describes manner in which the control image pairs are included within the annotations for validations.
Figure 3.1: Clothing and body shape attributes in the dataset. There are 11 clothing and 4 body shape categories. An image represents a configuration where a person with a certain body shape wears specific top and bottom clothing.

3.1.1 Dataset and labelling

From the literature review it was found that there are some existing datasets related to clothing and fashion, but none of them are suitable (cf. Chapter 2 Section 2.4.3) [127, 91, 78, 79]. Therefore, a new dataset is collected across the Internet from different clothing retailers, including high-end, high-street, budget and dedicated online shopping retailers\(^1\).

Various categories with a total number of 15 that include 11 categories for clothing and 4 for body shape attributes are used and shown in Figure 3.1. For the top clothing attributes, 5 categories including top: fitted, loose, ruffled and jacket: fitted and loose are used. The 6 bottom clothing attributes include trousers: flared, fitted and straight and skirt: flared, fitted and straight. The body shape attributes are drawn from the common categories of apple, column, hourglass and pear. There are a total of 120 configurations constituted from the 5, 6 and 4 categories of top clothing, bottom clothing and body shape attributes respectively.

It is not straightforward to find examples of the same clothing configurations and different body shapes. Image examples are therefore collected for all possible configurations of clothing attributes and images are then carefully warped to reflect each type of body shape according to body proportions given on fashion websites. Points in the waist-hip region are manually selected with reference to a specific body shape to obtain realistic images. Some examples from the dataset that collectively consists of 1064 images labelled with states for 120 configurations can be seen in Figure 3.2. The face-related bias is eliminated by excluding this part from the images.

\(^1\)The data is available from http://kahlan.eps.surrey.ac.uk/featurespace/fashion and further details for this can be found in Appendix A.
3.2. Aesthetic comparisons

A large number of pairwise comparisons utilizing images from the dataset introduced in Section 3.1.1 are required for this work. Image pairs are presented to a group of annotators who label these in accordance with their judgement of what they regard as being aesthetically pleasing. An image represents a specific state for the attributes of body shape, top clothing and bottom clothing which were labelled as part of the dataset collection process in Section 3.1.1. For the comparison an annotator must choose the image they find more aesthetically pleasing from an image pair. These perspectives are utilized to extract underlying pattern and correlation from which an objective score for an image is constructed. Further details of how the annotators are selected for performing this task and details of the annotation task are given in Section 3.3.1 and Section 3.3.2.

Annotation noise: The quality of annotated data is crucial to research undertaken here, so certain measures need to be adopted to address the issues of annotation noise. There are several factors that may result in annotation noise, one being that certain tasks are more appropriate for
some workers compared to others. Given that worker is adept at a specific task and performs consistently over time, then the expected level of noise should be low. To determine eligibility of applicant annotators, a specific selection criteria is required and discussed in Section 3.3.1. To minimize the impact on consistency over time, annotators were allowed to complete the tasks at their leisure. Additionally, they were allowed to revisit and alter their decisions if they so choose. This is to reduce the overall pressure associated with performing the task. The next issue affecting quality requires addressing the mixed abilities of annotators. For this a control set of image pairs are utilized across all annotators and described in detail in Section 3.2.1. Outliers are filtered by an expert annotator who determines when an annotator deviates significantly from the fashion consensus of experts. Criteria for exclusion on this basis is discussed in Section 3.4.2.

### 3.2.1 Control sets

In order to measure the annotation consistency between different annotators several control sets of image pairs are introduced. Specifically, two subsets within a large list of image pairs presented to each annotator are used. One is annotated by an expert and the other one is repeated in all lists for annotators. In the study conducted in [95] it was found that expert annotators showed a high consistency in annotations measured on agreement and correlation. In the experiments conducted as part of this work one expert is utilized for annotating the reference control set which makes the process more cost-effective and feasible. It is important to have an expert so that consistency checks can be performed and stronger and weaker group of annotators can be established.

![Figure 3.3: This illustrates the inclusion of control sets within 10 annotation lists. An annotation list $L_a$ is formed of three sets $L_{e.a}$, $L_{r.a}$ and $L_{u,a}$ and the control image pairs are sampled at regular intervals within a list.](image-url)

Figure 3.3: This illustrates the inclusion of control sets within 10 annotation lists. An annotation list $L_a$ is formed of three sets $L_{e.a}$, $L_{r.a}$ and $L_{u,a}$ and the control image pairs are sampled at regular intervals within a list.
A list of images $L_a$ presented to annotator $a$ consists of $L_a = L_{e,a} + L_{r,a} + L_{u,a}$ as shown in Figure 3.3. $L_{e,a}$ is the set where image pairs are annotated by all $A$ annotators including the expert $e$. List $L_{r,a}$ is used for the inter-annotator analysis and includes additional image pairs that are repeated for each annotator $a$. $L_{u,a}$ are image pairs that are unique to every annotator $a$. The control image pairs $L_{e,a}$ and $L_{r,a}$ are sampled at regular intervals within an annotation list and an annotator is not aware of where these have been positioned. The image pairs for the annotators can be drawn from the complete comparative dataset. They can also include a more practical subset of all possible image pairs depending on the application. The size of $L_{e,a}$, $L_{r,a}$ and $L_{u,a}$ will then be set according to the number of image pairs that are annotated.

3.3 Crowdsourcing for qualitative assessments

There are $N = 1064$ images in the dataset which give $(N^2 - N)/2 = 565516$ unique image pairs. Based on various constraints an annotator was asked to score $|L_a| = 7000$ pairs which corresponds to 7h of work. $A = 10$ annotators who follow fashion trends were recruited which allowed a total of 70000 image pairs to be scored. The size of the expert and repeated control pairs was set to $0.1 \cdot |L_a|$ that is $|L_{e,a}| = 700$ and $|L_{r,a}| = 700$. With the remaining $|L_{u,a}| = 5600$ of unique pairs for every annotator a total of 57400 unique image pairs were scored for the whole dataset. This is a small subset of all possible 565516 pairs therefore the subsets have to be carefully selected to assure that each image occurs a comparable number of times in all image pairs and the distribution is balanced amongst the annotators. Given the upper triangle of $N \times N$ matrix of all possible pairs of images, pairs from the longest diagonals of the triangle were selected.

In the context of this work, task of annotating 70000 image pairs by $A = 10$ annotators and 700 image pairs by an expert is defined as crowdsourcing. These 10 annotators are mid-level annotators since they possess knowledge of fashion and its principles. Specifics of the criteria used in the selection of annotators and the annotation task are described next.

3.3.1 Selection criteria

This work requires a consensus of fashion enthusiasts to analyse aesthetics based on general fashion and so it is possible to establish a general criteria. It does not target obtaining inputs from a large number of fashion professionals working in industry, i.e., experts due to various
constraints with cost being the primary reason. The criteria for choosing annotators for performing the crowdsourcing task described in Section 3.3 included “people who follow western fashion with a keen interest”. Annotators were selected based on details provided in their CV, in particular they had to show work experience, hobbies or an active interest in the field of fashion. The advertisement that was used to advertise the annotation task to interested participants is shown in Figure 3.4. People who could not demonstrate an interest or knowledge or experience in fashion were filtered out. This procedure of recruitment was followed because it is a standard process to select participants for a task. As a limitation it may exclude some competent individuals but it would mostly include people with a fashion know-how.

Fashion Apparel Assessor
Location:
University of Surrey
Role Information:
The task involves comparing fashion-based paired images to decide which is more aesthetically pleasing.
Criteria used in the decision process should include:
1. How does the overall clothing look?
2. How well does the person suit the clothing that they are wearing?
Comparisons will be made using an interface in Matlab, which can be accessed freely across campus. The interface allows the user to complete the comparisons at their leisure with options to go back and change results if they so choose.
There are a total of 7000 images pairs for each participant. A total number of 10 annotators are required who will each finish a set of experiments out of the total number of 10 experiments. Normally a fair unbiased judgement of an image for similar testing takes a few seconds, therefore, it is expected that approximately 1000 image comparisons can be performed in an hour. The annotators will need to spend a total of 7 hours in order to compare 7000 images.
Additional Information:
Participants who follow western fashion with a keen interest are desirable.
The main task requires access to Matlab which is freely available on campus. However, the interface is self-explanatory and does not need specific knowledge of the software. There will also be personal help available if the participant desires any sort of assistance with using Matlab.
To Apply:
Please ensure that you have your CV uploaded to your account and your contact details are up to date prior to submitting an application. Please ensure you submit a covering statement along with your application.

Figure 3.4: Selection criteria and information provided to the participants for performing the crowdsourcing task.
3.3.2 Annotation task

The expert as well as the annotators were provided with an assessment tool (see Figure 3.5) to obtain binary scores. Images in grayscale were presented to each annotator and they were asked to select the one that created a better overall aesthetic impression from each pair. Images used in performing these pairwise comparisons originate from the dataset introduced in Section 3.1.1. An image presented to an annotator is a configuration labelled with a state for clothing and body shape. These labellings were performed at the time of collecting and post-processing images in the dataset. The instructions provided to the annotators included how well did the overall clothing look and how well did the clothing suit the body shape.

Figure 3.5: The tool that was used to perform annotations by both mid-level annotators and expert. An annotator clicks on the preferred image from the image-pair in this user-friendly tool to perform 7000 comparisons.

3.4 Evaluation of annotations

In this section evaluation measures used to assess agreement and correlation of the annotations are first described. And then the criteria used to validate the annotations is explained.
3.4.1 Agreement amongst annotators

An intuitive method to test the consistency amongst the annotators can be based on the agreement between the annotated lists. First, agreement between the expert list $L_{e,e}$ and annotated list $L_{e,a}$ for all $A$ annotators (expert-annotator) is computed. Then the inter-annotator agreement between all pairs of lists $L_{r,a}$ is measured.

**Accuracy** [20] between two sets of annotated image pairs $L_{e,a}$ and $L_{e,e}$ is given by:

$$\text{acc}(L_{e,a}, L_{e,e}) = \frac{|\{\forall i, L_{e,a}\{i\} = L_{e,e}\{i\}\}|}{|L_e|} \quad (3.1)$$

where $L_{e,a}\{i\}$ is the outcome of the comparison for image-pair $i$ in list $L_{e,a}$, and $|\cdot|$ is the size of a set. The $\text{acc}$ value of 0.5 corresponds to a random score and 1 to identical lists.

**Cohen’s $\kappa$** [35] coefficient is a statistical measure that is used to evaluate the agreement amongst annotators [95]. It is used to compare agreement between two annotators who each classify assignments into mutually exclusive categories. This measure is considered to be more robust compared to accuracy since it also considers the agreement that occurs as a result of chance (shown in Equation 3.2).

$$\kappa(L_{e,a}, L_{e,e}) = \frac{P(A) - P(C)}{1 - P(C)} \quad (3.2)$$

where $P(A)$ is the proportion of times the annotators $L_{e,a}$ and $L_{e,e}$ agree and is equal to $\text{acc}(L_{e,a}, L_{e,e})$ and $P(C)$ is the proportion of times the annotators $L_{e,a}$ and $L_{e,e}$ are expected to agree by chance only. The values for $\kappa$ range between $[-1, 1]$ for perfect disagreement to a perfect agreement with 0 implying that relationship between the annotators is due to chance alone (random).

**Kendall’s $\tau$** [67] coefficient is used to assess the correlation between ranked lists [86]. It is based on a non-parametric statistic to measure the level of association between two rankings and to test the corresponding significance of this association. Note that a ranking has to be generated from pairwise comparisons in order to use Kendall’s $\tau$ measure. If there are two samples $x$ and $y$ each of size $N$ then the pairs can be ordered by both $x$ and $y$ values. The total number of unique pairings that can be formed of $x$ with $y$ is $(N^2 - N)/2$. If $x_1 > y_1$ when ordered based on both $x$ and $y$ the first pair is concordant else it is discordant. Kendall’s $\tau$ can then be given by:

$$\tau = \frac{n_c - n_d}{(N^2 - N)/2} \quad (3.3)$$
where \( n_c \) are the concordant pairs from observation and \( n_d \) are the discordant pairs. The correlation score varies between 1 for identical rankings, 0 for independent ones to \(-1\) for perfect inverse of two rankings.

In addition to the strength of correlation between two rankings being considered, Kendall’s \( \tau \) also provides a distribution free test of uncorrelation. For testing the null hypothesis, p-value needs to be calculated. In this context, null hypothesis represents the hypothesis of no correlation or mutual independence between two rankings being evaluated. The standard value where the null hypothesis is rejected at the significance level of 0.05 for two-tailed test is employed. This means that the possibility of both positive or negative correlation is considered. So if the p-value is less than 0.05 then there is statistically significant lack of independence, i.e., correlation score is significantly different from zero between the rankings, e.g., that generated from an annotator and the expert.

### 3.4.2 Annotator validation

Based on the agreement scores from all the three measures the pairwise comparisons performed by an annotator are validated. If the values for an annotator \( a \) are above a threshold, then the annotated list \( L_a \) is considered valid. The expert \( L_{e,a} \) and the repeated \( L_{r,a} \) lists are used to evaluate annotation agreements. The minimum required agreement with the expert annotator \( L_{e,a} \) is set higher than for the inter-annotator \( L_{r,a} \). Based on this validation a group of strong and weak annotators can be identified.

### 3.5 Rankings based on pairwise comparisons

To generate rankings from \( L_{r,a} \), \( L_{e,a} \) and \( L_{u,a} \) all the pairwise decisions are ordered into overall preference orders, this problem falls in the domain of rank aggregation.

**Rank aggregation**: It is the problem of determining an ordering \( \pi \) of a ground set \( S \) with a total number of \( n \) elements and combines information from the list \( \pi_1, \pi_2, \ldots, \pi_k \) which is formed from multiple input rankings (votes) [39]. One of the main limitation of generating a full ranking is that obtaining full ranking information of \( S \) from all the voters is potentially too much of a task. However, it finds usage in multiple applications such as meta-searching,
ranking players in sports and competitions, collaborative filtering, e-commerce and many more. But most extensively this problem has been studied in the domain of electoral voting where a natural step for rank aggregation was taken by Kemeny. A Kemeny optimal $SK$ is given by [46, 45]:

$$SK(\pi, \pi_1, \pi_2, \ldots, \pi_k) = \sum_{i=1}^{k} K(\pi, \pi_i)$$ (3.4)

So if there are multiple partial lists $\pi_1, \pi_2, \ldots, \pi_k$ and a full list $\pi$, then a permutation $\sigma$ is a Kemeny optimal aggregation for the partial lists $\pi_1, \pi_2, \ldots, \pi_k$ if it minimizes $SK(\pi, \pi_1, \pi_2, \ldots, \pi_k)$ over all permutations $\pi$. In this case $K$ is the Kendall $\tau$ distance (or bubble-sort distance) and so informally a Kemeny optimal ordering minimizes the sum of the bubble sort distances. Therefore, Kemeny optimal solutions result in the best compromise orderings.

3.5.1 Kemeny-Young method

The Kemeny-Young method combines ordered preference lists from multiple sources into one overall preference order and is utilized for obtaining rankings in this work [66, 132, 105]. As a voting algorithm it not only computes the winner but also determines an entire ranked list. This Condorcet method will order $M$ candidates such that the winners will be ranked at top $N$. In a Condorcet method the candidate with the first ranking would win against the remaining candidates in individual pairwise comparisons given that a candidate with such a property exists. And this will continue such that the candidate with the second ranking would win against all the other candidates except the first ranked candidate. Mathematically it minimizes the disagreements amongst the voters in their pairwise preferences between all the candidates. This means that if a certain candidate is ranked ahead of another candidate in a Kemeny-Young ordering, then a voter who made a contradictory preference for these candidates will contribute +1 to the value of the total discrepancy. So it does not use any information about how much higher a candidate is ranked compared to another by a voter but only uses relative ordering.

The Kemeny-Young ordering may not be unique because more than one ordering can be tied for the minimum number of discrepancies. It is also known that this ranking problem is NP hard even for a single minimum discrepancy ordering. Thus real-world usage of this approach can only be heuristic but fast heuristics that perform well for practical problems exist and one such
implementation is employed [105]. It consists of \( N_{\text{try}} \) independent greedy minimizations of the Kemeny-Young score for which mean rank is chosen as the starting value to record resulting ranks for each candidate. For the application being considered, voters correspond to the pairwise comparisons being performed and candidates are the independent images that form the image pairs. The quality of the ranking output by the Kemeny-Young method is indicated by the number of pairwise decisions that were honoured in the generated ranking.

**Algorithm:** For a single greedy minimization some steps are executed [105]. First a random permutation is drawn of the candidates. This permutation order is followed to re-arrange the candidates in the order position so that the Kemeny-Young score is minimized. In the instance of a tie, a random position is chosen. Following the movement of all the candidates, starting at the top and moving down, a window of a certain number of candidates is considered. All permutations for the candidates within the window are tried to choose the permutation that minimizes the Kemeny-Young score. If the steps do not decrease the Kemeny-Young score then the indefinite greedy minimization loop is exited. Since each minimization is partly stochastic, results can be different on each try.

### 3.5.2 Rankings from annotators

Several ranked lists can be produced from the annotated image pairs using the Kemeny-Young method.

- A ranking of images within each annotator list \( L_a \) combined from the three subsets \( L_e,a \), \( L_r,a \) and \( L_u,a \).

- A ranking of all images combined from several or all annotators.

For the image pairs from the repeated list \( L_{r,a} \) that are scored several times by different annotators the more frequent score is selected.

Note that each image is an example of a configuration of clothing and body shape therefore each ranking of images can be compressed into a ranking of 120 configurations by averaging the rank of images that belong to the same configuration (cf. Section 3.1.1). A full ranking, i.e., ordering over image space for all the images in the dataset will consist of 1064 elements.
and a compressed ranking over attribute space will have 120 configurations. These are both utilized when performing evaluations in this work.

### 3.6 Experimental results

In this section performance of the approach is evaluated. The criteria used for measures is first determined in Section 3.6.1. An analysis of the performance of Kemeny-Young method to generate rankings is then presented in Section 3.6.2. Next, agreement and correlation between rankings generated from the expert and other annotators is analysed in Section 3.6.3. The inter-annotator analyses is then performed in Section 3.6.4 by comparing rankings from the annotators. Based on these evaluations rankings obtained from various group of annotators are discussed in Section 3.6.5.

**Data:** A total of 57400 image pairs from 1064 images in the dataset are included in the crowdsourcing experiment. The pairs are formed such that each image is included in at least one pair. A set of pairs for each annotator include \( |L_e| = 700 \) image pairs that are also annotated by the expert, \( |L_r| = 700 \) image pairs that are repeated for all annotators and \( |L_u| = 5600 \) unique pairs. To establish a baseline an experiment on a simulated dataset is also performed. This dataset is obtained for 57400 image pairs by randomly generating the comparison score for the 10 annotation sets.

#### 3.6.1 Criteria for agreement

In this section the relations between accuracy, Cohen’s \( \kappa \) and Kendall’s \( \tau \) are analysed as well as the thresholds that will be used to validate the annotators are determined. This is performed using two sets of 700 image pairs with simulated binary scores such that agreement between the two sets increases from 0% to 100%. Agreement results based on accuracy and Cohen’s \( \kappa \) can be calculated from the binary scores in contrast to Kendall’s \( \tau \) which compares full rankings of images generated with Kemeny-Young method.

Figure 3.6 displays accuracy, Kendall’s \( \tau \) and Cohen’s \( \kappa \) when increasing agreement between two sets. The results for Cohen’s \( \kappa \) are closely correlated with those for accuracy. There is a non linear increase of Kendall’s \( \tau \) at the end of the agreement range. This is due to the fact
that a full ranking cannot honour all randomly generated binary scores as some are conflicting, which becomes apparent from Kendall’s $\tau$ values at high levels of agreements. It should be noted that the p-value associated with Kendall’s $\tau$ have a value of 0 for the range of agreement except when the agreement is 50%. A value of 0.56 is found for this instance. A positive agreement and correlation in which the agreement exceeds 50% so that the correlation score is significantly different from zero is required to meet the criteria. Arbitrarily chosen agreement of 65% with the expert and 60% with the repeated sets provides a reasonable threshold to identify strong and weak annotators.

![Figure 3.6: This shows accuracy, Cohen’s $\kappa$ and Kendall’s $\tau$ when comparing decision results from the two sets of 700 image pairs. For the two sets, overlap in output decisions is increased from 0% to 100%.](image)

### 3.6.2 Ranking using Kemeny-Young method

The effect of increasing $N_{try}$ on the number of preferences honoured for the pairwise comparisons performed by the annotators is evaluated in this section. This includes an evaluation of the improvement in performance obtained by increasing $N_{try}$. A general investigation into Kendall’s $\tau$ and its associated p-values is also done to facilitate experiments in the following sections. The evaluation set which has also been annotated by the expert is used for this study. Ranking is performed by Kemeny-Young method over a range of $N_{try}$ of 10, 20, 40, 80, 160 for the 711 images from the 700 image pairs. Ranked lists are obtained for image pairs annotated by annotators and image pairs for which the scores are randomly generated.

**Preferences honoured:** The results for the preferences honoured for $N_{try} = 10, 160$ are shown in Table 3.1. It is observed that the number of preferences honoured is higher for the real
Chapter 3. The Dataset

annotations when compared with the random annotations. This is because the random nature of assessments results in disagreement in the comparisons. On increasing $N_{try}$ to 160 there is a slight increase in the number of preferences honoured for some annotators. It should be noted that the number of preferences honoured may decrease generally with more comparisons. This is because it might not be possible to satisfy all comparisons due to disagreements from different annotators or even from the same annotator.

<table>
<thead>
<tr>
<th>Set</th>
<th>$N_{try}$</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>7</th>
<th>8</th>
<th>9</th>
<th>10</th>
</tr>
</thead>
<tbody>
<tr>
<td>Real</td>
<td>10</td>
<td>700</td>
<td>698</td>
<td>694</td>
<td>700</td>
<td>698</td>
<td>700</td>
<td>698</td>
<td>700</td>
<td>694</td>
<td>698</td>
</tr>
<tr>
<td></td>
<td>160</td>
<td>700</td>
<td>700</td>
<td>698</td>
<td>700</td>
<td>700</td>
<td>700</td>
<td>700</td>
<td>700</td>
<td>696</td>
<td>700</td>
</tr>
<tr>
<td>Rand</td>
<td>10</td>
<td>690</td>
<td>682</td>
<td>684</td>
<td>692</td>
<td>684</td>
<td>694</td>
<td>682</td>
<td>686</td>
<td>690</td>
<td>694</td>
</tr>
<tr>
<td></td>
<td>160</td>
<td>696</td>
<td>690</td>
<td>688</td>
<td>692</td>
<td>698</td>
<td>690</td>
<td>696</td>
<td>690</td>
<td>688</td>
<td>694</td>
</tr>
</tbody>
</table>

Table 3.1: Preferences honoured from 700 total preferences when using the Kemeny-Young method with $N_{try}$ of 10 and 160 for the real and random annotations.

Kendall’s $\tau$: Values from the rankings between expert and annotators are given in Table 3.2. It is seen that the random annotations have a value close to zero because they are not correlated with the expert annotations. Moreover, high p-values mostly greater than 0.05 are found which indicates that the correlation found between random and expert rankings is not significantly different from zero. However, this is not the case for real annotations since Kendall’s $\tau$ is greater than zero with zero-valued p-values. In particular, a high $\tau$ is seen for annotators 1, 2, 4, 6, 10.

<table>
<thead>
<tr>
<th>Set</th>
<th>$N_{try}$</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>7</th>
<th>8</th>
<th>9</th>
<th>10</th>
</tr>
</thead>
<tbody>
<tr>
<td>Real</td>
<td>10</td>
<td>0.31</td>
<td>0.29</td>
<td>0.12</td>
<td>0.23</td>
<td>0.19</td>
<td>0.26</td>
<td>0.14</td>
<td>0.08</td>
<td>0.19</td>
<td>0.24</td>
</tr>
<tr>
<td></td>
<td>160</td>
<td>0.30</td>
<td>0.27</td>
<td>0.11</td>
<td>0.21</td>
<td>0.19</td>
<td>0.25</td>
<td>0.15</td>
<td>0.09</td>
<td>0.19</td>
<td>0.24</td>
</tr>
<tr>
<td>Rand</td>
<td>10</td>
<td>0.00</td>
<td>0.02</td>
<td>-0.01</td>
<td>0.00</td>
<td>-0.03</td>
<td>-0.01</td>
<td>-0.06</td>
<td>-0.02</td>
<td>-0.01</td>
<td>0.06</td>
</tr>
<tr>
<td></td>
<td>160</td>
<td>(0.00)</td>
<td>(0.48)</td>
<td>(0.60)</td>
<td>(0.94)</td>
<td>(0.28)</td>
<td>(0.71)</td>
<td>(0.02)</td>
<td>(0.51)</td>
<td>(0.62)</td>
<td>(0.02)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>(0.37)</td>
<td>(0.19)</td>
<td>(0.34)</td>
<td>(0.47)</td>
<td>(0.65)</td>
<td>(0.66)</td>
<td>(0.01)</td>
<td>(0.80)</td>
<td>(0.43)</td>
<td>(0.02)</td>
</tr>
</tbody>
</table>

Table 3.2: Kendall’s $\tau$ (with the associated p-values) when $N_{try}$ of 10 and 160 are used for the real and random annotations; all real annotations have p-value = 0.

Marginal differences are observed between the correlation scores when using different values of $N_{try}$. For this reason and to ensure consistency, ranking in the following experiments which
3.6. Experimental results

Involving validating annotations is performed with $N_{try} = 10$.

An interesting observation is made when the preferences honoured remain the same for $N_{try} = 10$ and $N_{try} = 160$, although very slight differences are seen for Kendall’s $\tau$. As an example, consider annotator 4 where Kendall’s $\tau$ shows marginal reduction from 0.23 to 0.21 even though all the 700 preferences are honoured. This is most likely due to the procedure followed to achieve an independent greedy minimization explained in Section 3.5 [105]. As it was described each minimization is partly stochastic so the results can be different on each try resulting in slightly different Kendall’s $\tau$.

3.6.3 Annotator vs. expert agreement

This section investigates the quality of annotations by measuring agreement between annotators and expert. The performance of Kemeny-Young by comparing the agreements between full rankings generated with Kemeny-Young to the agreements of binary scores that are input to Kemeny-Young is also measured. The results for random and real annotations for the three measures are presented in Figure 3.7. Average accuracy, Cohen’s $\kappa$ and Kendall’s $\tau$ when using random annotations are 0.50, 0 and $-0.01$. This is as expected because these annotations are not correlated with those of the expert. Furthermore, a high p-value of 0.52 is seen for Kendall’s $\tau$ as was found previously in Section 3.6.2.

It can be seen that higher values are found for the real instances. In addition, a zero-valued p-value for Kendall’s $\tau$ was found for all the annotations. The highest agreement is observed for annotator 1 with accuracy and Cohen’s $\kappa$ of 0.73 and 0.45. Based on accuracy, the pairwise comparisons that are in agreement can be determined which equals 511 out of the 700 performed. Similar measurements can be derived from Cohen’s $\kappa$ which represents an agreement of 70%–80% in the comparisons performed. On the contrary the lowest agreement is seen for annotator 8 with accuracy and Cohen’s $\kappa$ of 0.56 and 0.12. This corresponds to a low agreement of 50%–60% which means that this annotation may not be much better than that performed randomly. Besides annotator 1, annotators 2, 4, 6, 10 are seen to be in high agreement with the expert. All of these annotators have an overlap above 65% for the pairwise comparisons. This corresponds to accuracy and Cohen’s $\kappa$ of 0.65 and 0.30. In summary, annotators 1, 2, 4, 6, 10 show a higher degree of agreement with the expert and form a strong
group in contrast to 3, 5, 7, 8, 9 which fall into a weak group.

The Kemeny-Young method is used to obtain ranking lists of 711 images which formed the set of 700 image pairs. The results obtained for full rankings and Kendall’s $\tau$ measure are consistent with the above observations. Similarly, for annotator 1, Kendall’s $\tau = 0.31$ which in Figure 3.6 corresponds to an agreement of $70\%-80\%$. The smallest correlation is seen for annotator 8 with Kendall’s $\tau$ of 0.08 which corresponds to $50\%-60\%$ agreement.

Another interesting observation is that for the randomly generated data, accuracy, Cohen’s $\kappa$ and Kendall’s $\tau$ are not consistent with each other. This result is as expected as this data is generated randomly. Moreover, accuracy ranges between 0 and 1 while Cohen’s $\kappa$ and Kendall’s
$\tau$ range between $-1$ and $1$ which also contributes to the inconsistency. On the contrary, values obtained from the measures for the real annotations are seen to be mostly coherent with each other. There are slight differences that occur due to two annotators having the same value for a measure.

It can be established from the observations that in contrast to the randomly generated data, for the real annotations most of the binary relations provided as an input to Kemeny-Young remain unchanged. For real annotations the number of honoured binary scores is high because there are fewer conflicting decisions within each annotated set than in the case of the random scores. This demonstrates the effectiveness of Kemeny-Young in converting binary decisions into a ranking list.

### 3.6.4 Inter-annotator agreement

After comparing the annotations with the expert, agreement between the annotators is evaluated. This can be used to find which annotators apply similar criteria in scoring fashion images. The values of accuracy and Cohen’s $\kappa$ for the random and real annotations are shown in Table 3.3 and Table 3.4.

For random annotations, similar observations to those made previously are seen where there is a lack of consistency between accuracy, Cohen’s $\kappa$ and Kendall’s $\tau$. As expected values close to 0.5 for accuracy and those near 0 for Cohen’s $\kappa$ and Kendall’s $\tau$ are seen. Higher p-values are observed for various $\tau$ values. As an example, highest p-value of 0.98 when $\tau$ is zero-valued is seen for A2-A8.

For real annotations, the highest agreement is observed between annotators 1 and 2 with an accuracy of 0.70 and Cohen’s $\kappa$ of 0.40. This corresponds to an agreement of 490 image pairs out of the 700 in this set and 70% agreement. Smallest value of 0.54 and 0.06 is seen for accuracy and Cohen’s $\kappa$ for inter-annotator agreement. These represent low level of agreement that is close to random for these annotators.

It was found from previous testing that annotators 1, 2, 4, 6, 10 had the most agreement and correlation with expert annotations. On further inspection of Table 3.4, a similar trend is seen. As an example, annotator 1 has the highest agreement with fellow annotators 2, 4, 6, 10. Similarly, annotator 2 has agreement with 1, 4, 6, 10 and additionally with annotator 9.
Chapter 3. The Dataset

Table 3.3: Results for the inter-annotator agreement for the random set of annotations where the accuracy is shown in the upper triangle and Cohen’s $\kappa$ is given in the lower triangle.

<table>
<thead>
<tr>
<th></th>
<th>A1</th>
<th>A2</th>
<th>A3</th>
<th>A4</th>
<th>A5</th>
<th>A6</th>
<th>A7</th>
<th>A8</th>
<th>A9</th>
<th>A10</th>
</tr>
</thead>
<tbody>
<tr>
<td>A1</td>
<td>0.49</td>
<td>0.51</td>
<td>0.49</td>
<td>0.53</td>
<td>0.49</td>
<td>0.48</td>
<td>0.53</td>
<td>0.49</td>
<td>0.49</td>
<td>0.49</td>
</tr>
<tr>
<td>A2</td>
<td>-0.01</td>
<td>0.51</td>
<td>0.48</td>
<td>0.49</td>
<td>0.50</td>
<td>0.51</td>
<td>0.49</td>
<td>0.49</td>
<td>0.49</td>
<td>0.51</td>
</tr>
<tr>
<td>A3</td>
<td>0.02</td>
<td>0.02</td>
<td>0.55</td>
<td>0.49</td>
<td>0.51</td>
<td>0.48</td>
<td>0.47</td>
<td>0.48</td>
<td>0.44</td>
<td></td>
</tr>
<tr>
<td>A4</td>
<td>-0.02</td>
<td>-0.04</td>
<td>0.11</td>
<td>0.50</td>
<td>0.48</td>
<td>0.50</td>
<td>0.49</td>
<td>0.49</td>
<td>0.49</td>
<td></td>
</tr>
<tr>
<td>A5</td>
<td>0.06</td>
<td>-0.01</td>
<td>-0.02</td>
<td>0.01</td>
<td>0.51</td>
<td>0.49</td>
<td>0.50</td>
<td>0.51</td>
<td>0.53</td>
<td></td>
</tr>
<tr>
<td>A6</td>
<td>-0.01</td>
<td>-0.00</td>
<td>0.02</td>
<td>-0.05</td>
<td>0.02</td>
<td>0.49</td>
<td>0.50</td>
<td>0.51</td>
<td>0.51</td>
<td></td>
</tr>
<tr>
<td>A7</td>
<td>-0.05</td>
<td>0.02</td>
<td>-0.04</td>
<td>-0.00</td>
<td>-0.03</td>
<td>-0.03</td>
<td>0.50</td>
<td>0.51</td>
<td>0.51</td>
<td></td>
</tr>
<tr>
<td>A8</td>
<td>0.05</td>
<td>-0.01</td>
<td>-0.07</td>
<td>-0.02</td>
<td>-0.01</td>
<td>0.01</td>
<td>-0.00</td>
<td>0.48</td>
<td>0.50</td>
<td></td>
</tr>
<tr>
<td>A9</td>
<td>-0.03</td>
<td>-0.03</td>
<td>-0.04</td>
<td>-0.06</td>
<td>0.02</td>
<td>0.03</td>
<td>0.01</td>
<td>-0.04</td>
<td>0.51</td>
<td></td>
</tr>
<tr>
<td>A10</td>
<td>-0.01</td>
<td>0.03</td>
<td>-0.11</td>
<td>-0.02</td>
<td>0.06</td>
<td>0.01</td>
<td>0.02</td>
<td>0.00</td>
<td>0.02</td>
<td></td>
</tr>
</tbody>
</table>

Table 3.4: Inter-annotator agreement for the real annotations where accuracy is shown in the upper triangle and Cohen’s $\kappa$ is presented in the lower triangle. Highlighted values represent the pairwise annotations that met the criteria when compared to the expert (cf. Section 3.6.1). A1-A2 is highlighted because both A1 and A2 meet the criteria compared to the expert annotations. Accuracy and Cohen’s $\kappa$ follow a pattern where the annotators who are in agreement with the expert annotations also agree amongst each other.

For the full rankings inter-annotator correlation is measured using Kendall’s $\tau$ where the rankings include 741 independent images for the 700 image pairs. Kendall’s $\tau$ for the random and real annotations is shown in Table 3.5 and Table 3.6. Smaller p-values are observed for the real inter-annotator combinations which represent that the correlation observed is significantly different from zero. This is in contrast to the random annotations where high values that are nearly equal to 1 are seen. It was found from previous testing that annotators 1, 2, 4, 6, 10 had the most correlation with expert annotations. Inter-annotator assessments are coherent with
higher correlation coefficients observed for annotators 1, 2, 4, 6, 10 (cf. Table 3.6). For example, annotator 1 has the highest correlation with 2, 4, 6, 10. It can be observed that annotations which met criteria when analysed against the expert also show high correlation with one another. The observations from agreement and correlation can be put together to rule out the weaker annotations 3, 5, 7, 8, 9 when generating the overall rankings.

Table 3.5: The results for inter-annotator correlation using Kendall’s $\tau$ for the random set of annotations. Kendall’s $\tau$ is shown in the upper triangle and the associated p-value is given in the lower triangle.

<table>
<thead>
<tr>
<th></th>
<th>A1</th>
<th>A2</th>
<th>A3</th>
<th>A4</th>
<th>A5</th>
<th>A6</th>
<th>A7</th>
<th>A8</th>
<th>A9</th>
<th>A10</th>
</tr>
</thead>
<tbody>
<tr>
<td>A1</td>
<td>-0.01</td>
<td>0.01</td>
<td>0.01</td>
<td>0.01</td>
<td>0.03</td>
<td>-0.05</td>
<td>0.02</td>
<td>-0.05</td>
<td>-0.01</td>
<td></td>
</tr>
<tr>
<td>A2</td>
<td>0.72</td>
<td>0.00</td>
<td>-0.04</td>
<td>0.00</td>
<td>-0.01</td>
<td>0.01</td>
<td>0.00</td>
<td>-0.01</td>
<td>-0.01</td>
<td></td>
</tr>
<tr>
<td>A3</td>
<td>0.68</td>
<td>0.92</td>
<td>0.11</td>
<td>-0.02</td>
<td>0.02</td>
<td>-0.06</td>
<td>-0.05</td>
<td>-0.04</td>
<td>-0.08</td>
<td></td>
</tr>
<tr>
<td>A4</td>
<td>0.71</td>
<td>0.10</td>
<td>0.00</td>
<td>0.00</td>
<td>-0.02</td>
<td>-0.03</td>
<td>-0.02</td>
<td>-0.02</td>
<td>0.00</td>
<td></td>
</tr>
<tr>
<td>A5</td>
<td>0.67</td>
<td>0.95</td>
<td>0.38</td>
<td>0.87</td>
<td>0.03</td>
<td>-0.06</td>
<td>-0.01</td>
<td>0.00</td>
<td>0.04</td>
<td></td>
</tr>
<tr>
<td>A6</td>
<td>0.27</td>
<td>0.69</td>
<td>0.49</td>
<td>0.35</td>
<td>0.23</td>
<td>0.00</td>
<td>0.01</td>
<td>0.03</td>
<td>0.01</td>
<td></td>
</tr>
<tr>
<td>A7</td>
<td>0.04</td>
<td>0.83</td>
<td>0.02</td>
<td>0.25</td>
<td>0.02</td>
<td>0.87</td>
<td>-0.02</td>
<td>0.02</td>
<td>0.02</td>
<td></td>
</tr>
<tr>
<td>A8</td>
<td>0.48</td>
<td>0.98</td>
<td>0.03</td>
<td>0.52</td>
<td>0.67</td>
<td>0.74</td>
<td>0.45</td>
<td>-0.02</td>
<td>0.04</td>
<td></td>
</tr>
<tr>
<td>A9</td>
<td>0.04</td>
<td>0.68</td>
<td>0.15</td>
<td>0.31</td>
<td>0.96</td>
<td>0.24</td>
<td>0.46</td>
<td>0.46</td>
<td>0.00</td>
<td></td>
</tr>
<tr>
<td>A10</td>
<td>0.58</td>
<td>0.73</td>
<td>0.00</td>
<td>0.91</td>
<td>0.12</td>
<td>0.74</td>
<td>0.51</td>
<td>0.07</td>
<td>0.87</td>
<td></td>
</tr>
</tbody>
</table>

Table 3.6: Inter-annotator correlation between the rankings from the real annotations using Kendall’s $\tau$ with its associated p-values. Kendall’s $\tau$ is given in the upper triangle and the p-values are given in the lower triangle where zero-valued p-values are removed. Highlighted values represent the annotations that are validated by the expert. For example, A1-A2 is highlighted with a value of 0.26 which corresponds to both A1 and A2 meeting the criteria compared to the expert. $\tau$ of 0.26 corresponds to a large overlap of 70% for the decisions of the pairwise comparisons.
3.6.5 Rankings from annotators

Based on the experiments performed in Section 3.6.3 and Section 3.6.4, annotators can be divided into two groups of stronger annotations 1, 2, 4, 6, 10 and weaker annotations 3, 5, 7, 8, 9. These give two rankings strong-5 and weak-5 using the Kemeny-Young implementation ($N_{try} = 2000$) described in Section 3.5 following the procedure in Section 3.5.2. An additional ranking all-10 is also generated by incorporating all the 10 annotations. This results in three groups of rankings from the annotations that can be utilized within the evaluation framework for comparison because they are sufficiently varied. As mentioned previously in Section 3.5.2, a full ranking i.e. ranking in the image space with 1064 elements and a compressed ranking i.e. ranking based on 120 configurations (attributes) are generated.

Some samples of the ordered dataset with top-10 ranked images for strong-5, weak-5 and all-10 are shown in Figure 3.8. Specific states for the images in the figure can be determined from Appendix A Table A.1. As an example image 64_10 is the top ranked image for strong-5 ranking and is the $10^{th}$ image of the 64$^{th}$ configuration. The 64$^{th}$ configuration represents states (4 1 4) for body shape, top and bottom clothing as determined from Table A.1. Specifically these represent a person with a pear body shape wearing a fitted top with a flared skirt (cf. Appendix A Section A.1). A set of pictures that illustrate the ordering over attribute space are given in Figure 3.9 showing top-5 and lower-5 ranked configurations for strong-5, weak-5 and all-10. For a specific attribute configuration, a number of example images for that configuration
are shown. The 98th configuration is the top ranked configuration for strong-5, this means that overall a preference for column body shape wearing a ruffled top combined with a fitted skirt is indicated and this has been validated by the expert.

![Image examples representing top-5 and lower-5 ordered configurations for strong-5, weak-5 and all-10 rankings.](image)

Figure 3.9: Image examples representing top-5 and lower-5 ordered configurations for strong-5, weak-5 and all-10 rankings.

Overall it was seen that based on validation criteria, annotations from some of the annotators were ruled out as part of a weaker set. And a stronger group of annotators was seen in which the annotations were validated by the expert. This indicates that there is a significant consistency in assessing images based on aesthetic qualities so there exists an aesthetics based global criteria.
This can be used to represent an image using an objective score and it should be possible to automatically learn these criteria from the provided rankings. Therefore this problem is not completely subjective and a general criteria for analysing fashion can be formed which represents the consensus of the industry.

3.7 Conclusions

An approach for obtaining an objective ranking using judgements made from the visual assessments of static images was presented. It used the knowledge of crowdsourced annotators captured in pairwise comparisons. Images for the comparisons were obtained from the dataset that was introduced in this chapter. This comprised of 1064 images fully labelled with various attributes of 11 clothing and 4 body shapes. The method was evaluated on a large number of 57400 image pairs that were annotated by 10 annotators and an expert. On validating annotations it was found that the validation procedure performed well and demonstrated that not all annotations were consistent with that of the expert and with one another. Specifically two groups of annotators, a stronger with annotators 1, 2, 4, 6, 10 and a weaker group with annotators 3, 5, 7, 8, 9 were observed. These groups were used to generate strong-5 and weak-5 rankings and an additional ranking all-10 was also obtained from all the 10 annotators. This was done by utilizing an implementation of the Kemeny-Young method after following a procedure for combining different annotations. Validation of some of the annotators by expert indicated that there was a consistency in the assessment of images based on aesthetics and so a global criteria can be established. This criteria was used to give rankings strong-5 and weak-5, which represent a general criteria for fashion.
Chapter 4

Detecting attributes automatically

This chapter presents method for obtaining an automatic labelling for attributes of clothing and body shape. The recognition method is first presented in Section 4.1. Then, a lookup model based on the estimated attributes and objective ranking is described in Section 4.2. Finally, performance of recognition and assessment model is presented in Section 4.3.

4.1 Attribute recognition

For recognition SVM classifiers with BoV approach based on SIFT features are used [69]. Local image descriptors (SIFT) are extracted from grey-value images to learn a visual vocabulary from a training dataset to form visual words similar to described in Chapter 2 Section 2.3.1. Then, feature coding is performed which results in mid-level features where each descriptor is represented by a subset of visual words from the visual vocabulary space. The mid-level features from an image are then transformed to an image signature vector by pooling. Finally, training and classification is performed on these image signatures with the use of SVM classifier [69].

There are 15 different attributes including 11 clothing and 4 for body shape. The training and testing of the attribute recognition is done using the dataset of 1064 images that was discussed in Chapter 3 Section 3.1.1 and illustrated in Chapter 3 Figure 3.2. Height normalized upper, lower and mid-level body crops for estimating the top clothing, bottom clothing and body
shapes respectively are used. For each category, the data is randomly split into training and test sets with 532 images in each. The positive training images for one category are used as negative examples for all the other categories. Similarly the positive test images for one category are used as negative test images for all the other categories. The classification decision can then be made based on a hard threshold using the confidence score output by the classifier. It can also be made by utilizing the label of classifier with the highest score. For example, for top clothing categories this would include all the attributes that would be worn on upper half of the body. Both the techniques are considered in the experimental evaluation when making classification decisions.

4.1.1 Performance evaluation

The performance for the clothing and body shapes is reported with the standard recall and precision measures with the confidence interval for the mean at 95% confidence level.

Recall and Precision: These measures vary between 0 and 1 and are given in Equations 4.1 and 4.2, where $TP, FN, FP$ are the true positive, false negative and false positive [97]. Recall is the number of true positives divided by the total examples that belong to the positive class. In the recall formulation, sum of the true positive and false negative is also known as the ground truth, which is the true label of a test pattern. Precision is the number of true positives divided by the total examples that are classified by the classifier as being part of the positive class. As an example, consider a classification problem in which there is a positive class $P$, a negative class $N$ and a classifier $C$. A recall of 1 for $P$ represents that all test patterns that belong to $P$ are labelled as positive by $C$. However, this does not give any information about the test patterns of $N$ that are assigned to $P$ by $C$. A precision of 1 for $P$ indicates that all test patterns classified by $C$ as being part of $P$ indeed belong to it. But it does not give any indication of the test patterns within $P$ that are missed by $C$. Generally recall and precision are inversely related i.e. one can be increased at a cost of reducing the other.

\[
\text{Recall} = \frac{TP}{(TP + FN)} \quad (4.1)
\]
4.2 Rankings based on attribute recognition

A number of rankings were obtained from the crowdsourcing procedure in Chapter 3 Section 3.6.5 so that judgements made by annotators can be utilized. As a baseline approach such a ranking can be used as a lookup model for automatic assessment of images based on aesthetic qualities. Each configuration of clothing and body shape attribute is represented within the ranking. Using a recognition system trained for every attribute, the configuration of a test image can be identified and a rank can be assigned to that image. Two independent images can be ranked and compared with each other thus pairwise comparisons can be automatically performed. In the real-world scenario there would generally be an error associated with recognition of various attributes, this error can be included within the lookup model. To obtain this, validated annotations along with an estimate of attribute errors are used as the lookup ranking.

\[
\text{Precision} = \frac{TP}{(TP + FP)} \tag{4.2}
\]

**Confidence intervals for means:** A confidence interval gives an estimated range of values which is likely to include an unknown population parameter. The estimated range is determined using a given set of sample data [5]. A confidence level \(C\) corresponds to the percentage of area in the normal density curve therefore 95\% confidence interval encompasses 95\% of the normal curve. This means that probability of observing a value outside this area is less than 0.05. For the evaluation being performed, the interval contains averaged performance measures of recall and precision with 95\% confidence. In this work and most of other practical research the standard deviation is not known therefore the standard error is estimated which is the estimated value of standard deviation. So the sample mean follows the \(t\) distribution and the \(t\) distribution is described by its degrees of freedom. A confidence interval for the population mean for a population with unknown mean and unknown standard deviation is given by \(X \pm t^* \frac{s}{\sqrt{n}}\), where \(X\) is the sample mean, \(n\) is the sample size, \(t^*\) is the upper \((1 - C)/2\) critical value for the \(t\) distribution and has \(n - 1\) degrees of freedom. For a 95\% confidence interval for the mean, the 0.025 critical value, i.e., \(t^*\) for 4 degrees of freedom \((n = 5\) here\) is given by 2.776.
Chapter 4. Detecting attributes automatically

4.2.1 Lookup model

The stronger annotations 1, 2, 4, 6, 10 (strong-5) consist of 29400 comparisons which includes 1400 control image pairs (cf. Chapter 3 Section 3.6.5). These along with an estimate of errors based on recognition are employed to generate ranking for the lookup model. Besides strong-5, other rankings such as weak-5 from the weaker group of annotators 3, 5, 7, 8, 9 and all-10 from all the annotators will also be utilized for evaluation. In addition to rankings from annotators, a simulated dataset with random annotations is also used. This gives a random ranking rand-10 for the 10 annotators which will provide a baseline when testing the performance.

4.2.2 Accuracy measure

Any pair of images can be ranked based on scores that are output by the method for each of the images. The performance of the method can be evaluated by measuring the fraction of all pairs that were correctly ranked. The output of this measure is correlated with Kendall’s $\tau$ but it is more intuitive. A value of 1 for accuracy corresponds to the scenario when all the pairs were ranked correctly and a value of 0 indicates that none of the pairs were ranked correctly, while a value of 0.5 indicates random ranking.

For this instance, accuracy of the lookup model is determined using rankings of the various annotation groups, such as, strong-5, weak-5 and all-10. The evaluation is performed using compressed ranking of 120 configurations, the procedure for obtaining such a ranking was outlined in Chapter 3 Section 3.5. So if there are $C = 120$ configurations it gives $(C^2 - C)/2 = 7140$ total number of paired-configurations which is the normalization factor. Therefore, accuracy of 1 indicates that all 7140 paired-configurations are correctly scored. It should be noted that it is this accuracy that is used in the following chapters as an evaluation measure.

4.3 Experimental results

In this section performance of the recognition approach and lookup model is evaluated. The evaluation for recognition of various attributes is first presented in Section 4.3.1. And the overall approach is tested in Section 4.3.2, which incorporates a recognition of clothing and body shape attributes into the rankings that meet the evaluation criteria for fashion interpretation.
4.3. Experimental results

Figure 4.1: Examples of images from a split when evaluating the recognition system that have top clothing attributes categorised correctly and falsely. Correctly recognized images are marked with a tick and false categorisations are represented with a cross.

4.3.1 Attribute recognition

For an automatic ranking of images the attributes are recognised with the use of BoV approach using 128D SIFT features from $10 \times 10$ dense sampling [69]. The results are shown in Table 4.1 using recall and precision with confidence interval for the mean at 95% confidence level (cf. Section 4.1). These results are for averaged 5 runs of random splits using different thresholds for classifier scores as well as the maximum response of the set of classifiers. A high performance for the 11 clothing categories is obtained with different threshold settings and in particular when taking the label of the maximum prediction value $p_{max}^*$. That is why for measuring the performance of attribute recognition, values from $p_{max}^*$ will be used. Examples of images for clothing attributes which have been categorised correctly and falsely are shown in Figure 4.1 and Figure 4.2. Both, precision and recall are very high for the clothing attributes due to their distinctive shape characteristics but much lower for body shapes with an average recall and precision of $0.30 \pm 0.08$. 
Specifically the reason for this performance decline for body shapes comes from the very subtle differences in features extracted from them. In addition, the quantisation of SIFT features and spatial bins of the pyramid do not allow to capture these variations. Experiments with different variants of feature extractors were also performed but no significant improvement was observed. The best performance of $0.43 \pm 0.06$ was observed for apple body where the shape differences are the largest compared to column or hourglass. Overall, the classification error is below 10% for most of the other attributes and such an error is realistic for state-of-the-art visual classification systems. As an example, from one of the recent results lowest error of 11% was found for the classification task of the recognition challenge organized by ImageNet [3].
4.3. Experimental results

Figure 4.2: Examples of images from a split when evaluating the recognition system that have bottom clothing attributes categorised correctly and falsely. As done previously correct and false categorisations are marked with a tick and cross respectively.

4.3.2 Rankings with attribute recognition

The error due to attribute recognition has an impact on the accuracy of assessments made from a ranking. This impact is evaluated in this experiment by incorporating the attribute errors due to recognition. The errors are included using four different methods. The actual errors for all the 15 influencing attributes with the performances measured from Table 4.1 are included first. In the latter scenarios averaged performance for different attributes with and without the body shape attributes are used.
Chapter 4. Detecting attributes automatically

**Lookup model:** The accuracy of automatic comparison when testing using strong-10, all-10, weak-5 and rand-10 is displayed in Figure 4.3. The averages for errors are measured at 7%, 10% and 23%. The 7% error is calculated by averaging the performance for all the attributes except body shape. In the 23% error, performance average including the body shape estimate is used. A value of 10% is chosen to reflect approximate performance based on the 7% error. When the errors are included in recognition it is possible that the paired configuration being tested could be mapped to the same configuration. In this case when the scoring is compared in ranking, both configurations will go to the same position. The goal is to automatically compare two images from different configurations, therefore for calculating accuracy such a result is taken to be an incorrect one.

![Figure 4.3: Correctly ranked pairs between 0 and 1 when using ranking with the recognition based error estimates. The actual error estimates along with the errors when an averaged performance is considered are shown.](Figure_4.3.png)

Overall, highest scores are seen when the error is measured at 7% (highest value of 0.96 for strong-5). This result is as expected because the error included in the ranking for the lookup is the lowest when compared with other cases. Another observation that can be made is that highest scoring is achieved for strong-5 in contrast to all-10, weak-5 and rand-10. This is due to this group of annotators (strong-5) applying similar criteria on being tested against the expert. The all-10 ranking produces reasonably high scores of 0.69, 0.85, 0.83, 0.78 for the actual and error estimates of 7%, 10% and 23% compared to scores of 0.61, 0.72, 0.71, 0.67 for weak-5. This is an indication that on using a large enough number of pairwise comparisons, some of the noise from the weaker annotations can be reduced.

The weaker annotations show a lower score compared to strong-5 and all-10 as expected. This is most likely due to these annotators applying different criteria when making fashion judge-
ments. For random ranking rand-10, a value of 0.54 is seen which is the lowest value for this experiment. This is because this ranking is not correlated with any of the annotations. In summary, it can be established that the annotated rankings provide a significant gain in performance when compared with the baseline random annotations. It was also found that the validated annotation strong-5 provides better results when compared to weak-5 and all-10.

To serve as a baseline in the following chapters, accuracy at 10% error will be utilized because this error represents an averaged performance. Accuracy of 0.94, 0.83, 0.71 and 0.54 is obtained for strong-5, all-10, weak-5 and rand-10 at this error and these values represent a baseline performance.

4.4 Conclusions

In this chapter method for detecting the attributes automatically was presented and a lookup model based on the estimates from this recognition was evaluated. The recognition demonstrated a high performance for the 11 clothing attributes with the best results obtained when using the label of the maximum prediction estimate. However, a decline in performance was seen for the body shape attributes with the apple body shape showing better results compared to the other body shapes. This was because the shape differences are the largest for apple body shape compared to other shapes. The evaluation was performed using measures of recall and precision with associated 95% confidence interval. High values of recall and precision with an average of 0.93 ± 0.05 for the 11 clothing attributes were observed. And the averaged value for body shape was reduced to 0.30 ± 0.08. Due to the classification error being low for attributes except body shape, prediction estimates from those attributes without body shape will be utilized when evaluating the ranking approaches proposed in the following chapters.

On testing the lookup model effectiveness of the approach was seen where the results indicated that on using a sufficiently large number of comparisons, noisy assessments made by non-expert annotators could be filtered out. This model also provides an effective baseline for testing pairwise comparisons with reference to accuracy. Particularly, accuracy values of 0.94, 0.83, 0.71, 0.54 obtained for strong-5, all-10, weak-5 and rand-10 in Section 4.3.2 at 10% error estimate will be used.
Chapter 5

Ranking images using matching

In this chapter rankings obtained from crowdsourced annotators are utilized to generate rankings using the nearest neighbour search approach. An overview of how the various matching techniques are incorporated is presented in Section 5.1. Matching is performed using the Bag of Visual Words (BoV) and Local Descriptor Matching (LDM) approaches. The procedure for obtaining a global ranking using the matches obtained is also discussed. Next, dataset, parameters and measures for evaluation are explained in Section 5.2. Finally, performance of matching and how that determines the global rankings is presented in Section 5.3.

5.1 Background

For a dataset of $N$ images, each image is assigned an absolute score where the set of all the scores is given by $s = s_1, s_2, ..., s_N$. This ranking is based on the judgements made by annotators using the ranking approach outlined in Chapter 3 Section 3.5. An intuitive method to utilize such rankings would be to obtain the nearest neighbour match for an image. Such an approach acts as a brute force method for the ranking problem and therefore does not rely on automatic recognition of attributes performed in Chapter 4. Images obtained after nearest neighbour matching can be compared with their ordering in rankings such as strong-5, weak-5 and all-10 that were obtained in Chapter 3 Section 3.6.5. And this comparison can be used to determine which image is ranked better to perform a comparison of aesthetics within images.

In this work two leading approaches of BoV model and LDM are considered to evaluate ranking
based on matching [31, 93, 104, 114, 72]. A brief description of these is given in the following sections. This is done with reference to parameters and settings utilized to test the performance. The BoV and LDM approaches can be seen as variations of the basic algorithm shown in Figure 5.1. Images from the database are presented to the feature extractor and feature vectors are generated which are utilized to construct a data structure. Similar procedure is performed at the query end which results in the query feature vectors. Nearest neighbour search is performed with the aid of the data structure and matches are obtained after sorting the database.

5.1.1 Matching with Bag of Visual Words (BoV)

The query and database images are represented as a sparse vector of visual word occurrences and on search the similarity between the image vector and query vector is estimated [114]. The steps for achieving a BoV representation for describing visual content are shown in Figure 5.2 [62].

Figure 5.1: General process of achieving image matching.

Figure 5.2: An illustration of steps for creating BoV as given in [62].
5.1. Background

The tf-idf weighting scheme has been shown to perform well and is used. If there is a vocabulary of $w$ words given by a vector of length $w$, $V = (v_1, v_2, ..., v_w)$, then each element of this vector is defined as in Equation 5.1 under the tf-idf scheme. The weighting is a product of two terms, the words frequency, $\frac{n_{wd}}{n_d}$, where $n_{wd}$ is the number of times the word $w$ occurs in $d$, $n_d$ is the total number of words in $d$ and the inverse document frequency, $\log \frac{N}{n_w}$, where $N$ is the total number of images (or documents in the database) and $n_w$ is the number of occurrence of word $w$ in the complete database.

$$v_w = \frac{n_{wd}}{n_d} \log \frac{N}{n_w} \tag{5.1}$$

**Inverted file:** In the inverted list, there is an entry for a visual word followed by the image indices of the images in which that visual word is present [114]. Using this technique avoids overheads and comparisons between high dimensional descriptors along with reducing the number of candidate images. This especially offers a speed-up for sparse queries since only database images containing the visual words need to be checked rather than checking all the images. Similarity is found by storing scores for all the relevant images.

**Min-Hash algorithm:** Efficient matching relies on the chosen representation of an image and the similarity metric used for comparison when retrieving. This makes hashing algorithms a powerful tool for use in large scale search applications. The min-Hash algorithm originates from the text retrieval community where it is used for returning duplicate documents [34]. It has been shown in literature to be effective for visual data due to several reasons. Firstly, it results in a compact image representation with a constant amount of data needed to be stored. It also allows fast search time that is proportional to the number of retrieved matches and can generate accurate results with reasonable probability.

The BoV model represents an image as a fixed length vector of length $|V|$, where $|V|$ is the size of the vocabulary. Each element (or visual word) of this vector is a representation of the number of features that are encoded by that visual word. A set representation which is a weaker representation is used for min-Hash given by $A_i \subset V$. This uses only limited information of whether a feature is present or not resulting in a binary vector. The distance between two images is computed using Jaccard measure for set similarity. This is defined by the ratio of the number of elements in intersection and union as shown in Equation 5.2:

$$\text{sim}(A_1, A_2) = \frac{|A_1 \cap A_2|}{|A_1 \cup A_2|} \tag{5.2}$$
similarity is determined when this measure exceeds a certain threshold. The min-Hash is a LSH function that selects an element from each set \( A_i \) of visual words for an image \( i \). A number of \( N \) independent hash functions \( f_j \) are used and min-Hashes generated are recorded. Every min-Hash function infers a certain ordering on the set and min-Hash of the set is then defined as where \( v \in A_i \):

\[
m(A_i, f_j) = \arg\min f_j(v)
\]  

(5.3)

The method is based on probability of two sets having the same min-Hash being equal to the similarity of the two sets. This has been defined by the aforementioned Jaccard measure in Equation 5.4.

\[
P(m(A_1, f_j) = m(A_2, f_j)) = \frac{|A_1 \cap A_2|}{|A_1 \cup A_2|} = \text{sim}(A_1, A_2)
\]  

(5.4)

A single min-Hash being a visual word is no more distinctive than a single visual word. In order to increase the discriminativeness multiple min-Hashes are grouped into \( s \)-tuples called sketches. This results in increased precision, but also leads to a lower recall. To increase the recall this process is usually repeated \( n \) times. This allows the computation of \( n \) sketches with \( n \times s \) independent hash functions. The probability of sketch collision and retrieval, where at least one of the \( n \) sketches (hash tables) must collide are given by Equation 5.5 and Equation 5.6.

\[
P(\text{sketch collision}) = \text{sim}(A_1, A_2)^s
\]  

(5.5)

\[
P(\text{retrieval}) = 1 - (1 - \text{sim}(A_1, A_2)^s)^n
\]  

(5.6)

When indexing, each sketch is inserted into a hash table such that the \( i \)-th sketch is inserted into the \( i \)-th hash table. For querying, min-Hashes and sketches are generated with the same hash functions used for indexing. The \( i \)-th hash table is queried to check if it contains the \( i \)-th query sketch. If this is the case then there is similarity between the images.

5.1.2 Matching with Local Descriptor Matching (LDM)

The LDM representation involves computation of the full set of descriptors from all images in the database. Then the most similar feature is found by employing \( k \)-nearest neighbours search over these descriptors. A voting criterion using the \( k \)-nearest descriptor distances found is finally used to output a list of matches.
5.1.3 Global ranking based on matching

Matches obtained from different approaches can be used to establish a global ranking in which an image configuration is assigned a score. An approach that can be utilized to obtain such a ranking would be to view this process as a two-fold problem. In the first stage, every query is assigned to a unique output image after matching. Next from this output, correctly ranked pairs based on the annotated rankings can be used to achieve the final ranking.

The goal of finding a unique match for every test image can be viewed as a combinatorial optimization problem. A member of the set that optimizes an objective function in a solution space that is large but finite can be found. This can be considered as an assignment problem for which Munkres assignment algorithm can be used to solve it in polynomial time [70]. In this case, there will be a non-negative $\frac{N}{2} \times \frac{N}{2}$ matrix, where an element in the $i$th row and $j$th column represents the cost of assigning the $j$th match to the $i$th query.

To achieve a matching a sorted list of all the images for a query is retrieved. A matrix is populated with cost assignments that vary in accordance with the matching method utilized. For example, it can be given by a function that is a representation of the distance between query and its match. It could also be a value derived from such a distance after a voting scheme has been applied. For a voting-based cost function, goal of the problem is altered and it is required to maximise the cost in order to find an assignment. That is why each cost is replaced by the maximum cost subtracted from the cost. After a complete matrix has been formed for all the queries, Munkres assignment algorithm is applied.

The comparisons from such a matching output can be used as an input to the Kemeny-Young method for ranking. Matches obtained from a pair of query images are compared to their ranking in the annotated lists. This is done to establish which image is ranked better as shown in Figure 5.3. Matches with their pairwise order serve as a comparison input to the Kemeny-Young method. If there are $N/2$ images in the database and $N/2$ query images, then in an ideal scenario Munkres algorithm would result in $N/2$ matches. This would result in a total number of $\left(\frac{N^2}{4} - \frac{N}{2}\right)/2$ unique comparisons for the voting input. After a ranking is obtained, rank of images part of the same configuration can be averaged to obtain a configuration list similar to Chapter 3 Section 3.5.2. A configuration-based ranking can be used to perform comparisons of image configurations based on aesthetics. This averages the ranking scores of different images.
that belong to the same configuration to give a configuration-based ranking and hence provides an averaged solution when comparing image configurations.

Figure 5.3: A pair of query images can be checked for their ranking in an annotated ranking. Matches for this query are also checked for their ranking in the ranking. If the rankings agree then that is used as a correct output in the voting file and if not it is included as an incorrect match. This voting file is input to the Kemeny-Young method to obtain an output ranking.

5.2 Evaluation design for ranking

This section presents details of the evaluation approach used to assess the performance of various matching methods. Firstly, it describes the data used to carry out the experiments together with how it has been incorporated. Next, it discusses the three evaluation measures used to test the rankings generated. Lastly, it outlines the different testing scenarios with their associated parameters.

5.2.1 Dataset

The prototypes for performing evaluations are formed by dividing the dataset into five random splits. This is done by forming sets with equal number of test and training images based on the configuration of an image. For example, consider a configuration of a person with an apple body shape dressed in fitted skirt and loose top clothing. Then from the total $M$ images for this configuration, the test and train set will have $M/2$ images each. If in an instance where such a distribution is not possible, sets will be formulated by considering the required number of images for even sets. This ensures that for every test image there is a matching image in the training set. Images from the dataset that was introduced in Chapter 3 Section 3.1.1 are
5.2. Evaluation design for ranking

used for the random splits and the results from the matching approaches tested are averaged over 5 runs of random splits. The test sets are labelled with the particular states of the body shape, top clothing and bottom clothing, i.e, image configuration based on the dataset collection performed previously (cf. chapter 3 Section 3.1.1). Generation of input rankings for matching requires three groups of annotator rankings. These are strong-5, weak-5 and all-10 computed after various consistency checks were applied as described in Chapter 3 Section 3.6.5.

5.2.2 Measures

The test images are matched to images within the training set. From the matches found local or global rankings can be computed to assess performance. Three main measures for evaluation are utilized which test different scenarios of rankings.

**Absolute accuracy**: A pair of test images with their associated matches can be compared to their ranking in the annotated lists (strong-5, weak-5 and all-10) as shown in Figure 5.4. The performance can be evaluated from this as discussed in Chapter 4 Section 4.2.2. In contrast to the accuracy measure described in the previous chapter, both image-based and configuration-based rankings are employed. This is because they provide different measures for evaluating the matching performance. Furthermore variation also comes from the normalization factors when computing accuracy. For image-based ranking, if there are \( N \) images then there are \( (N^2 - N) / 2 \) unique pairs, so this is the normalization factor for image-based accuracy. However, this factor is reduced for configuration-based accuracy by the total number of paired-images within a configuration. So the formulation for this is \( \frac{(N^2 - N)}{2} - \sum_i C \cdot \frac{(c_i^2 - c_i)}{2} \), where \( C \) are the total number of configurations and \( c_i \) is the number of images in each configuration. When each configuration has only one image, this equates to \( (N^2 - N) / 2 \) pairs. It can be seen that this is indeed the normalization constant for image-based ranking.

The image-based ranking enables methods to be evaluated more comprehensively since the image pairs retrieved can be compared directly. This is particularly beneficial when testing different matching methods. On the contrary, configuration-based measure provides a more global measure. This is because only image pairs from different configurations are considered, therefore making this less prone to noise. It is also more tailored to the application scenario of comparing images that belong to different configurations.
Figure 5.4: A pair of test images with their associated matches can be compared to their ranking in the annotated lists.

For a test-pair there are three instances for the matches found. The first one being when matches can be found for both the inputs. Secondly there is a match obtained for only one of the inputs. The last scenario involves when no matches can be determined for either of the images. For the latter cases when there are missing matches, the decision is considered to be incorrect in accuracy. When matches are obtained for both the input images, there are two further possibilities. These include that they are either mutually exclusive or not. Only matches that are mutually exclusive are considered correct. The remaining matches are regarded as incorrect similar to that done previously in Chapter 4 Section 4.3.2.

**Accuracy measure as assignment problem:** There are equal number of images in the training and test set. So there is a match possible for every test image. This makes it possible to deduce a unique matched list in which an image is not repeated. This measure is in contrast to absolute accuracy because in this case two test images can not be matched to the same image. Munkres assignment algorithm can be used to solve it in polynomial time. After this is applied to matches, accuracy described above can be computed. The valid test pairs are used as in absolute accuracy. Only this time it is done after every test image has been assigned an independent image from the training set. Due to non-repeated matches there are no instances when a pair of test images are matched to the same outcome for the image-based ranking. However, they still exist for configuration-based ranking. This is because a test-pair can be matched to two different images that belong to the same configuration. To maintain consistency it is first attempted to match every test image to an independent image from the training set. When a match can not be found then that instance counts as incorrect result which lowers the accuracy. Such an occurrence is a result of the matching method’s inability to return all the
N/2 matches from the training set for a particular test image.

**Kendall’s \( \tau \):** Kendall’s \( \tau \) evaluates correlation between the ranked configurations found by annotators and that generated after matching. To compute this global rankings need to found using the method presented in Section 5.1.3. These rankings can be compared with strong-5, weak-5 and all-10 using Kendall’s \( \tau \) as described in Chapter 3 Section 3.4.1.

**Accuracy measure:** Rankings found above when computing Kendall’s \( \tau \) can be utilized to test pairs of configurations using accuracy detailed in Chapter 4 Section 4.2.2.

### 5.2.3 Setup

The BoV and LDM approaches are based on extracting local features from an image. SIFT features extracted from grey value images using \(10 \times 10\) dense sampling as in Chapter 4 Section 4.3.1 are employed for this. The evaluation uses an implementation by [8, 7] and [72] for the BoV and LDM approaches. An overall parameter setup for these is given in Table 5.1 which are explained next.

<table>
<thead>
<tr>
<th>Method</th>
<th>Parameters</th>
</tr>
</thead>
<tbody>
<tr>
<td>BoV</td>
<td>(</td>
</tr>
<tr>
<td></td>
<td>1: (none, (l_1), (l_2)), 2: (none, (l_2), (l_2)), 3: (tf-idf, (l_2), cos), 4: (bin, (l_1), (l_1)) and 5: (bin, (l_2), (l_2))</td>
</tr>
<tr>
<td></td>
<td>(1: (1, 1), 2: (1, 2), 3: (1, 3), 4: (5, 1), 5: (5, 2), 6: (5, 3), 7: (25, 1), 8: (25, 2), 9: (25, 2), 10: (100, 1), 11: (100, 2), 12: (100, 3))</td>
</tr>
<tr>
<td>LDM</td>
<td>(k = 50, \text{dist} = \ell_2)</td>
</tr>
</tbody>
</table>

Table 5.1: The parameters used in the setup when evaluating matching approaches of BoV and LDM. They vary for BoV representation with 5 and 12 settings for the inverted file and min-Hash structures.

**BoV representation:** The parameter variation for this representation occurs at two levels. In the first, parameters associated with building the vocabulary in the BoV implementation are varied. The next one includes variations as a result of the inverted file and min-Hash data structures, which are used to search through the histograms within the training set.

The parameters pertaining to the vocabulary include how the vocabulary is generated and the number of visual words in the vocabulary:
Based on [8, 7] a flat vocabulary built from approximate k-means (AKM) using a set of random kd-trees to perform the nearest neighbour search is employed.

A larger vocabulary generally results in better performance therefore for comparison this parameter is assigned values of 1K, 10K, 50K, 100K, 500K and 1M.

For the inverted file structure parameters varied include weighting of the histogram counts, normalization of the histograms and distance function [8, 7]. All the three parameters have three cases each:

- The weighting of the histogram counts includes the use of raw histogram counts, binarizing the histogram counts and tf-idf weighting.

- The histograms can be used unnormalized, use $l_1$ normalization i.e. $\sum_i h_i = 1$ or $l_2$ normalization i.e. $\sum_i h_i^2 = 1$.

- For distance functions $l_1$ i.e. $d_{l_1}(h_x, h_y) = \sum_z |h_{xz} - h_{yz}|$, $l_2$ i.e. $d_{l_2}(h_x, h_y) = \sqrt{\sum_z (h_{xz} - h_{yz})^2}$ and cosine distance i.e. $d_{cos}(h_x, h_y) = 1 - \sum_z h_{xz}h_{yz}$ functions are used.

- Based on [8, 7] combinations for these parameters of weighting, normalization and distance function are (none,$l_1$,l_1), (none,$l_2$,l_2), (tf-idf,$l_2$,cos), (bin,$l_1$,l_1) and (bin,$l_2$,l_2).

The min-Hash method has two parameters including the number of hash tables and number of hash functions:

- The number of hash tables is varied from 1, 5, 25, and 100.

- There are 1, 2, and 3 hash functions for the hash tables.

**LDM:** The implementation employed for the evaluation of LDM is used as given by [72], where k=50 for k-nearest neighbours search with utilization of the $l_2$ distance function.

### 5.3 Experimental results

In this section performance of the matching methods is evaluated. The $BoV$ representation with inverted file data structure is first tested in Section 5.3.1. Then, performance of the minHash algorithm to determine its suitability is assessed in Section 5.3.2. LDM is next tested and
analysed in Section 5.3.3. This is followed by evaluating the performance of global ranking in Section 5.3.4. This has been generated based on evaluations performed in the previous experiments of this section. Finally, results of the evaluation along with suitability of various matching methods for ranking are summarized in Section 5.3.5.

**Data:** To measure performance, matches generated for the test images are compared with strong-5, weak-5 and all-10 rankings using both image and configuration-based rankings (cf. Section 5.2.2). The dataset with 120 configurations and 1064 images is split into 524 images in the training set and test set due to reasons mentioned in Section 5.2.1. For the image-based ranking, there are 524 images in the training set resulting in $(524 \times 523)/2$, i.e., 137026 unique pairs. The number of these pairs is reduced to 136050 for the configuration-based ranking. This is because in a matching pair each image must be from a different configuration.

There are five random splits of training and test images for every test and results for these are averaged.

**Measures:** Absolute accuracy, accuracy after assignment, Kendall’s $\tau$ and accuracy as described in Section 5.2.2 are used for evaluation.

**Setup:** The BoV representation and the LDM setup is employed as outlined in Section 5.2.3.

### 5.3.1 BoV with inverted file

In this experiment performance of the BoV representation with inverted file data structure is tested using measures of absolute accuracy and accuracy after assignment over a range of parameter settings.

**Absolute accuracy:** For vocabulary of size 1K with different BoV settings, this is given in Table 5.2. As expected the scores are better compared to random scoring. Highest values of 0.683 and 0.630 are obtained for the image-based and configuration-based ranking. These are found for setting 2 which is the (none,$l_2$,$l_2$) parameter when using strong-5. It can also be noted that the values vary marginally for different parameter settings of the inverted file. There is also little variation when using strong-5, weak-5 and all-10 rankings with highest values for strong-5. Moreover, it can be seen that the accuracy for image-based ranking is higher when compared with its configuration-based counterpart. This is because the configuration-based
ranking is a more global version of the image-based ranking. Due to which the percentage of correctly ranked pairs is likely to be higher when evaluating test image pairs as opposed to configuration test image pairs.

<table>
<thead>
<tr>
<th>meth</th>
<th>strong-5</th>
<th>weak-5</th>
<th>all-10</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>im</td>
<td>config</td>
<td>im</td>
</tr>
<tr>
<td>(none, l₁, l₁)</td>
<td>0.679</td>
<td>0.624</td>
<td>0.668</td>
</tr>
<tr>
<td>(none, l₂, l₂)</td>
<td>0.683</td>
<td>0.630</td>
<td>0.665</td>
</tr>
<tr>
<td>(tf-idf, l₂, cos)</td>
<td>0.683</td>
<td>0.627</td>
<td>0.666</td>
</tr>
<tr>
<td>(bin, l₁, l₁)</td>
<td>0.679</td>
<td>0.623</td>
<td>0.667</td>
</tr>
<tr>
<td>(bin, l₂, l₂)</td>
<td>0.670</td>
<td>0.619</td>
<td>0.653</td>
</tr>
</tbody>
</table>

Table 5.2: Image and configuration-based absolute accuracy for BoV with vocabulary of size $|V| = 1K$ for different parameters of inverted file (averaged over five runs of random splits).

On testing the parameters for increased vocabulary sizes of $|V| = 10K, 50K, 100K$, it was found that there were no significant improvements in the accuracy. To show the minimal increase in accuracy when larger vocabularies are used, accuracy for 100K words is shown in Table 5.3.

<table>
<thead>
<tr>
<th>meth</th>
<th>strong-5</th>
<th>weak-5</th>
<th>all-10</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>im</td>
<td>config</td>
<td>im</td>
</tr>
<tr>
<td>(none, l₁, l₁)</td>
<td>0.684</td>
<td>0.622</td>
<td>0.673</td>
</tr>
<tr>
<td>(none, l₂, l₂)</td>
<td>0.687</td>
<td>0.624</td>
<td>0.663</td>
</tr>
<tr>
<td>(tf-idf, l₂, cos)</td>
<td>0.687</td>
<td>0.625</td>
<td>0.660</td>
</tr>
<tr>
<td>(bin, l₁, l₁)</td>
<td>0.684</td>
<td>0.627</td>
<td>0.672</td>
</tr>
<tr>
<td>(bin, l₂, l₂)</td>
<td>0.683</td>
<td>0.622</td>
<td>0.672</td>
</tr>
</tbody>
</table>

Table 5.3: Image and configuration-based absolute accuracy for BoV with vocabulary of size $|V| = 100K$ for different parameters of inverted file (averaged over five runs of random splits).

To improve the accuracy of the inverted file structure, the vocabulary size is increased to larger sizes of $|V| = 500K$ and $1M$. This has shown to increase performance in literature [59, 103]. There are minimal differences between $|V| = 500K$ and $1M$. Results for $|V| = 1M$ are shown in Table 5.4. From these it can be seen that accuracy for parameter settings 2 has values of 0.689 and 0.625 for the image and configuration-based ranking (strong-5). For comparison, these values can be compared to 0.683 and 0.630 when $|V| = 1K$. It can be noted that increasing the vocabulary size to $1M$ is only marginally beneficial for some cases when compared with 100K. Moreover, it can be observed that the accuracy for configuration-based ranking generally
5.3. Experimental results

decreases when vocabulary words are increased. A reason could be that matching performance for images increases to some degree with a larger vocabulary. However, this may not apply to the configurations. An image in the original dataset was warped multiple times to reflect the 4 body shapes. This means that each image occurs 4 times in the dataset. Hence, it is likely that an image may possibly be matched to its warped version for a different body shape resulting in an incorrect configuration match.

<table>
<thead>
<tr>
<th>method</th>
<th>strong-5</th>
<th>weak-5</th>
<th>all-10</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>im</td>
<td>config</td>
<td>im</td>
</tr>
<tr>
<td>(none, l₁, l₁)</td>
<td>0.685</td>
<td>0.620</td>
<td>0.669</td>
</tr>
<tr>
<td>(none, l₂, l₂)</td>
<td>0.689</td>
<td>0.625</td>
<td>0.667</td>
</tr>
<tr>
<td>(tf-idf, l₂, cos)</td>
<td>0.688</td>
<td>0.625</td>
<td>0.667</td>
</tr>
<tr>
<td>(bin, l₁, l₁)</td>
<td>0.684</td>
<td>0.618</td>
<td>0.668</td>
</tr>
<tr>
<td>(bin, l₁, l₂)</td>
<td>0.682</td>
<td>0.618</td>
<td>0.666</td>
</tr>
</tbody>
</table>

Table 5.4: Image and configuration-based absolute accuracy for BoV with vocabulary of size \(|V| = 1M\) for different parameters of inverted file (averaged over five runs of random splits).

Accuracy after assignment: Accuracy for vocabulary with 100K words is given in Table 5.5. It is observed that the values are lower when compared with absolute accuracy of the same vocabulary. This is possibly because every test image is assigned a unique match so a ranking without any repeated images can be obtained. As a result of which a test image may not be assigned to its true match in some cases.

As noted previously (absolute accuracy), there are only slight differences between accuracy for different parameters of the inverted file. Moreover, marginally higher accuracy is generally observed for strong-5. When comparing vocabularies with 10K, 50K and 100K words, minimal differences are observed. The primary improvement occurs when these vocabularies are compared with 1K words where the accuracy is mostly lower. For example, accuracy of 0.629 and 0.586 with parameter 3 for image and configuration-based ranking is observed with 1K words (strong-5). This increases to 0.647 and 0.601 when the vocabulary is increased to 100K words.

Accuracy is tested for larger vocabularies of size 500K and 1M to assess improvement in performance. Accuracy with 1M words is given in Table 5.6. Interestingly, there is a slight decline in performance when such larger vocabularies are used. This is most likely due to the fact that an increased size of the vocabulary results in missing some matches for a test image. As a
Table 5.5: Image and configuration-based accuracy after assignment for BoV with vocabulary of size $|V| = 100K$ for different parameters of inverted file (averaged over five runs of random splits).

Table 5.6: Image and configuration-based accuracy after assignment for BoV with vocabulary of size $|V| = 1M$ for different parameters of inverted file (averaged over five runs of random splits).

### 5.3.2 BoV with min-Hash

The performance of the min-Hash algorithm to determine its suitability is assessed in this section. Similar to that done previously in Section 5.3.1, measures of absolute accuracy and accuracy after assignment are employed.

**Absolute accuracy:** For a vocabulary with 1K words, this is shown in Table 5.7. For the 12 min-Hash parameters, accuracy is close to 0.5 which is a representation of its proximity to random scoring. This reflects a decline in performance when compared with the inverted file structure for the same number of visual words. Moreover, it can seen that there are slight
differences in accuracy for the 12 parameters. Amongst the rankings, strong-5 mostly shows better performance while weak-5 displays the lowest accuracy.

Table 5.7: Image and configuration-based absolute accuracy for BoV with vocabulary of size $|V| = 1K$ for different parameters of min-Hash (averaged over five runs of random splits).

<table>
<thead>
<tr>
<th>(T,H)</th>
<th>strong-5</th>
<th>weak-5</th>
<th>all-10</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>im</td>
<td>contig</td>
<td>im</td>
</tr>
<tr>
<td>(1,1)</td>
<td>0.582</td>
<td>0.539</td>
<td>0.579</td>
</tr>
<tr>
<td>(1,2)</td>
<td>0.562</td>
<td>0.526</td>
<td>0.561</td>
</tr>
<tr>
<td>(1,3)</td>
<td>0.548</td>
<td>0.509</td>
<td>0.541</td>
</tr>
<tr>
<td>(5,1)</td>
<td>0.585</td>
<td>0.547</td>
<td>0.575</td>
</tr>
<tr>
<td>(5,2)</td>
<td>0.585</td>
<td>0.549</td>
<td>0.577</td>
</tr>
<tr>
<td>(5,3)</td>
<td>0.584</td>
<td>0.547</td>
<td>0.580</td>
</tr>
<tr>
<td>(25,1)</td>
<td>0.584</td>
<td>0.547</td>
<td>0.575</td>
</tr>
<tr>
<td>(25,2)</td>
<td>0.584</td>
<td>0.547</td>
<td>0.575</td>
</tr>
<tr>
<td>(25,3)</td>
<td>0.585</td>
<td>0.547</td>
<td>0.575</td>
</tr>
<tr>
<td>(100,1)</td>
<td>0.584</td>
<td>0.547</td>
<td>0.575</td>
</tr>
<tr>
<td>(100,2)</td>
<td>0.584</td>
<td>0.547</td>
<td>0.575</td>
</tr>
<tr>
<td>(100,3)</td>
<td>0.584</td>
<td>0.547</td>
<td>0.575</td>
</tr>
</tbody>
</table>

Larger vocabularies of size 10K, 50K and 100K are next evaluated. It was observed that typically accuracy reduces with increasing vocabulary size. Absolute accuracy for 100K words vocabulary is given in Table 5.8. Some interesting observations are made particularly when more than one hash function is used. An example would be when there are 3 hash functions for 1 hash table. Accuracy drops from close to 0.5 to 0 when the number of words are increased from 1K to 100K. This is because when the number of hash functions are increased, it boosts the discriminativeness of the algorithm but reduces the recall. Furthermore, when using larger vocabularies this lower recall becomes more apparent due to the increased number of visual words. As a result of which there are many test images that do not get matched to any images from the training set. This corresponds to incorrectly ranked pairs when computing the accuracy measure.

The vocabulary size is further increased to 500K and 1M. A similar observation to previous results is made where the accuracy drops when the vocabulary size is increased to 1M from 500K. Accuracy for the 1M words vocabulary is displayed in Table 5.9. Values for the 4 variations of number of hash tables of 1, 5; 25, 100 have dropped much below random scoring when using more than one hash function. This observation corresponds with the previous results where the accuracy drops when more hash functions are used with larger vocabularies. Furthermore, ac-
Table 5.8: Image and configuration-based absolute accuracy for BoV with vocabulary of size $|V| = 100K$ for different parameters of min-Hash (averaged over five runs of random splits).

Accuracy shows various zero-valued or nearly zero-valued outcomes. These are a consequence of missed matches for test images resulting in incorrect ranked-pairs when computing accuracy. For example, when using 1 hash table with 3 hash functions, from five random splits, two of the test sets of 524 images do not have any matches. While for the other two, 2 test images from 524 have a match and for the last, 3 test images are matched to an image.

Table 5.9: Image and configuration-based absolute accuracy for BoV with vocabulary of size $|V| = 1M$ for different parameters of min-Hash (averaged over five runs of random splits).

**Accuracy after assignment:** Testing is performed over the range of visual vocabulary sizes and several observations are made in this experiment. Firstly it was noted that not all test images are assigned a match after Munkres assignment algorithm has been applied. This is
5.3. Experimental results

a result of incomplete matches found for a test image. It occurs when no matches have been
found or only few of the matches from the 524 images in the training set are returned. The
number of missed matches generally increases as the size of the visual vocabulary is increased.
It should be recognized that in order to generate a $\frac{N}{2} \times \frac{N}{2}$ matrix for the Munkres assignment
algorithm with all the $i$th row and $j$th column elements, every test image must be matched
to a sorted list of 524 images in the training set. However, the min-Hash algorithm performs
sub-optimally and in some instances is not able to return the matches. As stated before missed
matches count as incorrect ranking when computing accuracy which causes the lower scoring.

Accuracy for vocabulary of size 100K is given in Table 5.10. The missed matches occur nor-
mally when there is more than one hash function for a given number of hash tables which was
also seen previously. For example, it can be seen from Table 5.10 when there are 3 hash func-
tions for 1, 5, 25 tables, accuracy is nearly zero. However, accuracy when one hash function is
used in these cases has a higher value. In particular, there are two main scenarios. Firstly, the
number of test images for which a match can not be found increases when the number of hash
functions increases with a constant number of hash tables. Secondly, the number of test images
with missed matches reduces as the number of hash tables is increased for a constant number
of hash functions. For example, the number of test images with missed matches increases from
125.4 to 516.2 when hash functions are increased from 1 to 3 with the use of a single hash ta-
ble. On the contrary, this number reduces from 516.2 to 209.8 when the number of hash tables
is increased from 1 to 100 with 3 hash functions. It should be noted that the total number of
test images is 524.

On using larger vocabularies with 500K and 1M words, similar observations to those made
before are seen. Accuracy when 1M words are used is shown in Table 5.11. The maximum
value is found to be close to 0.5. Moreover, such accuracy only occurs when higher number
of 25 and 100 hash tables are combined with a single hash function. When smaller number
of tables are combined with more hash functions the accuracy drops. Lower performance is
a result of the higher number of words being used which causes the min-Hash algorithm to
miss matches. From amongst the 12 parameter settings for min-Hash, a complete list of 524
matches were found for the 524 test images for only 2 of the parameters. These include the use
of 25 and 100 tables with a single hash function.
Chapter 5. Ranking images using matching

Table 5.10: Image and configuration-based accuracy after assignment for BoV with vocabulary of size $|V| = 100K$ for different parameters of min-Hash (averaged over five runs of random splits).

<table>
<thead>
<tr>
<th>(TH)</th>
<th>strong-5</th>
<th>weak-5</th>
<th>all-10</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>im</td>
<td>config</td>
<td>im</td>
</tr>
<tr>
<td>(1,1)</td>
<td>0.305</td>
<td>0.300</td>
<td>0.305</td>
</tr>
<tr>
<td>(1,2)</td>
<td>0.017</td>
<td>0.016</td>
<td>0.017</td>
</tr>
<tr>
<td>(1,3)</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
</tr>
<tr>
<td>(5,1)</td>
<td>0.555</td>
<td>0.533</td>
<td>0.545</td>
</tr>
<tr>
<td>(5,2)</td>
<td>0.202</td>
<td>0.188</td>
<td>0.195</td>
</tr>
<tr>
<td>(5,3)</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
</tr>
<tr>
<td>(25,1)</td>
<td>0.529</td>
<td>0.515</td>
<td>0.533</td>
</tr>
<tr>
<td>(25,2)</td>
<td>0.550</td>
<td>0.526</td>
<td>0.530</td>
</tr>
<tr>
<td>(25,3)</td>
<td>0.036</td>
<td>0.033</td>
<td>0.037</td>
</tr>
<tr>
<td>(100,1)</td>
<td>0.518</td>
<td>0.498</td>
<td>0.538</td>
</tr>
<tr>
<td>(100,2)</td>
<td>0.563</td>
<td>0.536</td>
<td>0.563</td>
</tr>
<tr>
<td>(100,3)</td>
<td>0.203</td>
<td>0.189</td>
<td>0.203</td>
</tr>
</tbody>
</table>

Table 5.11: Image and configuration-based accuracy after assignment for BoV with vocabulary of size $|V| = 1M$ for different parameters of min-Hash (averaged over five runs of random splits).

<table>
<thead>
<tr>
<th>(TH)</th>
<th>strong-5</th>
<th>weak-5</th>
<th>all-10</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>im</td>
<td>config</td>
<td>im</td>
</tr>
<tr>
<td>(1,1)</td>
<td>0.116</td>
<td>0.112</td>
<td>0.110</td>
</tr>
<tr>
<td>(1,2)</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
</tr>
<tr>
<td>(1,3)</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
</tr>
<tr>
<td>(5,1)</td>
<td>0.418</td>
<td>0.401</td>
<td>0.408</td>
</tr>
<tr>
<td>(5,2)</td>
<td>0.003</td>
<td>0.003</td>
<td>0.003</td>
</tr>
<tr>
<td>(5,3)</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
</tr>
<tr>
<td>(25,1)</td>
<td>0.570</td>
<td>0.547</td>
<td>0.559</td>
</tr>
<tr>
<td>(25,2)</td>
<td>0.047</td>
<td>0.042</td>
<td>0.045</td>
</tr>
<tr>
<td>(25,3)</td>
<td>0.001</td>
<td>0.001</td>
<td>0.001</td>
</tr>
<tr>
<td>(100,1)</td>
<td>0.551</td>
<td>0.534</td>
<td>0.540</td>
</tr>
<tr>
<td>(100,2)</td>
<td>0.205</td>
<td>0.192</td>
<td>0.200</td>
</tr>
<tr>
<td>(100,3)</td>
<td>0.009</td>
<td>0.008</td>
<td>0.009</td>
</tr>
</tbody>
</table>

5.3.3 Local Descriptor Matching (LDM)

In this section the LDM approach is evaluated using measures of accuracy and accuracy after assignment to determine its suitability for obtaining a ranking.

Absolute accuracy and accuracy after assignment for matching when using LDM are shown in Table 5.12. Values higher than baseline of 0.5 are obtained for the image-based and configuration-based ranking. This is true for strong-5, weak-5 and all-10 rankings. Highest absolute accuracy
of 0.694 and 0.624 for image and configuration-based ranking is seen for strong-5. These values are nearest to accuracy obtained with BoV representation when a higher number of 1M words were used. This involved setting 2 (none, l2, l2) with values of 0.689 and 0.625. The LDM matching uses full set of feature descriptors for matching by employing k-nearest neighbour search. That is why results here are similar to those achieved when using the larger visual vocabularies. Furthermore, it can be seen that highest accuracy is obtained with strong-5 while the lowest is found for weak-5. When comparing absolute accuracy with that observed after assignment, the former shows higher values. This is most likely due to every image being assigned an independent match. Incorrect matching can occur from Munkres algorithm for matching after assignment.

<table>
<thead>
<tr>
<th>measure</th>
<th>strong-5</th>
<th>weak-5</th>
<th>all-10</th>
</tr>
</thead>
<tbody>
<tr>
<td>abs acc</td>
<td>0.694</td>
<td>0.624</td>
<td>0.598</td>
</tr>
<tr>
<td>acc after assign</td>
<td>0.647</td>
<td>0.595</td>
<td>0.641</td>
</tr>
</tbody>
</table>

Table 5.12: Image and configuration-based absolute accuracy and accuracy after assignment averaged over five runs of random splits when matching using LDM.

5.3.4 Global ranking using LDM

In this experiment performance of global ranking using LDM is tested and analysed. Measures of Kendall’s τ and accuracy are utilized for performing the evaluations, this is because global rankings are generated here.

To demonstrate a full ranking with 120 configurations matching approach based on LDM is used. This is because according to evaluation performed LDM generated better accuracy compared to most other approaches evaluated i.e. BoV for smaller vocabularies with inverted file and BoV with the min-Hash data structure. Similar to previous testing voting files are constructed using rankings from strong-5, weak-5 and all-10.

Kendall’s τ and accuracy for the three scenarios is presented in Table 5.13. It can be seen that coefficients for Kendall’s τ are nearly 0. This is an indication that output ranking which takes matching into account is approximately independent of input ranking (strong-5, weak-5 and all-10). Furthermore, complementary observations can be made from accuracy, where value
is below 0.5. This represents that matching does not perform well and is unable to return true matches in most instances. Due to this appropriate comparisons are not made in the voting input to Kemeny-Young.

<table>
<thead>
<tr>
<th>measure</th>
<th>strong-5</th>
<th>weak-5</th>
<th>all-10</th>
</tr>
</thead>
<tbody>
<tr>
<td>Kendall’s $\tau$</td>
<td>-0.026</td>
<td>0.015</td>
<td>0.080</td>
</tr>
<tr>
<td>acc</td>
<td>0.360</td>
<td>0.365</td>
<td>0.353</td>
</tr>
</tbody>
</table>

Table 5.13: Kendall’s $\tau$ and accuracy averaged over five runs of random splits when matching using LDM.

### 5.3.5 Summary

An evaluation of image matching methods was performed using various measures including accuracy based on correctly ranked pairs and Kendall’s $\tau$. From the performance analysis, it was found that matching using LDM and BoV with inverted file (with a higher number of 1M visual words) provided better results compared to other approaches. This is as expected because LDM uses the full set of feature descriptors from the training set for nearest neighbour search. And in the BoV model, an image is represented by a histogram of occurrences of visual words. These visual words are part of a vocabulary which is built by clustering the features of the training images and so a higher number of visual words should generally improve the performance.

In particular, BoV with min-Hash was shown to give an inferior performance with slight improvements after the number of hash tables are increased. This needs to be done in conjunction with a minimum number of hash functions. The min-Hash algorithm works by using independently chosen visual words as global descriptors. A higher number of similar features in two images will lead to higher probability of them having the same min-Hash. However, in the context of the datasets used, the global nature of this algorithm is prohibitive because testing is based on image configurations. So this algorithm is more likely to be suitable for near duplicate image detection because differences between test and training images are minimum. Compared to the min-Hash algorithm, BoV with the inverted file data structure was shown to give better performance. An increment was particularly seen for larger vocabulary sizes of 500K and 1M.
On varying the parameters of the inverted file structure, such as, weighting of the histogram counts, normalization of the histograms and distance function, slight differences were seen in the performance.

Another interesting observation was made in the consistency of image-based and configuration-based absolute accuracy when evaluating. The highest values obtained for \textit{BoV} with inverted file and LDM were consistent for the image and configuration-based accuracy. These were moreover obtained for strong-5 ranking in all the cases, cf. highlighted accuracy in the tables of results. However, accuracy obtained for the min-Hash algorithm was not consistent between image and configuration-based measures. A constant high scoring for strong-5 was also not seen. This was most likely due to lower performance of the min-Hash algorithm where the accuracy obtained was mostly near 0.5 (random value). In summary, better results were obtained when using LDM, that is why it was used for generating a global ranking.

\section{5.4 Conclusions}

An evaluation of matching approaches to achieve ranking was performed and it was established that LDM and \textit{BoV} with inverted file (1M visual words) gave a better performance. Therefore matching from LDM was used to construct an overall configuration ranking. On testing this it was found that performance achieved was prohibitive in this context. This is most likely because although LDM generates better matches compared to other approaches, the quality of matching achieved is still not adequate. On comparing accuracy for global ranking from matching with that generated in Chapter 4 Section 4.3.2, matching based ranking shows weaker performance. For example, comparison of an accuracy of 0.94 for strong-5 with 10\% error for attribute recognition from Section 4.3.2 with a value of 0.360 for strong-5 in Section 5.3.4. Therefore, performance of ranking i.e. accuracy obtained from Table 5.13 will be used as a basis to compare with the approach proposed in Chapter 6.
Chapter 5. Ranking images using matching
Chapter 6

Ranking images by applying graphical modelling

In Chapter 3, the process of obtaining ranked lists of images from manual annotation based on comparative scoring was described. Rankings achieved from this can be utilized as reference sets for learning and evaluation. Given this data, a graphical model is trained to capture the attributes and relations between them. Using this model comparisons can be performed based on visual aesthetics when image pairs are presented at the input. The graph based approach alongside representation of the rankings using learning is first presented in Section 6.1. Next, an example based on the learning method is demonstrated in Section 6.2, which shows the attributes and their relation potentials for a toy example. Then, results from the experiments performed to test the graphical model and overall ranking which includes the recognition system estimates are discussed in Section 6.4. Finally, accuracy from the graphical modelling approach is compared with the baseline methods in Section 6.5.

6.1 Image model: learning and recognition

An overview of the approach for ranking images based on their aesthetic impression is first given. Next, procedure of how the global ranking is utilized to model various attributes and rating criteria is described. Finally, specifics of the model for usage within application of fashion assessment is explained.
Chapter 6. Ranking images by applying graphical modelling

6.1.1 Graph based model

The objective of the approach is to rank images based on the aesthetic impression they make. This can be simplified to producing an absolute rating where the approach is presented with a single image and generates a score within a normalized range of values. The automatic scoring method should be based on the same attributes and criteria that humans take into account when assessing an image. Building a model requires identifying the essential attributes as well as complex relations between them and then learning the weights with which they influence the score.

![Object representation model for modelling the ranked lists with 3-nodes each at a range of states.](image)

Figure 6.1: Object representation model for modelling the ranked lists with 3-nodes each at a range of states. In particular, this figure depicts two nodes for the clothing attributes of top \(T\), bottom \(B\) clothing and another for the body shape attribute \(S\) with the associated node \(\psi_i\) and edge \(\psi_w\) potentials.

To model the attributes and their relations the use of undirected graphical modelling (UGM) is proposed here [17]. There are two basic types of graphical models given by directed and undirected graphs. A node in a probabilistic graphical model represents a random variable and the edges between nodes give the probabilistic relationships between the variables. In a directed graphical model (DGM) edges between the nodes have a particular orientation given by arrows. On the contrary edges in an UGM do not have any directional significance and there are no arrows associated with them. An UGM is chosen because for the problem being addressed there is no natural directionality between the variables and so it is not straightforward to choose the direction of edges. Particularly in fashion aesthetics one cannot specify orientations between top clothing, bottom clothing and body shape. Therefore an UGM is more natural and suitable to represent the potentials of attributes as states of nodes here. Relations between the various attributes are represented by edges between the graph nodes. The states of each node and the relations between the states have certain potentials that contribute to the overall score of
aesthetic appearance. In the fashion assessment application, the nodes correspond to body parts and the states of the nodes correspond to clothing and body attributes. Figure 6.1 illustrates the model that is adopted, where edges between the states of the nodes represent relations between the attributes.

### 6.1.2 Learning image ranking

To facilitate the modelling and rating of images, the position in the ranking is considered as a joint potential of nodes being at given states. The higher the individual potentials of the states, the higher the position of their configuration in the global ranking. The probabilistic scenario where the dependencies within the graph involve \( N \) nodes is considered. The joint probability for this instance is represented using a model based on undirected graphical modelling. The overall rating of an image can therefore be expressed as a product of the attribute potentials and their relations that are present in the image. For all the nodes at states \( y_i \), this is given by the normalized product of non-negative potential functions \( \psi \) as:

\[
p(y_1, y_2, \ldots, y_N) = \frac{1}{Z} \prod_{i=1}^{N} \psi_i(y_i) \prod_{w=1}^{W} \psi_w(y_q, y_v),
\]

where potential function \( \psi_i \) is associated with node \( i \) and \( \psi_w \) is associated with edge \( w \) connecting nodes \( y_q \) and \( y_v \). This distribution is normalized with constant \( Z \) given by:

\[
Z = \sum_{y_1} \sum_{y_2} \cdots \sum_{y_N} \prod_{i=1}^{N} \psi_i(y_i) \prod_{w=1}^{W} \psi_w(y_q, y_v)
\]

### 6.1.3 Learning attributes potentials

The potentials for all the nodes and edges are estimated by learning the model from a training data. This training data is in the form of a ranked list of images that can be obtained by manual annotation. Providing objective ranking by manual annotation, in particular when there can be hundreds of possible configurations, is not straightforward. The process of obtaining such a ranking was described in Chapter 3 Section 3.5 and several rankings were generated in Chapter 3 Section 3.6.5. The process of estimation of the potentials is described below.

For a ranked list \( L \), which may include \( E \) examples of the same configuration of nodes at states \( y_1, y_2, \ldots, y_N \), the joint potential of this particular combination is represented as:

\[
p(y_1, y_2, \ldots, y_N) = \frac{1}{E} \sum_{i=1}^{E} p(y_{1i}, y_{2i}, \ldots, y_{Ni})
\]
where $p(y_1, y_2, \ldots, y_N)$ is a rating of an individual example at this particular configuration of states. This estimation of $p$ allows to accommodate for unbalanced datasets with different number of examples per configuration. Once this estimate for each unique configuration of node states has been obtained, it can be used to learn the node potentials $\psi_i$ and edge potentials $\psi_w$. For $\psi(y_1)$ an average over all configurations that include state $y_1 = z_1$ of node 1 is used as follows:

$$\psi(y_1 = z_1) = \sum_{y_2} \sum_{y_3} \cdots \sum_{y_N} p(y_1 = z_1, y_2, \ldots, y_N) \quad (6.4)$$

For edge potentials e.g. $\psi(y_1, y_2)$ states $y_1 = z_1$ and $y_2 = z_2$ are used:

$$\psi(y_1, y_2) = \sum_{y_3} \sum_{y_4} \cdots \sum_{y_N} p(y_1 = z_1, y_2 = z_2, y_3, \ldots, y_N) \quad (6.5)$$

### 6.1.4 Fashion aesthetics

In the application of aesthetic assessment a 3-node model with 4 states for the body shape, 5 for top and 6 for bottom clothing attributes is considered. The attributes are listed in Figure 6.3. One could add more nodes to represent shoes, jewellery, purse and other accessories as well as more states such as colour and texture. However, this requires a large training set where each state is included in various configurations of attributes. Furthermore, the study from [91] shows that colour has little impact on the overall dressing attractiveness. For example, in the specific case, the potential $\psi(S_s)$ for body shape at state $s$ is:

$$\psi(S_s) = \sum_t \sum_b p(S = s, T_t, B_b) \quad (6.6)$$

This means that to determine the node potential for body shape node $S$ at state $s$, i.e. $\psi(S_s)$, probabilities are marginalized over the top and bottom clothing nodes; an example for this is shown in Figure 6.2. Correspondingly the remaining node potentials $\psi(T_t)$ and $\psi(B_b)$ for top and bottom clothing attributes can be determined as:

$$\psi(T_t) = \sum_s \sum_b p(S_s, T = t, B_b) \quad (6.7)$$

$$\psi(B_b) = \sum_s \sum_t p(S_s, T_t, B = b) \quad (6.8)$$
Figure 6.2: An example for learning the node and edge potential from a ranking is shown in this figure. For estimating the potential for the body shape node $S$ at state 1 (i.e. $\psi(S_1)$), highlighted values in red i.e. $S_1T_1B_3$, $S_1T_2B_1$ and so on are utilized. And for estimating the edge potential between body shape node $S$ at state 2 and bottom clothing node $B$ at state 1 (i.e. $\psi(S_2, B_1)$), highlighted values in green i.e. $S_2T_1B_1$, $S_2T_2B_1$ and so on are employed. Rankings for this procedure are obtained using crowdsourced judgements as described in Chapter 3 Section 3.5; as an example a strong-5 or a weak-5 ranking can be utilized (cf. Chapter 3 Section 3.6.5).

The edge potential $\psi(S_s, B_b)$ between the body shape node $S$ and bottom clothing node $B$ at state $s$ and $b$ can be computed as follows:

$$\psi(S_s, B_b) = \sum_t p(S = s, T_t, B = b)$$ (6.9)

So to determine the edge potential of body shape node $S$ at state $s$ and bottom clothing node $B$ at state $b$, i.e. $\psi(S_s, B_b)$, probabilities are marginalized over the top clothing as illustrated in Figure 6.2. Edge potentials $\psi(S_s, T_t)$ and $\psi(T_t, B_b)$ can be calculated along similar lines:

$$\psi(S_s, T_t) = \sum_b p(S = s, T = t, B_b)$$ (6.10)

$$\psi(T_t, B_b) = \sum_s p(S_s, T = t, B = b)$$ (6.11)

### 6.2 Graph based model example

In this section a toy example for the graphical model learnt using the above equations is presented to illustrate its usage in practice. This model uses three nodes, one each for the body
shape, top clothing and bottom clothing that are shown in Figure 6.1. Each node has 3 states therefore there are a total number of \(3 \times 3 \times 3 = 27\) configurations possible. The breakdown of the states along with the initial product potential \(\psi_0(S_s, T_t, B_b)\) is shown in Table 6.1. The initial product potential which is used for the learning process represents the joint potential. Values for this are chosen arbitrarily in this toy example. In comparison, for the real-world scenario the initial product potential for the configurations is determined from ranking obtained using crowdsourcing procedure described in Chapter 3 Section 3.5. A higher value of the potential denotes a higher joint probability of the configuration and vice versa. This is an indication that the lowest preferred configuration in Table 6.1 occurs when state configuration \((s, t, b)\) is equal to \((1, 1, 1)\) and the best possible configuration is found at \((3, 3, 3)\).

<table>
<thead>
<tr>
<th>No.</th>
<th>Shape state (s)</th>
<th>Top state (t)</th>
<th>Bottom state (b)</th>
<th>Initial prod pot (\psi_0(S_s, T_t, B_b))</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>0.01</td>
</tr>
<tr>
<td>2</td>
<td>1</td>
<td>1</td>
<td>2</td>
<td>0.02</td>
</tr>
<tr>
<td>3</td>
<td>1</td>
<td>1</td>
<td>3</td>
<td>0.03</td>
</tr>
<tr>
<td>4</td>
<td>1</td>
<td>2</td>
<td>1</td>
<td>0.04</td>
</tr>
<tr>
<td>5</td>
<td>1</td>
<td>2</td>
<td>2</td>
<td>0.05</td>
</tr>
<tr>
<td>6</td>
<td>1</td>
<td>2</td>
<td>3</td>
<td>0.06</td>
</tr>
<tr>
<td>7</td>
<td>1</td>
<td>3</td>
<td>1</td>
<td>0.07</td>
</tr>
<tr>
<td>8</td>
<td>1</td>
<td>3</td>
<td>2</td>
<td>0.08</td>
</tr>
<tr>
<td>9</td>
<td>1</td>
<td>3</td>
<td>3</td>
<td>0.09</td>
</tr>
<tr>
<td>10</td>
<td>2</td>
<td>1</td>
<td>1</td>
<td>0.10</td>
</tr>
<tr>
<td>11</td>
<td>2</td>
<td>1</td>
<td>2</td>
<td>0.11</td>
</tr>
<tr>
<td>12</td>
<td>2</td>
<td>1</td>
<td>3</td>
<td>0.12</td>
</tr>
<tr>
<td>13</td>
<td>2</td>
<td>2</td>
<td>1</td>
<td>0.13</td>
</tr>
<tr>
<td>14</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>0.14</td>
</tr>
<tr>
<td>15</td>
<td>2</td>
<td>2</td>
<td>3</td>
<td>0.15</td>
</tr>
<tr>
<td>16</td>
<td>2</td>
<td>3</td>
<td>1</td>
<td>0.16</td>
</tr>
<tr>
<td>17</td>
<td>2</td>
<td>3</td>
<td>2</td>
<td>0.17</td>
</tr>
<tr>
<td>18</td>
<td>2</td>
<td>3</td>
<td>3</td>
<td>0.18</td>
</tr>
<tr>
<td>19</td>
<td>3</td>
<td>1</td>
<td>1</td>
<td>0.19</td>
</tr>
<tr>
<td>20</td>
<td>3</td>
<td>1</td>
<td>2</td>
<td>0.20</td>
</tr>
<tr>
<td>21</td>
<td>3</td>
<td>1</td>
<td>3</td>
<td>0.21</td>
</tr>
<tr>
<td>22</td>
<td>3</td>
<td>2</td>
<td>1</td>
<td>0.22</td>
</tr>
<tr>
<td>23</td>
<td>3</td>
<td>2</td>
<td>2</td>
<td>0.23</td>
</tr>
<tr>
<td>24</td>
<td>3</td>
<td>2</td>
<td>3</td>
<td>0.24</td>
</tr>
<tr>
<td>25</td>
<td>3</td>
<td>3</td>
<td>1</td>
<td>0.25</td>
</tr>
<tr>
<td>26</td>
<td>3</td>
<td>3</td>
<td>2</td>
<td>0.26</td>
</tr>
<tr>
<td>27</td>
<td>3</td>
<td>3</td>
<td>3</td>
<td>0.27</td>
</tr>
</tbody>
</table>

Table 6.1: Initial product potential for 3-node graphical model with 3 states per node resulting in 27 configurations.
6.2.1 Node potentials

In the first step node potentials for the shape, top and bottom clothing are calculated. Each calculation involves computation of the specific node at a specific state. So three sets of calculations for each node need to be performed. To normalize each potential a normalizing constant is calculated. The value for this can be found from Table 6.1 and is equal to 378. For the three nodes of shape, top clothing and bottom clothing the potentials $\psi(S_s)$, $\psi(T_t)$ and $\psi(B_b)$ are calculated using Equations 6.6, 6.7 and 6.8. The computations for these are shown below.

Shape node: $\psi(S_s)$

\[
\begin{align*}
\psi(S = s_1) &= \frac{1+2+3+4+5+6+7+8+9}{378} = 0.1190 \\
\psi(S = s_2) &= \frac{10+11+12+13+14+15+16+17+18}{378} = 0.3333 \\
\psi(S = s_3) &= \frac{19+20+21+22+23+24+25+26+27}{378} = 0.5476 
\end{align*}
\]

Top clothing node: $\psi(T_t)$

\[
\begin{align*}
\psi(T = t_1) &= \frac{1+2+3+10+11+12+19+20+21}{378} = 0.2619 \\
\psi(T = t_2) &= \frac{4+5+6+13+14+15+22+23+24}{378} = 0.3333 \\
\psi(T = t_3) &= \frac{7+8+9+16+17+18+25+26+27}{378} = 0.4048 
\end{align*}
\]

Bottom clothing node: $\psi(B_b)$

\[
\begin{align*}
\psi(B = b_1) &= \frac{1+4+7+10+13+16+19+22+25}{378} = 0.3095 \\
\psi(B = b_2) &= \frac{2+5+8+11+14+17+20+23+26}{378} = 0.3333 \\
\psi(B = b_3) &= \frac{3+6+9+12+15+18+21+24+27}{378} = 0.3571 
\end{align*}
\]

6.2.2 Edge potentials

Edge potentials for the three edges given by $\psi(S_s, T_t)$, $\psi(S_s, B_b)$ and $\psi(T_t, B_b)$ will now be calculated. For every edge, this makes use of the state of the nodes that the edge is connecting.

Edge 1: $\psi(S_s, T_t)$

This edge connects the shape node and top clothing node. So the computations for $\psi(S_s, T_t)$ will be based on the states of these nodes as shown in Equation 6.10.
\[ \psi(S = s_1, T = t_1) = \frac{1+2+3}{378} = 0.0159 \]
\[ \psi(S = s_1, T = t_2) = \frac{4+5+6}{378} = 0.0397 \]
\[ \psi(S = s_1, T = t_3) = \frac{7+8+9}{378} = 0.0635 \]
\[ \psi(S = s_2, T = t_1) = \frac{10+11+12}{378} = 0.0873 \]
\[ \psi(S = s_2, T = t_2) = \frac{13+14+15}{378} = 0.1111 \]
\[ \psi(S = s_2, T = t_3) = \frac{16+17+18}{378} = 0.1349 \]
\[ \psi(S = s_3, T = t_1) = \frac{19+20+21}{378} = 0.1587 \]
\[ \psi(S = s_3, T = t_2) = \frac{22+23+24}{378} = 0.1825 \]
\[ \psi(S = s_3, T = t_3) = \frac{25+26+27}{378} = 0.2063 \]

Edge 2: \( \psi(S_s, B_b) \)

Likewise, for \( \psi(S_s, B_b) \) states of the shape node and the bottom clothing node will be utilized as in Equation 6.9.

\[ \psi(S = s_1, B = b_1) = \frac{1+4+7}{378} = 0.0317 \]
\[ \psi(S = s_1, B = b_2) = \frac{2+5+8}{378} = 0.0397 \]
\[ \psi(S = s_1, B = b_3) = \frac{3+6+9}{378} = 0.0476 \]
\[ \psi(S = s_2, B = b_1) = \frac{10+13+16}{378} = 0.1032 \]
\[ \psi(S = s_2, B = b_2) = \frac{11+14+17}{378} = 0.1111 \]
\[ \psi(S = s_2, B = b_3) = \frac{12+15+18}{378} = 0.1190 \]
\[ \psi(S = s_3, B = b_1) = \frac{19+22+25}{378} = 0.1746 \]
\[ \psi(S = s_3, B = b_2) = \frac{20+23+26}{378} = 0.1825 \]
\[ \psi(S = s_3, B = b_3) = \frac{21+24+27}{378} = 0.1905 \]

Edge 3: \( \psi(T_t, B_b) \)

For \( \psi(T_t, B_b) \) the states of the top clothing node and the bottom clothing node are used as in Equation 6.11.
\[ \psi(T = t_1, B = b_1) = \frac{1+10+19}{378} = 0.0794 \]
\[ \psi(T = t_1, B = b_2) = \frac{2+11+20}{378} = 0.0873 \]
\[ \psi(T = t_1, B = b_3) = \frac{3+12+21}{378} = 0.0952 \]
\[ \psi(T = t_2, B = b_1) = \frac{4+13+22}{378} = 0.1032 \]
\[ \psi(T = t_2, B = b_2) = \frac{5+14+23}{378} = 0.1111 \]
\[ \psi(T = t_2, B = b_3) = \frac{6+15+24}{378} = 0.1190 \]
\[ \psi(T = t_3, B = b_1) = \frac{7+16+25}{378} = 0.1270 \]
\[ \psi(T = t_3, B = b_2) = \frac{8+17+26}{378} = 0.1349 \]
\[ \psi(T = t_3, B = b_3) = \frac{9+18+27}{378} = 0.1429 \]

### 6.2.3 Final product potential

The computations above can be used to generate the node and edge potential matrices as follows.

**Node potential matrices**

\[ \psi(S_s) = 0.1190 \quad 0.3333 \quad 0.5476 \]
\[ \psi(T_t) = 0.2619 \quad 0.3333 \quad 0.4048 \]
\[ \psi(B_b) = 0.3095 \quad 0.3333 \quad 0.3571 \]

**Edge potential matrix**

\[
\begin{array}{ccc}
0.0159 & 0.0397 & 0.0635 \\
0.0873 & 0.1111 & 0.1349 \\
0.1587 & 0.1825 & 0.2063 \\
0.0317 & 0.0397 & 0.0476 \\
0.1032 & 0.1111 & 0.1190 \\
0.1746 & 0.1825 & 0.1905 \\
0.0794 & 0.0873 & 0.0952 \\
0.1032 & 0.1111 & 0.1190 \\
0.1270 & 0.1349 & 0.1429
\end{array}
\]
It can be noted from the above matrices that the different states of a specific node and edge sum to 1 as expected. As example of this, consider the shape node where the different states \((0.1190 + 0.3333 + 0.5476)\) add to 1. Likewise, consider an edge-based example of the first edge that connects the shape and top clothing node \((0.0159 + 0.0397 + 0.0635 + 0.0873 + 0.1111 + 0.1349 + 0.1587 + 0.1825 + 0.2063)\) which also sums to 1.

The node and edge potentials can be used to calculate the final product potential \(\psi(S_s, T_t, B_b)\). For a given configuration in the state table, this is given by the product of the relevant potentials in that row. This product results in unnormalized potential for that configuration. The probability can be found by dividing this potential with the normalizing constant of 0.0027 (cf. Equation 6.2). An example for probability calculation for the configuration of states \((1, 1, 1)\) is presented below based on Equation 6.1. The complete set of values can be found in Table 6.2.

\[
p(S = s_1, T = t_1, B = b_1) = \frac{0.1190 \times 0.2619 \times 0.3095 \times 0.0159 \times 0.0317 \times 0.0794}{0.0027} = 0.0001
\]

To calculate the final product potential for real usage, rankings obtained from crowdsourcing are utilized within the graph-based model in the calculations above. This learns the attribute potentials which are the node potentials and attribute relation potentials which are the edge potentials. These can then be incorporated within Equation 6.1.

### 6.2.4 Correlation Coefficient

Final potentials can be arranged in order to achieve an output ranking. This ranking is compared with the input ranking to determine the correlation using Kendall’s \(\tau\) described previously in Chapter 3 Section 3.4.1. A correlation of 0.9943 is found for the toy example demonstrated above. It can be seen from Table 6.2 that the ranked ordering differs when the initial product potentials have a value of 18 and 19. The final potential calculated when the initial product potential is 18 is 0.048. This is in fact higher than 0.0360, which is the potential at 19. That is why correlation observed is lower than 1.000. In the real scenario Kendall’s \(\tau\) is calculated between rankings from the annotators and a modified version obtained after the learning procedure described in Section 6.1.4 has been performed.
6.3. Ranking based on automatic labelling

The recognition of attributes performed in Chapter 4 Section 4.1 is used for labelling attributes of clothing and body shape. For reference performance using recall and precision is summarized in Section 6.4.3. The error due to recognition has an impact on the performance of the graphical model and to evaluate this impact a controlled experiment is performed. In this the percentage of misclassified attributes in the classifiers is increased by a certain value and the previous misclassification is used with the next error. A detailed description of this process with the experimental results is given in Section 6.4.5 and Section 6.4.9.

Table 6.2: Final product potential for 3-node graphical model with 3 states per node to obtain the ranking based comparison.
6.4 Experimental results

In this section the performance of the representation model alongside the overall ranking approach is evaluated. The learning process for the representation requires a list of images that objectively ranks the various configurations of the attributes. These rankings were obtained previously using the Kemeny-Young method [66, 132] in Chapter 3 Section 3.5.

The rankings of images obtained from individual annotators and the representation model trained from these rankings is first assessed in Section 6.4.1. Next, global rankings from groups of annotators are utilized to train and test the approach, based on assumption that the states of all the nodes within the graphical model are known in Section 6.4.2. Then, performance of attribute recognition is briefly summarized in Section 6.4.3. This is followed by testing the approach assuming random predictive estimates for attributes in Section 6.4.4. Next, real recognition estimates are utilized for evaluation in Section 6.4.5. A different subdivision of data described later is employed for evaluation in the following experiments. For this data split, a correlation between different groups of rankings is first measured in Section 6.4.6. The attributes potentials and attribute relations potentials obtained from the representation model are then discussed in Section 6.4.7. Finally, rankings with known node states and from prediction estimates based on recognition are evaluated in Section 6.4.8 and Section 6.4.9.

**Data:** The annotations performed by the 10 annotators were validated using the procedure described in Chapter 3 Section 3.4.2. From the results it was established that the annotations could be divided into two groups. One containing the stronger annotations that met the criteria and the other that consisted of the weaker annotations that did not meet the criteria. To test the approach this data is utilized for the training and test rankings using different methods.

First the evaluation is performed using a complete group of either strong, weak or all the annotations; the groups were determined in Chapter 3 Section 3.6.5 following validation criteria. This is formed by including annotators 1, 2, 4, 6, 10 for the strong-5, 3, 5, 7, 8, 9 for the weak-5 and all the annotators for the all-10 ranking. Next, to validate the approach the 10 annotators are further split into four sub-groups: two strong and two weak ones, with two or three annotators in each group (strong-2, strong-3, weak-2, weak-3). This split is performed so that testing can be performed using sub-groups that were not used when training, but would still follow similar criteria. Such as, strong-2 tested when training using strong-3. In addition, random an-
notations for the paired-images presented to annotators are also generated. These give random rankings of rand-10, rand-s-2 and rand-w-2 in the all-10, strong-2 and weak-2 groups. Random rankings will serve as a baseline in the experiments.

**Kendall’s $\tau$ and accuracy measure:** These measures are calculated between the input and output rankings from the model and were described previously in Chapter 3 Section 3.4.1 and Chapter 4 Section 4.2.2 respectively.

### 6.4.1 Individual annotator ranking correlation

The performance of the approach is first measured for the annotated lists independently using the rankings obtained from each annotator. Alongside measuring how well the annotations are represented within the model, this can be used to investigate how consistently annotators perform the comparisons. For this experiment, rankings from Kemeny-Young are obtained over a range of $N_{try} = 10, 20, 40, 80, 160$. Each annotated list comprises of 7000 paired-image comparisons. Rankings from each of the 10 annotators are input into the representation model to measure Kendall’s $\tau$ between the input and output rankings. Testing on the random annotations is performed similarly to the real data.

<table>
<thead>
<tr>
<th>$N_{try}$</th>
<th>A1</th>
<th>A2</th>
<th>A3</th>
<th>A4</th>
<th>A5</th>
<th>A6</th>
<th>A7</th>
<th>A8</th>
<th>A9</th>
<th>A10</th>
</tr>
</thead>
<tbody>
<tr>
<td>10</td>
<td>$\tau$</td>
<td>0.81</td>
<td>0.78</td>
<td>0.68</td>
<td>0.73</td>
<td>0.78</td>
<td>0.74</td>
<td>0.66</td>
<td>0.66</td>
<td>0.78</td>
</tr>
<tr>
<td></td>
<td>Pref</td>
<td>6108</td>
<td>6188</td>
<td>5716</td>
<td>5978</td>
<td>6356</td>
<td>6424</td>
<td>6064</td>
<td>5996</td>
<td>5882</td>
</tr>
<tr>
<td>160</td>
<td>$\tau$</td>
<td>0.81</td>
<td>0.78</td>
<td>0.67</td>
<td>0.72</td>
<td>0.79</td>
<td>0.74</td>
<td>0.68</td>
<td>0.66</td>
<td>0.77</td>
</tr>
<tr>
<td></td>
<td>Pref</td>
<td>6116</td>
<td>6216</td>
<td>5750</td>
<td>6002</td>
<td>6362</td>
<td>6450</td>
<td>6082</td>
<td>6004</td>
<td>5910</td>
</tr>
</tbody>
</table>

Table 6.3: Kendall’s $\tau$ when different runs of the real-world and simulated datasets are ranked using $N_{try} = 10$ and 160 and input into the representation model.

Kendall’s $\tau$ along with preferences honoured when $N_{try} = 10, 160$ is shown in Table 6.3. An average value of 0.74 for Kendall’s $\tau$ is seen for the 10 real-world annotated lists when $N_{try} = 160$. It may not be possible to meet all the pairwise constraints in a global ranking for an annotator since the same configurations may be scored differently by the annotator. As expected the simulated annotations have a much lower mean of 0.41. An interesting observation is made where annotators 5, 9 have a high Kendall’s $\tau$ for this test even though they are part of the weak group of annotators. This could possibly be due to differences in their preferences when compared with the expert and high confidence annotators. However, this is not the case.
with other low confidence annotators $3, 7, 8$. This indicates a lack of consistency in their individual annotations. An average of 3955.8 preferences get honoured when $N_{try} = 160$ for the random annotations. This means that 56.51% preferences from 7000 are met. On the contrary, for the real annotations this value increases to 6084 which means that 86.91% preferences are met. However, differences in the measures over the range of $N_{ry}$ are small.

**6.4.2 Rankings from learnt attribute potentials**

In this section, performance of the approach is tested using grouped rankings all-10, strong-5, weak-5 and rand-10. This is based on the assumption that a perfect recognition is obtained for all the attributes. The testing is performed by measuring Kendall’s $\tau$ and accuracy between the input and output rankings. It was earlier observed that there were small differences when $N_{try}$ was increased to 160. To further test this on a larger scale, evaluations are also performed on higher $N_{ry}$ of 250, 500, 1000, 2000. This is particularly relevant because global rankings from annotation groups are based on a larger number of pairwise comparisons. Performance from this is used to determine the appropriate $N_{try}$ for the other experiments.

Kendall’s $\tau$ between rankings over the range of $N_{try}$ is presented in Table 6.4. As expected, Kendall’s $\tau$ is higher for the annotated lists compared to rand-10. The highest Kendall’s $\tau$ is seen for strong-5. This is as expected because this group represents the high confidence annotators who have been validated by the expert. So they show a high degree of agreement with the expert. Lower values are seen for rand-10 since this ranking group possibly has a high degree of disagreement. This is due to the pairwise comparisons generated at random.

<table>
<thead>
<tr>
<th>Ann</th>
<th>$\tau$</th>
<th>$\tau$</th>
<th>$\tau$</th>
<th>$\tau$</th>
<th>$\tau$</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>160</td>
<td>250</td>
<td>500</td>
<td>1000</td>
<td>2000</td>
</tr>
<tr>
<td>all-10</td>
<td>0.74</td>
<td>0.74</td>
<td>0.74</td>
<td>0.74</td>
<td>0.74</td>
</tr>
<tr>
<td>strong-5</td>
<td>0.78</td>
<td>0.77</td>
<td>0.77</td>
<td>0.76</td>
<td>0.78</td>
</tr>
<tr>
<td>weak-5</td>
<td>0.70</td>
<td>0.70</td>
<td>0.70</td>
<td>0.70</td>
<td>0.70</td>
</tr>
<tr>
<td>rand-10</td>
<td>0.42</td>
<td>0.43</td>
<td>0.43</td>
<td>0.45</td>
<td>0.47</td>
</tr>
</tbody>
</table>

Table 6.4: Kendall’s $\tau$ for the representation model when different runs of the real-world and simulated datasets are ranked using larger values of $N_{try}$ to be used as input.

The accuracy of ranked pairs when $N_{try} = 160, 250, 500, 1000, 2000$ is presented in Table 6.5. The overall accuracy over the range of $N_{try}$ is marginally different from each other. From the results, it can be seen that highest accuracy is obtained for strong-5 and the lowest
for rand-10 (random chance score is 0.5). This result is as expected and is consistent with the correlation from Kendall’s \( \tau \). An accuracy of 0.89 is achieved for strong-5 when \( N_{try} = 2000 \) which is the highest value. It shows that 89\% image pairs were correctly ranked.

<table>
<thead>
<tr>
<th>Ntry</th>
<th>160</th>
<th>250</th>
<th>500</th>
<th>1000</th>
<th>2000</th>
</tr>
</thead>
<tbody>
<tr>
<td>all-10</td>
<td>0.87</td>
<td>0.87</td>
<td>0.87</td>
<td>0.87</td>
<td>0.87</td>
</tr>
<tr>
<td>strong-5</td>
<td>0.89</td>
<td>0.88</td>
<td>0.89</td>
<td>0.88</td>
<td>0.89</td>
</tr>
<tr>
<td>weak-5</td>
<td>0.85</td>
<td>0.85</td>
<td>0.85</td>
<td>0.85</td>
<td>0.85</td>
</tr>
<tr>
<td>rand-10</td>
<td>0.71</td>
<td>0.71</td>
<td>0.71</td>
<td>0.72</td>
<td>0.74</td>
</tr>
</tbody>
</table>

Table 6.5: Accuracy of ranked pairs of images for the representation model following perfect attribute estimates, i.e., with known node states for all the 15 influencing categories on increasing \( N_{try} \).

Overall, it can be seen that weak-5 performs worse than all-10 and strong-5 as expected. In addition, it can be seen that there are marginal differences in the correlation between the different values of \( N_{try} \). Due to this the higher \( N_{try} = 2000 \) will be utilized in performing the ranking using the Kemeny-Young method.

### 6.4.3 Attribute recognition

Performance using recall and precision is shown in Figure 6.3 for the maximum response of a set of classifiers that are part of the same region (i.e., upper, middle and lower body regions for reference). A more in-depth account of the performance is given in Chapter 4 Section 4.3.1. Overall, it can be seen from Figure 6.3 that high recall and precision is obtained for most of the attributes.

![Figure 6.3: Performance of attribute recognition using recall and precision averaged over five runs of random splits.](attachment:image.png)
6.4.4 Rankings with random predictive estimates of attributes

The recognition of the clothing and body shape attributes have an impact on how well the overall ranking works. In this section the impact is determined by performing a controlled experiment. The percentage of misclassified attributes for a category of attributes (top clothing, bottom clothing or body shape) is increased by a constant value for every consequent test. While the remaining two categories are assumed to have perfect recognition rate. The constant value for the error is chosen as the random prediction estimate for the attributes. As an example, there are 5 categories of top clothing so the random recognition rate for this is 20% (used as the error). Previous misclassification is kept and used with added error for the next test. So for the top clothing this would lead to an increase in the error by 20%. For every test, the performance is estimated by comparing the paired-configurations at the input and output using the accuracy measure.

<table>
<thead>
<tr>
<th>Error %</th>
<th>all-10</th>
<th>strong-5</th>
<th>weak-5</th>
<th>rand-10</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>0.87</td>
<td>0.89</td>
<td>0.85</td>
<td>0.74</td>
</tr>
<tr>
<td>top</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>20</td>
<td>0.81</td>
<td>0.82</td>
<td>0.81</td>
<td>0.70</td>
</tr>
<tr>
<td>40</td>
<td>0.76</td>
<td>0.76</td>
<td>0.76</td>
<td>0.66</td>
</tr>
<tr>
<td>80</td>
<td>0.64</td>
<td>0.62</td>
<td>0.67</td>
<td>0.57</td>
</tr>
<tr>
<td>100</td>
<td>0.59</td>
<td>0.57</td>
<td>0.63</td>
<td>0.53</td>
</tr>
<tr>
<td>bottom</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>17</td>
<td>0.85</td>
<td>0.88</td>
<td>0.82</td>
<td>0.71</td>
</tr>
<tr>
<td>34</td>
<td>0.82</td>
<td>0.86</td>
<td>0.79</td>
<td>0.68</td>
</tr>
<tr>
<td>85</td>
<td>0.76</td>
<td>0.81</td>
<td>0.71</td>
<td>0.60</td>
</tr>
<tr>
<td>100</td>
<td>0.74</td>
<td>0.80</td>
<td>0.68</td>
<td>0.57</td>
</tr>
<tr>
<td>bshape</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>25</td>
<td>0.81</td>
<td>0.82</td>
<td>0.79</td>
<td>0.70</td>
</tr>
<tr>
<td>50</td>
<td>0.75</td>
<td>0.76</td>
<td>0.72</td>
<td>0.65</td>
</tr>
<tr>
<td>75</td>
<td>0.69</td>
<td>0.70</td>
<td>0.65</td>
<td>0.61</td>
</tr>
<tr>
<td>100</td>
<td>0.62</td>
<td>0.64</td>
<td>0.58</td>
<td>0.57</td>
</tr>
</tbody>
</table>

Table 6.6: Accuracy of ranked pairs of images using the representation model following attribute estimates with 20%, 17% and 25% error for top clothing, bottom clothing and body shape attributes with increasing % error.

The accuracy when performance is estimated based on random error rates for an attribute are shown in Table 6.6. Several observations can be made from this table. Firstly, it can be seen that the rate of decline in performance is lower than the error included. For example, for top clothing at an error of 20% the accuracy reduces from 0.87 to 0.81 for all-10. This means that 81% image pairs are correctly ranked. Another interesting observation is that the top clothing category is affected more in terms of the decline in performance when compared with bottom clothing and body shape. As an example, for strong-5 at 100% error the accuracy for top
clothing drops to 0.57. While that for bottom clothing and body shape drops to 0.80 and 0.64. This observation varies slightly for weak-5 since the annotators in this group apply different criteria. A uniform rate of decline is observed when rand-10 is used. This is as expected since no annotators were used when generating these rankings. In summary, some interesting observations were seen, such as, a higher impact of the top clothing as compared with bottom clothing for strong-5. This is in contrary to weak-5 where the highest impact was seen for the body shape category.

### 6.4.5 Rankings with attribute recognition

The performance of attribute recognition was evaluated in Section 6.4.3. It was determined that the error for most of the attributes was below 10%. This error is utilized to measure the overall impact of errors due to attribute recognition. The experiment is performed similarly to that done in the previous Section 6.4.4, but using different error percentages for the attributes. The percentage of the misclassified attributes is increased at a constant error rate of 10%. Error is induced in all the individual classifiers for the 15 influencing attributes. This is in contrast to the previous experiment where the error was included in one main category, e.g., top clothing and the other two were assumed to have perfect recognition. The performance is measured using the accuracy measure between the input and output rankings.

<table>
<thead>
<tr>
<th>Ann</th>
<th>0%</th>
<th>10%</th>
<th>40%</th>
<th>70%</th>
</tr>
</thead>
<tbody>
<tr>
<td>all-10</td>
<td>0.87</td>
<td>0.83</td>
<td>0.69</td>
<td>0.57</td>
</tr>
<tr>
<td>strong-5</td>
<td>0.89</td>
<td>0.84</td>
<td>0.70</td>
<td>0.56</td>
</tr>
<tr>
<td>weak-5</td>
<td>0.85</td>
<td>0.81</td>
<td>0.69</td>
<td>0.57</td>
</tr>
<tr>
<td>rand-10</td>
<td>0.74</td>
<td>0.71</td>
<td>0.64</td>
<td>0.56</td>
</tr>
</tbody>
</table>

Table 6.7: Accuracy of ranked pairs of images using the representation model following attribute estimates with 10% error and increasing % error for each of the 15 influencing categories.

The results for the different groups of annotators are given in Table 6.7. Similar to the previous experiment, rate of decline in performance is lower than the error introduced in the individual classifiers. A high accuracy of 0.84 is seen for strong-5 at 10% error which means that a high number of 84% pairs are correctly ranked. Moreover, performance obtained compares well to the baseline of automatic assessment based on lookup model evaluated in Chapter 4 Section 4.3.2. This had an accuracy of 0.94 for strong-5 with 10% error in attribute recognition.
In summary, it is found that the overall ranking performs well even when a realistic error is introduced in all the influencing categories.

### 6.4.6 Rankings correlations

After evaluation of the performance of rankings obtained using groups of annotators, the focus of analysis will now be on the sub-groups of annotators. In this section correlation between rankings that resulted from different sub-groups of annotators is measured. By using Kendall’s $\tau$ measure, every pair of groups i.e. strong-2, strong-3, weak-2, weak-3 is evaluated.

The correlations between groups is displayed in Figure 6.4. Diagonal self-correlations were removed to reduce clutter. As expected strong groups have better correlation than weak groups. The random rankings have significantly lower scores as they are not correlated with any of the annotator groups. Interestingly, there is a slight correlation between strong-2 and rand-s-2. This may be due to the fact that these two global rankings were generated from the same subset of pairwise scored images. Any random correlation would have been amplified when generating global rankings. In summary, the agreements amongst strong as well as weak annotators indicate that they apply common fashion criteria and that it should be possible to automatically learn these criteria from the provided rankings.

![Figure 6.4: Kendall’s $\tau$ correlation between strong, weak and random annotators. The values vary from 0 to 0.45.](image)

### 6.4.7 Attributes potentials

The rankings generated with Kemeny-Young method are used to learn the attribute and relations potentials in the graph model as discussed in Section 6.1. In this section, attributes potentials, i.e., which states of the nodes are preferred and relations potentials, i.e., the states of different nodes that complement each other are investigated. To better visualise the learnt
potentials the corresponding potential from a random ranking is subtracted from each estimated potential. Thus, negative potentials indicate lower than random influence of a body shape or a clothing part on the overall rating of the image.

**Attributes potentials:** The node potentials for different attributes are shown in Figure 6.5. It can be seen that apple body shape, loose jacket and straight skirt have the lowest potentials. On the contrary column body shape and flared trousers show higher potentials. This indicates as an example a preference for flared trousers in comparison to straight skirt.

A higher potential for column body shape, fitted and ruffled tops and flared trousers is seen for the stronger group of annotators strong-2. In contrast the weaker group weak-3 show a higher potential for pear body shape, loose top and fitted trousers. This shows the variation of the criteria applied by the different groups of annotators when performing pairwise comparisons. However, there are also some instances which do not show this behaviour. Such as straight skirt which has low potential for both the groups of annotators.

![Figure 6.5: Attributes potentials learnt from rankings: strong-2 w.r.t. rand-s-2 and weak-3 w.r.t. rand-s-2.](image)

For the attribute relation potential between body shape and top clothing, it can be seen that a stronger positive or negative impact is observed for strong-2 when compared to weak-3. Another interesting observation is the preference for fitted top by strong-2 as opposed to loose top by weak-3 for apple body shape. Due to the contrasting shape of loose and fitted top, it is strengthened that these two groups of annotators apply different criteria when making judgements.
Chapter 6. Ranking images by applying graphical modelling

Figure 6.6: Attribute relations potentials learnt from rankings: (left) strong-2 w.r.t rand-s-2 and (right) weak-3 w.r.t rand-s-2. Top: between body shape and top clothing, Middle: between body shape and bottom clothing and Bottom: between top clothing and bottom clothing.

The highest positive impact between the shape and bottom clothing nodes is seen in the preference for flared trousers for column body shape by strong-2. On the contrary for weak-3 the most preferred bottom clothing for column body shape is observed to be that of fitted trousers. These two judgements are also quite contradicting since the shape for these two clothing is different. An intriguing observation is seen when an apple body shape wears a straight skirt.
This combination is not recommended by either strong-2 or weak-3 showing the alignment of the assessments made for this scenario.

The edge potential between the top and bottom clothing also show some interesting observations. In general, the strong-2 group preferred fitted jacket with most bottom clothing. While weak-3 show a preference for loose jacket for different bottom attributes. Particularly, strong-5 recommends wearing flared trousers with fitted jacket and weak-3 would like fitted or straight trousers with loose jacket. A high negative impact is observed for both strong-2 and weak-3 when loose jacket is worn with straight skirt.

In summary, the above observations show that the attributes potentials and attribute relations potentials support each other to establish which configurations work well together. As an example, for strong-2 group apple body shape, loose jacket and straight skirt have the lowest attribute potential. Similar observation was made for edge potentials for these three attributes. Apple body shape with loose jacket or straight skirt is not preferred by strong-2. In addition, a high negative impact is also seen when loose jacket is combined with straight skirt for this group.

### 6.4.8 Ranking based on learnt attribute potentials

In order to validate the model it is trained on image ranking that resulted from one group of annotators and tested on another one. This is measured with the average accuracy of pairwise preference ratings of images. In this experiment it is assumed that all the states of the nodes, that is the attributes present in the image are known (also done in Section 6.4.2). In this way the disagreements between training and test are only due to the limitations of the proposed model and differences between test and training data. It should be noted again that not all pairwise constraints given by the annotators can be satisfied in one global ranking. Some of them may contradict each other i.e. same configurations can be scored differently by different annotators or even by the same annotator.

Table 6.8 shows the percentage of pairs correctly ranked for different training and testing sets. The highest score is obtained when trained and tested on rankings validated by expert i.e. strong-2 and strong-3. The score of 0.91 indicates that the level of contradictions within the pairwise rankings is low (< 10%). These results also show that the model captures the annotation criteria very well and reflects the ranking of image pairs with high accuracy. It is
Chapter 6. Ranking images by applying graphical modelling

Table 6.8: Accuracy of ranked pairs of images using the representation model and assuming the node states are known.

<table>
<thead>
<tr>
<th>test \train</th>
<th>strong-2</th>
<th>strong-3</th>
<th>weak-2</th>
<th>weak-3</th>
<th>rand-s-2</th>
</tr>
</thead>
<tbody>
<tr>
<td>strong-2</td>
<td>0.91</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>strong-3</td>
<td>0.76</td>
<td>0.87</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>weak-2</td>
<td>0.75</td>
<td>0.76</td>
<td>0.88</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>weak-3</td>
<td>0.62</td>
<td>0.66</td>
<td>0.84</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>rand-s-2</td>
<td>0.58</td>
<td>0.59</td>
<td>0.58</td>
<td>0.55</td>
<td>0.72</td>
</tr>
<tr>
<td>rand-w-2</td>
<td>0.55</td>
<td>0.58</td>
<td>0.54</td>
<td>0.52</td>
<td>0.56</td>
</tr>
</tbody>
</table>

observed that the results gradually decrease when training and testing on weak sets with the lowest results for randomly generated rankings.

6.4.9 Image ranking with attribute recognition

The attribute recognition error has an impact on the performance of the entire ranking system. To assess this impact with the sub-group of annotators a controlled experiment is performed. The percentage of the misclassified attributes in the individual classifiers is increased by a constant value for every consequent test (cf. Section 6.4.5). For every test, the performance is estimated by comparing the training and testing paired-configurations using accuracy measure.

Table 6.9: Accuracy of ranked pairs of images using the representation model and attribute recognition with 10% error and increasing % error for each of the 15 influencing categories.

<table>
<thead>
<tr>
<th>train/test</th>
<th>0%</th>
<th>10%</th>
<th>40%</th>
<th>70%</th>
</tr>
</thead>
<tbody>
<tr>
<td>strong-2/strong-3</td>
<td>0.76</td>
<td>0.73</td>
<td>0.63</td>
<td>0.54</td>
</tr>
<tr>
<td>weak-2/weak-3</td>
<td>0.66</td>
<td>0.64</td>
<td>0.58</td>
<td>0.53</td>
</tr>
<tr>
<td>strong-2/rand-s-2</td>
<td>0.58</td>
<td>0.57</td>
<td>0.53</td>
<td>0.51</td>
</tr>
<tr>
<td>weak-2/rand-w-2</td>
<td>0.54</td>
<td>0.53</td>
<td>0.51</td>
<td>0.50</td>
</tr>
</tbody>
</table>

The results are presented in Table 6.9. Several observations are made from these results. The performance is only slightly lower compared to results with no error in attribute classification. Moreover, the rate of decline is lower than the actual error induced. For example, for the strong data, the performance decline is from 0.76 to 0.73 at 10% attribute recognition error. That is 73% of image pairs were correctly ranked. The classification error is far below 10% for most attributes except body shape. Even better performance can be achieved in certain application scenarios e.g. no pose or viewpoint variations in front of a mirror.
6.5 Comparison with baselines

The performance obtained using the graphical model approach is far superior to that achieved from matching in Chapter 5. Specifically, an accuracy of 0.36 was observed for matching in Section 5.3.4. This means that 36% image pairs were ranked correctly. In comparison an accuracy of 0.73 is obtained for strong-2/strong-3 at a realistic error of 10% for recognition using the graph based model; this means that 73% image pairs were ranked correctly. This also compares well against the lookup model baseline in Chapter 4 Section 4.3.2 in which 94% image pairs were ranked correctly for strong-5 ranking at 10% error for recognition.

6.6 Conclusions

In this chapter an effective approach for learning the ranking of images using qualitative assessments of visual aesthetics was introduced. A graph based representation where node and edge potentials captured the importance of visual attributes and their relations was proposed. The overall score was based on the potentials of these nodes and edges. This was represented by the product of the attribute and edge potentials; higher potentials have a higher position in the global ranking. A method for learning the model from pairwise preference scores via global ranking of images was also presented. This model was learnt by generating estimates of all the node and edge potentials from training data. Training data was in the form of a ranked list which was obtained previously in Chapter 3 Section 3.6.5. To rank an image based on attributes potentials, the attributes present within an image need to be identified; this was performed as in Chapter 4 Section 4.1. The effectiveness of the approach was demonstrated on a collection of fashion images that included different combinations of clothing and body shapes. From the assessments it was found that the model performed very well and was able to effectively capture the criteria for the attributes and their relations. This was also true when realistic prediction estimates for the attributes were incorporated to test the overall ranking. Moreover, it also compared well with the lookup model baseline in Chapter 4 Section 4.3.2 and very positively against the ranking from matching performed in Chapter 5 Section 5.3.4, where more than 100% improvement in performance was observed.
Chapter 7

Discussion

In this chapter contributions made within this thesis are discussed; firstly in Section 7.1 a summary of the problems addressed is given. Specific contributions made within chapters with findings from the work performed are then highlighted in Section 7.1.1, Section 7.1.2, Section 7.1.3 and Section 7.1.4. Finally, in Section 7.2 opportunities for future work are discussed.

7.1 Summary

The goal was to obtain an ordered list of image configurations based on aesthetics. An image configuration is given by states of clothing and body shape in an image. Utilizing such a ranking, pairwise comparison of images that belong to different configurations based on certain criteria can be performed. The work is motivated by the limited amount of research related to the study of aesthetic qualities, particularly in the context of fashion interpretation. However, aesthetics and fashion represent a significant driver in the economy outside the interest of academic research.

Rankings were modelled based on judgements from crowdsourced mid-level annotators and an expert. The judgements provided by annotators were filtered to determine inconsistencies so that these could be ruled out in Chapter 3. Based on this weaker and stronger group of annotations were established in order to generate rankings for these groups. For detecting attributes of clothing and body shapes automatically recognition was performed in Chapter 4. Objective
rankings from the annotators were incorporated within several matching approaches to establish which test image was ranked better from a test pair in Chapter 5. Finally an approach that learnt an image representation based on graphical modelling was proposed in Chapter 6. The contributions within these approaches are discussed below.

7.1.1 Crowdsourced judgements

An approach for obtaining an objective measure using annotators was proposed. For rating, images were presented in the form of pairwise comparisons from a dataset. This dataset was introduced as part of the work and provided appropriate images for performing comparisons. Specifically the dataset consisted of 1064 images which belonged to 120 configurations given by states of body shape and clothing. A large number of 57400 comparisons were performed by 10 annotators so that ranked lists could be obtained. Utilizing these comparisons several ranked lists such as those based on annotators who exhibited a high degree of agreement and correlation with the expert or a ranking that was inconsistent with the expert were generated. This was done by employing a ranking method based on Kemeny-Young preference aggregation [66, 132, 105].

The consistency checking using the measures of agreement and correlation was shown to effectively isolate 5 weaker annotations. Particularly following the verification annotators were split into two groups with annotators 1, 2, 4, 6, 10 representing a stronger group and annotators 3, 5, 7, 8, 9 which were part of the weaker group. This was based on observations made from annotator vs. expert agreement and the inter-annotator agreement. It was displayed that annotators who showed a large degree of correlation with the expert were also correlated with each other. On the contrary, weaker performing annotators showed agreement with other low confidence annotators. As an example annotator 1 showed a high accuracy of 0.73 with the expert which represents an agreement of 73%. A high value of 0.69 for accuracy was also seen for annotator 2 for this agreement. These two annotators also displayed the highest level of inter annotator accuracy of 0.73. This showed that these annotators applied similar criteria in scoring fashion images. The judgements made by the annotators were represented in the form of objective rankings based on the division of the annotations into a stronger and weaker group using the Kemeny-Young method. Several rankings were generated such as strong-5
from stronger group of annotators, weak-5 by employing weaker group of annotators and all-
utilizing all the 10 annotations. This was performed both in image space, i.e., ranking of the
1064 images part of the complete dataset and attribute space, i.e., a compressed ranking of 120
configurations obtained by averaging the rank of images that belong to the same configuration.

### 7.1.2 Attribute recognition

For detecting attributes of clothing and body shapes automatically recognition was performed
using the BoV approach [69]. This performed well for the 11 clothing categories, however, ex-
hibited a weaker performance for the body shape attributes. A high average recall and precision
of 0.93 ± 0.05 was obtained for the 11 clothing categories and a lower average of 0.30 ± 0.08
was observed for the body shape. Since the average error for most of the attributes is below
10% recognition performs well compared to state-of-the-art visual classification systems [3].
Mean prediction estimates from the better performing attributes with and without the body
shape estimates were incorporated within the evaluation framework.

Objective rankings obtained from crowdsourced judgements with the prediction estimates for
the attributes were incorporated within a lookup model for automatic assessment. Ranking
in this lookup model is based on judgements of the crowdsourced annotators and the error is
solely due to recognition of clothing and body shapes automatically. Therefore such a lookup
model is used to establish a baseline for the performance of ranking approaches proposed in the
remaining work in this thesis. When testing the lookup model, effectiveness of the model was
observed with a consistency in performance. Rankings from stronger and all of the annotation
groups performed better compared to weaker and random annotations. In addition a random
ranking was formed based on the pairwise comparisons for 10 annotations being performed
randomly. Such a rand-10 ranking was utilized as a baseline for comparative purposes so
testing could also be performed on a simulated dataset. The strong-5 ranking produced scores
of 0.73, 0.96, 0.94, 0.87 for the actual and error estimates of 7%, 10% and 23%. These are
much higher than the scores of 0.61, 0.72, 0.71, 0.67 which were obtained for weak-5. These
results are as expected because the stronger and the weaker group apply different criteria when
performing the pairwise comparisons. Specifically accuracy of 0.94, 0.83, 0.71, 0.54 obtained
for strong-5, all-10, weak-5 and rand-10 at 10% error estimate will serve as a baseline.
7.1.3 Matching based evaluation

A ranking of images obtained from annotators can be incorporated directly within a matching approach to obtain nearest neighbour matches. This was done by applying a matching approach to a pair of test images in which an image represents a configuration of clothing and body shape. Ordering of the matches obtained were compared with their ranking obtained from crowdsourcing such as strong-5, weak-5 and all-10 to check which configuration was ranked better. An evaluation using the approaches of Bag of Visual Words (BoV) and Local Descriptor Matching (LDM) was performed [114, 72]. For the BoV representation, inverted file and min-Hash data structures were employed and parameter variation was performed at the vocabulary and data structure level [8, 7]. A highest accuracy of 0.687 and 0.627 (image and configuration) was obtained when using strong-5 ranking with a visual vocabulary of size 100K on employing the inverted file data structure. On increasing the vocabulary to 1M visual words marginal variation was seen with values of 0.689 and 0.625. This indicated that increasing the vocabulary size did not display much improvement in matching. Similar observations were made when using different parameter settings for the inverted file representation. A decrease in accuracy was observed with the use of min-Hash data structure with highest values of 0.580 and 0.529 for 100K vocabulary. Moreover the highest values were obtained when a minimum number of hash functions were combined with more hash tables. This result is as expected since an increase in the number of hash tables increases the recall of the min-Hash algorithm. Next the LDM implementation was used as provided in [72] and showed marginal variation compared to matching with inverted file. Specifically accuracy of 0.694 and 0.624 was observed when using strong-5 ranking. Overall due to lower accuracy achieved with different matching approaches, its usage in ranking images based on aesthetics was shown to be prohibitive.

A procedure for obtaining a global ranking based on the matches was also proposed. This used the Munkres assignment algorithm to obtain a unique list of matches which were compared with the rankings obtained from the annotators. A voting file was then generated which was input to the Kemeny-Young method for a ranking. To test the global ranking, LDM was utilized and an accuracy of 0.360 was observed. This lower than random value of accuracy indicates that the global ranking generated was not able to capture the judgements of the annotators effectively. Due to the limiting results displayed by matching approaches, accuracy of 0.360
will be used as a baseline.

### 7.1.4 Graphical modelling

For generating a ranking based on aesthetics matching was not shown to perform well, therefore a different approach based on graphical modelling was proposed. This approach learnt image ranking and captured the users perspective of what they deemed as being aesthetically pleasing. The learnt image representation provides a more effective method for modelling the attributes of clothing and body shapes and their relations. Using a graph based model in which the states of the attributes were represented by the nodes and the relations between the attributes as edges, an overall score for a configuration was calculated. Rankings from crowdsourced annotators generated previously in Chapter 3 and utilized for ranking in Chapter 4 and Chapter 5 provided the training data for the model. This was used in the learning process to estimate all the node and edge potentials. The overall ranking is based on the learnt node and edge potentials in which a higher potential indicates a more favourable position in the global ranking. To identify the attributes within an image, attribute recognition system evaluated previously in Chapter 4 was used. The methods were extensively evaluated based on two main scenarios with the use of several groups of strong, weak, all and random rankings, i.e. strong-2, strong-3, weak-2, weak-3, rand-s-2 and rand-w-2. These rankings were obtained previously in Chapter 3 after the validation procedure. The first approach for evaluation was based on the instance when node states were known. This was done to test the model when error would be a result of difference between training and test set and due to limitations of the proposed model and not due to detecting attributes automatically. A high accuracy of 0.76 was observed when the graphical model was trained using strong-3 and tested with strong-2. This means that a high level of 76% agreement was obtained using the learning approach and different train/test rankings. It is also an indication that stronger group of annotators apply similar criteria. On the contrary much lower value of 0.52 was obtained for rand-w-2/weak-3. This is as expected and an indicative of a gradual decrease of accuracy from stronger to weaker groups with the lowest results for randomly generated rankings. Next prediction estimates from automatic recognition were included in the representation model in order to include error as a result of this in accuracy. As expected this resulted in a slightly lower accuracy of 0.73 for strong-3/strong-2 at 10% error for attribute recognition. Although accuracy here was lower than that observed when there was
no error, this decrease is small when compared with the large recognition error of 10\%. This showed that the proposed model was robust to realistic error due to recognition of attributes automatically. Therefore testing of learning approach and overall ranking demonstrated high performance rates. Methods proposed were able to effectively capture judgements made by annotators by extracting the criteria for various attributes and their relations. As a result of this rankings similar to that from experts were obtained. In addition this graph based approach compared very well against baselines in Chapter 4 and Chapter 5. An accuracy of 0.94 was obtained for strong-5 at 10\% error utilizing the lookup model baseline while an accuracy of 0.360 for strong-5 was obtained from the global ranking based on matching approach. In comparison an accuracy of 0.73 is obtained using the image representation model when it is trained using strong-2 and tested with strong-3 at a realistic error of 10\% for attribute recognition.

### 7.2 Future work

The work offers a broad range of opportunities for future work due to attracting attention only in recent times. The dataset introduced in Chapter 3 Section 3.1.1 can support future research, particularly in the fields of visual assessments and fashion based studies [91].

An implementation of the Kemeny-Young method was utilized to combine preference lists from multiple annotators into a single preference order as described in Chapter 3 Section 3.5. This voting scheme was employed due to two main reasons. Firstly, it is appropriate for combining annotations because the annotators can be intuitively viewed as voters. Secondly, its usage has long been established in the domain of electoral voting since being developed in 1959 by John Kemeny [66, 132]. An intriguing direction for the future could be to investigate implementations of alternative ranking methods of varying criteria. This would allow a comparative study so that an optimum can be obtained. For example, ranking based on the popular models for pairwise scores in statistics such as Thurstone model and Bradley-Terry model could be explored [118, 19, 28]. The performance of these can then be evaluated in this context.

The attribute recognition system described in Chapter 4 Section 4.1 to determine prediction estimates of the attributes was not addressed as part of this work. This BoV based approach performs well for the clothing categories of top and bottom clothing. However, it shows a
decline in performance for the body shape attributes making the recognition system prohibitive for recognition of body shapes. This reduction in performance is a result of subtle differences in the body shapes that are not captured by SIFT that was designed to be robust against small shifts. General shape features are not designed for such task as a body shape recogniser requires much more accurate measurements from the images focussed on the mid body regions and global shape proportions. An interesting area in the future could involve the design of a feature that can be used to facilitate reliable body shape recognition.

In Chapter 5 matching techniques were evaluated for performing pairwise comparisons. They were found to exhibit inefficiencies in this context. Some of the recent developments in matching could be explored in order to improve performance to some extent and achieve better ranking. An example of this can include the use of novel query expansion with image augmentation as in [12]. Another one can be the use of a newer IDF expression when weighting histogram counts for an image in the BoV representation presented in [135].

One of the future directions could also involve extending the image representation model in Chapter 6 Section 6.1.4 with other attributes such as shoes, colour, and different accessories. This can enable an interesting study into the level of impact these supporting attributes have on the aesthetics of images. It can also be used to determine which one of these attributes exhibits a stronger influence on the overall dressing for an aesthetically pleasing outfit. A comparison to [78, 79] in which a dataset that includes clothing annotation was introduced will also be interesting once the dataset is released.

The approaches proposed in the work are not limited to fashion only. They are applicable to other domains where a ranking of data can represent human preferences and the assessment criteria can be defined. So it would be interesting to study and extend the methods in a different domain that requires the use of subjective assessments within a complex set-up.
Appendix A

Dataset: Aesthetics Based on Fashion Images

An approach for ranking images by pooling from the knowledge and experience of crowdsourced annotators was presented in this thesis. To address the highly subjective and complex problem of fashion interpretation and assessment of aesthetic qualities of images, a dataset was introduced\(^1\). This dataset includes images fully labelled with attributes of body shape (s), top (t) and bottom (b) clothing, with states of the image configurations (s t b) in Table A.1. It also includes aesthetic pairwise assessments performed by utilizing these images.

A.1 Images

Collectively there are 1064 images for 120 configurations where a configuration represents an image of a person of a body shape with specific top and bottom clothing categories. An image is represented by its configuration (config) and the number of images in that particular configuration. So image 1_10.jpg belongs to the first configuration and is the 10\(^{th}\) image of this configuration. Specific states for the configurations as given in Table A.1 are as follows;

Body shape (s):

\(^1\)The data is available from http://kahlan.eps.surrey.ac.uk/featureSpace/fashion
1: apple, 2: column, 3: hourglass, 4: pear

Top clothing (t):
1: fitted top, 2: fitted jacket, 3: loose jacket, 4: loose top, 5: ruffled top

Bottom clothing (b):
1: flared trousers, 2: fitted trousers, 3: straight trousers, 4: flared skirt, 5: fitted skirt, 6: straight skirt

For example, image 1_10.jpg of config 1 (s t b: 1 1 1) in Table A.1 has an apple body shape with the person wearing a fitted top with flared trousers. The states for the remaining configurations can be determined in a similar manner.

A.2 Aesthetic comparisons

Ten annotators who follow fashion were recruited and allowed the scoring of a total of 70000 images. The size of the expert and repeated control pairs was set to 700 paired-images, details of which can be found in Chapter 3 Section 3.3.

There are 10 files containing aesthetic pairwise comparisons for the 10 annotators and a file with annotations from the expert. Annotations given by the 10 annotators are consistent with the evaluation performed in Chapter 3 Section 3.6. An image pair with the preferred choice is given where 1 indicates a preference for the first image (left) and 2 shows the second image (right) is preferred.

As an example, consider a total number of 7000 aesthetic comparisons in a file which have been provided by the first annotator. The first comparison in this includes;

27_6.jpg 28_3.jpg 1

Therefore the annotator chooses the left image 27_6.jpg from the image pair. Out of the 7000 comparisons, 700 have been performed by the expert. And 700 are repeated in all the annotations so they have been performed by the remaining 9 annotators as well. These control image
A.2. Aesthetic comparisons

<table>
<thead>
<tr>
<th>config</th>
<th>s</th>
<th>t</th>
<th>b</th>
<th>config</th>
<th>s</th>
<th>t</th>
<th>b</th>
<th>config</th>
<th>s</th>
<th>t</th>
<th>b</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>41</td>
<td>1</td>
<td>1</td>
<td>3</td>
<td>81</td>
<td>1</td>
<td>1</td>
<td>5</td>
</tr>
<tr>
<td>2</td>
<td>2</td>
<td>1</td>
<td>1</td>
<td>42</td>
<td>2</td>
<td>1</td>
<td>3</td>
<td>82</td>
<td>2</td>
<td>1</td>
<td>5</td>
</tr>
<tr>
<td>3</td>
<td>3</td>
<td>1</td>
<td>1</td>
<td>43</td>
<td>3</td>
<td>1</td>
<td>3</td>
<td>83</td>
<td>3</td>
<td>1</td>
<td>5</td>
</tr>
<tr>
<td>4</td>
<td>4</td>
<td>1</td>
<td>1</td>
<td>44</td>
<td>4</td>
<td>1</td>
<td>3</td>
<td>84</td>
<td>4</td>
<td>1</td>
<td>5</td>
</tr>
<tr>
<td>5</td>
<td>1</td>
<td>2</td>
<td>1</td>
<td>45</td>
<td>1</td>
<td>2</td>
<td>3</td>
<td>85</td>
<td>1</td>
<td>2</td>
<td>5</td>
</tr>
<tr>
<td>6</td>
<td>2</td>
<td>2</td>
<td>1</td>
<td>46</td>
<td>2</td>
<td>2</td>
<td>3</td>
<td>86</td>
<td>2</td>
<td>2</td>
<td>5</td>
</tr>
<tr>
<td>7</td>
<td>3</td>
<td>2</td>
<td>1</td>
<td>47</td>
<td>3</td>
<td>2</td>
<td>3</td>
<td>87</td>
<td>3</td>
<td>2</td>
<td>5</td>
</tr>
<tr>
<td>8</td>
<td>4</td>
<td>2</td>
<td>1</td>
<td>48</td>
<td>4</td>
<td>2</td>
<td>3</td>
<td>88</td>
<td>4</td>
<td>2</td>
<td>5</td>
</tr>
<tr>
<td>9</td>
<td>1</td>
<td>3</td>
<td>1</td>
<td>49</td>
<td>1</td>
<td>3</td>
<td>3</td>
<td>89</td>
<td>1</td>
<td>3</td>
<td>5</td>
</tr>
<tr>
<td>10</td>
<td>2</td>
<td>3</td>
<td>1</td>
<td>50</td>
<td>2</td>
<td>3</td>
<td>3</td>
<td>90</td>
<td>2</td>
<td>3</td>
<td>5</td>
</tr>
<tr>
<td>11</td>
<td>3</td>
<td>3</td>
<td>1</td>
<td>51</td>
<td>3</td>
<td>3</td>
<td>3</td>
<td>91</td>
<td>3</td>
<td>3</td>
<td>5</td>
</tr>
<tr>
<td>12</td>
<td>4</td>
<td>3</td>
<td>1</td>
<td>52</td>
<td>4</td>
<td>3</td>
<td>3</td>
<td>92</td>
<td>4</td>
<td>3</td>
<td>5</td>
</tr>
<tr>
<td>13</td>
<td>1</td>
<td>4</td>
<td>1</td>
<td>53</td>
<td>1</td>
<td>4</td>
<td>3</td>
<td>93</td>
<td>1</td>
<td>4</td>
<td>5</td>
</tr>
<tr>
<td>14</td>
<td>2</td>
<td>4</td>
<td>1</td>
<td>54</td>
<td>2</td>
<td>4</td>
<td>3</td>
<td>94</td>
<td>2</td>
<td>4</td>
<td>5</td>
</tr>
<tr>
<td>15</td>
<td>3</td>
<td>4</td>
<td>1</td>
<td>55</td>
<td>3</td>
<td>4</td>
<td>3</td>
<td>95</td>
<td>3</td>
<td>4</td>
<td>5</td>
</tr>
<tr>
<td>16</td>
<td>4</td>
<td>4</td>
<td>1</td>
<td>56</td>
<td>4</td>
<td>4</td>
<td>3</td>
<td>96</td>
<td>4</td>
<td>4</td>
<td>5</td>
</tr>
<tr>
<td>17</td>
<td>1</td>
<td>5</td>
<td>1</td>
<td>57</td>
<td>1</td>
<td>5</td>
<td>3</td>
<td>97</td>
<td>1</td>
<td>5</td>
<td>5</td>
</tr>
<tr>
<td>18</td>
<td>2</td>
<td>5</td>
<td>1</td>
<td>58</td>
<td>2</td>
<td>5</td>
<td>3</td>
<td>98</td>
<td>2</td>
<td>5</td>
<td>5</td>
</tr>
<tr>
<td>19</td>
<td>3</td>
<td>5</td>
<td>1</td>
<td>59</td>
<td>3</td>
<td>5</td>
<td>3</td>
<td>99</td>
<td>3</td>
<td>5</td>
<td>5</td>
</tr>
<tr>
<td>20</td>
<td>4</td>
<td>5</td>
<td>1</td>
<td>60</td>
<td>4</td>
<td>5</td>
<td>3</td>
<td>100</td>
<td>4</td>
<td>5</td>
<td>5</td>
</tr>
<tr>
<td>21</td>
<td>1</td>
<td>1</td>
<td>2</td>
<td>61</td>
<td>1</td>
<td>1</td>
<td>4</td>
<td>101</td>
<td>1</td>
<td>1</td>
<td>6</td>
</tr>
<tr>
<td>22</td>
<td>2</td>
<td>1</td>
<td>2</td>
<td>62</td>
<td>2</td>
<td>1</td>
<td>4</td>
<td>102</td>
<td>2</td>
<td>1</td>
<td>6</td>
</tr>
<tr>
<td>23</td>
<td>3</td>
<td>1</td>
<td>2</td>
<td>63</td>
<td>3</td>
<td>1</td>
<td>4</td>
<td>103</td>
<td>3</td>
<td>1</td>
<td>6</td>
</tr>
<tr>
<td>24</td>
<td>4</td>
<td>1</td>
<td>2</td>
<td>64</td>
<td>4</td>
<td>1</td>
<td>4</td>
<td>104</td>
<td>4</td>
<td>1</td>
<td>6</td>
</tr>
<tr>
<td>25</td>
<td>1</td>
<td>2</td>
<td>2</td>
<td>65</td>
<td>1</td>
<td>2</td>
<td>4</td>
<td>105</td>
<td>1</td>
<td>2</td>
<td>6</td>
</tr>
<tr>
<td>26</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>66</td>
<td>2</td>
<td>2</td>
<td>4</td>
<td>106</td>
<td>2</td>
<td>2</td>
<td>6</td>
</tr>
<tr>
<td>27</td>
<td>3</td>
<td>2</td>
<td>2</td>
<td>67</td>
<td>3</td>
<td>2</td>
<td>4</td>
<td>107</td>
<td>3</td>
<td>2</td>
<td>6</td>
</tr>
<tr>
<td>28</td>
<td>4</td>
<td>2</td>
<td>2</td>
<td>68</td>
<td>4</td>
<td>2</td>
<td>4</td>
<td>108</td>
<td>4</td>
<td>2</td>
<td>6</td>
</tr>
<tr>
<td>29</td>
<td>1</td>
<td>3</td>
<td>2</td>
<td>69</td>
<td>1</td>
<td>3</td>
<td>4</td>
<td>109</td>
<td>1</td>
<td>3</td>
<td>6</td>
</tr>
<tr>
<td>30</td>
<td>2</td>
<td>3</td>
<td>2</td>
<td>70</td>
<td>2</td>
<td>3</td>
<td>4</td>
<td>110</td>
<td>2</td>
<td>3</td>
<td>6</td>
</tr>
<tr>
<td>31</td>
<td>3</td>
<td>3</td>
<td>2</td>
<td>71</td>
<td>3</td>
<td>3</td>
<td>4</td>
<td>111</td>
<td>3</td>
<td>3</td>
<td>6</td>
</tr>
<tr>
<td>32</td>
<td>4</td>
<td>3</td>
<td>2</td>
<td>72</td>
<td>4</td>
<td>3</td>
<td>4</td>
<td>112</td>
<td>4</td>
<td>3</td>
<td>6</td>
</tr>
<tr>
<td>33</td>
<td>1</td>
<td>4</td>
<td>2</td>
<td>73</td>
<td>1</td>
<td>4</td>
<td>4</td>
<td>113</td>
<td>1</td>
<td>4</td>
<td>6</td>
</tr>
<tr>
<td>34</td>
<td>2</td>
<td>4</td>
<td>2</td>
<td>74</td>
<td>2</td>
<td>4</td>
<td>4</td>
<td>114</td>
<td>2</td>
<td>4</td>
<td>6</td>
</tr>
<tr>
<td>35</td>
<td>3</td>
<td>4</td>
<td>2</td>
<td>75</td>
<td>3</td>
<td>4</td>
<td>4</td>
<td>115</td>
<td>3</td>
<td>4</td>
<td>6</td>
</tr>
<tr>
<td>36</td>
<td>4</td>
<td>4</td>
<td>2</td>
<td>76</td>
<td>4</td>
<td>4</td>
<td>4</td>
<td>116</td>
<td>4</td>
<td>4</td>
<td>6</td>
</tr>
<tr>
<td>37</td>
<td>1</td>
<td>5</td>
<td>2</td>
<td>77</td>
<td>1</td>
<td>5</td>
<td>4</td>
<td>117</td>
<td>1</td>
<td>5</td>
<td>6</td>
</tr>
<tr>
<td>38</td>
<td>2</td>
<td>5</td>
<td>2</td>
<td>78</td>
<td>2</td>
<td>5</td>
<td>4</td>
<td>118</td>
<td>2</td>
<td>5</td>
<td>6</td>
</tr>
<tr>
<td>39</td>
<td>3</td>
<td>5</td>
<td>2</td>
<td>79</td>
<td>3</td>
<td>5</td>
<td>4</td>
<td>119</td>
<td>3</td>
<td>5</td>
<td>6</td>
</tr>
<tr>
<td>40</td>
<td>4</td>
<td>5</td>
<td>2</td>
<td>80</td>
<td>4</td>
<td>5</td>
<td>4</td>
<td>120</td>
<td>4</td>
<td>5</td>
<td>6</td>
</tr>
</tbody>
</table>

Table A.1: States of an image configuration (s t b) with attributes of body shape (s), top (t) and bottom (b) clothing.

Pairs have been sampled at regular intervals within an annotated dataset. The expert image pairs are included at positions 5, 15, 25, ..., 6995 and the repeated control pairs are included at 10, 20, 30, ..., 7000. For example, comparison 114_1.jpg 39_1.jpg occurs at the 25th position in the 10 aesthetic comparison files and at the 3rd position in the list where the pairwise comparisons have been performed by the expert. And comparison 102_7.jpg 80_2.jpg is present at the 20th position in the 10 aesthetic comparison files.
Bibliography


