Aesthetics based assessment and ranking of fashion images

A. Gaur*, K. Mikolajczyk

University of Surrey, Guildford, Surrey, GU2 7XH, UK

Abstract

We present an approach for ranking images by pooling from the knowledge and experience of crowdsourced annotators. Specifically, we address the highly subjective and complex problem of fashion interpretation and assessment of aesthetic qualities of images. To utilize the visual judgements, we introduce a novel dataset complete with labellings of various attributes of clothing and body shapes. Large scale pairwise comparisons of the order of tens of thousands are performed by annotators who follow fashion. We then apply various consistency measures to verify the agreement and correlation between the annotators to rule out inconsistencies amongst them. Based on the annotations we establish reliable rankings which are utilized to learn an image representation based on qualitative assessments of visual aesthetics. This relies on a multi-node multi-state model that represents image attributes and their relations. Bag-of-features object recognition is used for the classification of visual attributes such as clothing and body shape in an image. The attributes and their relations are then 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.

Keywords: Visual assessment, Aesthetics, Attributes, Annotation, Ranking, Graphical modelling

*Corresponding author

Email address: a.gaur@surrey.ac.uk (A. Gaur)

1. Introduction

The use of crowdsourcing to aid computer vision is a highly active area of research which has yet to reach maturity [1, 2, 3, 4, 5, 6]. The potential benefits of reducing model complexity and amount of processing via crowdsourcing outweighs the cost factor with visual assessment being a markedly recent example [1, 7, 8, 4]. The underlying pattern and correlation from these assessments can be used to gain perspectives of individuals from which an objective measure can be constructed. Assessing image quality based on visual perspective has gained momentum in recent years in computer vision, machine learning and image processing [9, 10, 11, 12]. 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 [13, 14], with specific applications for web image searching [15, 16]. Additional miscellaneous areas include object annotation [17], segmentation [18] and saliency detection [19]. A long established use of ranking is found in preferential voting systems [20, 21]. This is generally a much smaller domain and involves fewer number of candidates.

In this work we propose an approach for comparing images based on qualitative assessment of image content and aesthetic impression introduced in [22]. This is in contrast to ranking based on relevance to well defined image content where a similarity measure can be established between images. The aesthetic impression can be considered a hidden variable that is affected by various image attributes and relations between the attributes. We propose to construct an image representation using graphical modelling where the attributes and their relations are learnt from the ranked data. Fashion annotators provide the ratings based on pairwise preferences. From this, the annotated datasets are ranked and the underlying relations are extracted as part of the learning process that constructs the model. We apply this approach to fashion interpretation which has recently attracted more attention [1, 23, 24, 25] but it is also relevant to other computer vision areas such as assessing facial beauty [26], retrieval and

recognition [27, 2], annotating data [1, 25] and qualitative assessments [9, 10].

Our goal is to obtain an objective ranking based on certain criteria by comparing only two images at a time. The criteria can be varied with different levels of complexity for various applications and involve a large number of interrelated attributes as well as features alongside varied rules. Consider a person with a specific body shape who looks better with certain top and bottom clothing attributes e.g. a fitted top with flared skirt or loose top with fitted skirt. Other factors may also be considered like colour, texture, pattern and general fashion-related rules for dressing. One can use attribute recognition to label different parts automatically, however implementing general fashion rules to automatically obtain a ranking is very challenging in this context. That is why we adopt a different approach that can learn the influence and relations between many components from a ranked list of images. To produce a ranked list of many images we break the task to comparative scoring between two images at a time. The comparisons are performed by a number of mid-level annotators with knowledge of fashion and its principles. This method requires a binary decision from the annotator which can be made quickly with a low-level of ambiguity [26]. Following this we use various consistency checks to validate the annotations collected from the annotators and combine the pairwise preferences into global rankings. These rankings are then used as the reference sets for learning and evaluation. Given this data we train a graphical model that captures all the attributes and relations between them.

Publicly available datasets for performing fashion-related study of attributes are limiting in this context. As an example the data from [27, 1] does not address coordinating attributes worn on upper and lower half of the body. Therefore we introduce a novel dataset consisting of 1064 images [1]. Images have been fully labelled with attributes of body shape, top and bottom clothing as well as aesthetic pairwise assessments. We first outline the related work in Section 1.1. The dataset together with the attributes alongside the crowdsourcing procedure

[1]The data is available from http://kahlan.eps.surrey.ac.uk/featurespace/fashion

used within the approach is presented in Section 2. Next, evaluation design with the method for obtaining the objective ranking is described in Section 3. Then, method for recognizing the attributes along with automatic assessment model is outlined in Section 4. Graph based model and representation of the rankings using our learning approach is discussed in Section 5. Finally, experimental results for the evaluations performed are shown in Section 6.

1.1. Related work

Using crowdsourcing is becoming increasingly popular in the field of machine vision and introduces different challenges such as dealing with annotation noise. The level of difference between expert annotations was investigated in [3] to verify if the annotations need to be repeated. They also studied whether non-expert annotations can be reliably utilized for ground-truth annotations in a benchmark scenario. A method that combines preference and absolute judgements is proposed in [4]. Using a batch-mode active learning method, they construct a set of queries. Matching of tasks to workers from a mechanism design perspective is done in [6]. Using the bipartite graph between workers and task, they propose a uniform mechanism for the allocation. In a crowdsourcing scenario, [5] conduct experiments on Amazon Mechanical Turk to understand different voting rules.

There are several approaches to obtain rankings in this context. In absolute rating, the user when presented with an image has to assign it a score, usually between 1 and 10 [28]. When sorting a collection of images, the user is required to sort images based on some criteria where all the data has to be considered together [29]. For performing pairwise comparisons the user is presented with a binary decision [26]. In k-wise comparison ($k = 10$) the number of pairwise preferences attained from a k-wise rating is $\binom{k}{2}$ [1]. Pairwise comparisons reduce the level of challenge and ambiguity associated with the ratings. This approach was used to label facial beauty data in [26] with the assumption that people in general have a consistent opinion on facial attractiveness. The absolute scores are obtained by minimizing a cost function that penalizes images that are not in agreement with one of the pairwise preferences.

Visual assessment of the quality of images using their proposed regional and global features is done in [10] while [11] automatically assess the aesthetics of images using generic image descriptors, such as, SIFT and GIST. An image quality metric for auto-denoising is presented in [9]. Bag-of-colour-patterns approach that evaluates the colour harmony of photos with aesthetic quality classification is proposed in [12]. In [16] a re-ranking approach that automatically learns different offline visual semantic spaces is given. A graph-theoretical framework for noise resistant ranking is proposed in [15]. Facial beauty modelling was addressed in [26]. In [27] an effective method for parsing clothing in fashion photographs is presented. They also introduce a novel dataset for garment items and present results on using information about clothing estimates to improve pose identification. Cross-scenario clothing retrieval is addressed in [2] where using a human photo taken from the street they find similar clothing from online shops. Key components proposed here include human/clothing parts alignment and an auxiliary daily photo dataset. Closely related work was recently presented in [1] which discusses approaches to obtain image rankings and learn attribute based models. A cloth recommendation application is considered in [23, 24, 25]. However, [23] uses a common sense reasoning rather than vision based learning. In [24] a graphical model is used that given a cloth part proposes another one. Similar idea is exploited in [25] but introduces attributes and occasion components. Our objective is to learn a model directly from a ranked list of images and to rate outfits to reflect recommendations of fashion experts.

2. Crowdsourcing for qualitative assessments

In this section, we first present the dataset with the relevant attributes and labellings. Next, we discuss the structure of the dataset and included subsets of data for various consistency checks. Finally, we describe the process of acquiring annotations with the design of the tool, guidelines and image-pairs.

2.1. Dataset and labelling

There are several datasets for assessing facial beauty [26] but very few that are related to fashion. The datasets from [27, 2] that include clothing annota-

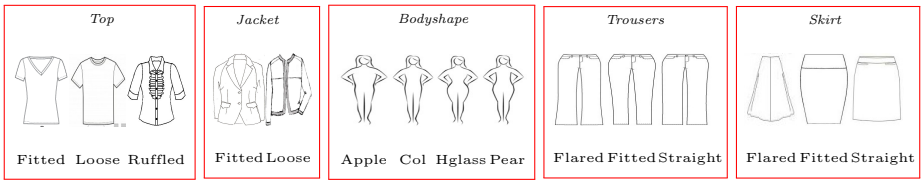

Figure 1: Clothing and body shape attributes in our dataset. There are 11 clothing and 4 body shapes categories. An image represents a configuration where a person with a certain body shape wears specific top and bottom clothing.

tion are not annotated according to aesthetic qualities by fashion experts. We therefore collected images suitable for performing comparative visual assessments from different clothing retailers, including high-end, high-street, budget and dedicated online shopping retailers. We use 15 different categories that include 11 categories for clothing and 4 for body shape attributes as shown in Fig. 1. For the top clothing attributes, we use 5 categories including top: fitted, loose, ruffled and jacket: fitted and loose. 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. We therefore collect image examples for all possible configurations of clothing attributes and warp the images 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.

Our dataset puts an emphasis on visual aesthetics related to the aforementioned attributes and coordinating these attributes for specific configurations. It imposes no restrictions in terms of the person's facial beauty, their ethnicity or age. To eliminate the face-related bias, we exclude this part from the images. Besides fashion-based studies our dataset can also be used in other clothing-related problems such as recognition and retrieval [27, 2]. Some examples from

the dataset that collectively consists of 1064 images for the 120 configurations can be seen in Fig. 2 (Left).

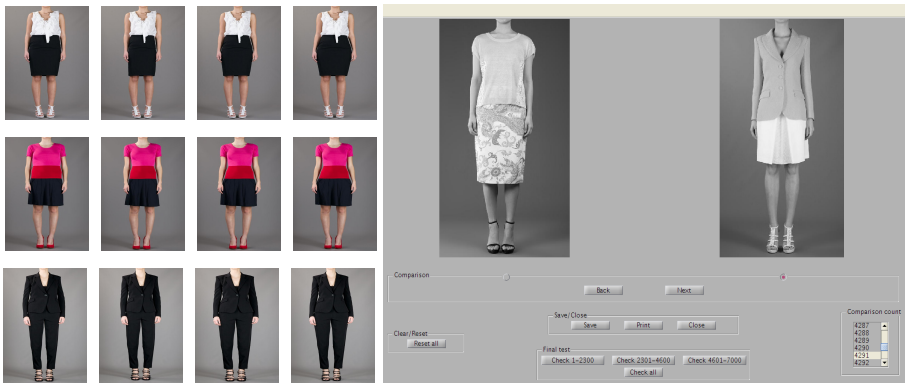

Figure 2: (Left) Example of images from the dataset where each image is warped to a specific body shape. From left to right the body shape attributes shown are apple, column, hourglass and pear. (Right) Interface used for crowdsourcing the pairwise comparisons.

2.2. Control sets

In order to measure the annotation consistency between different annotators we introduce several control sets of image-pairs. Specifically we use two subsets within a large list of pairs presented to each annotator. One is annotated by an expert and the other one is repeated in all lists for annotators. In the study conducted in [3] it was found that expert annotators showed a high consistency in annotations on measuring agreement and correlation. In our experiments one expert is utilized for annotating the reference control set which makes the process more cost-effective and feasible

A list of images L_a presented to annotator a consists of $L_a = L_{e,a} + L_{r,a} + L_{u,a}$. $L_{e,a}$ 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.

2.3. Collecting annotations

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 we selected pairs from the longest diagonals of the triangle.

The expert as well as the annotators were provided with an assessment tool (see Fig. 2 (Right)) to obtain binary scores. Images in grayscale were presented to each annotator and the one that created a better overall aesthetic impression in each pair was selected. The instructions provided to the annotators included how well did the overall clothing look on a given body shape.

3. Evaluation of annotations

In this section we first describe the evaluation measures that are used to assess the agreement and correlation of the annotations. We then outline the criteria used to validate annotations and explain the approach to generate rankings from individual lists. Finally, we describe the method for obtaining rankings from the validated annotations.

3.1. Agreement amongst annotators

An intuitive method to test the consistency amongst the annotators can be based on the agreement between the annotated lists. First, we compute the agreement between the expert list $L_{e,e}$ and annotated list $L_{e,a}$ for all A annotators. We then measure the inter-annotator agreement between all pairs of lists $L_{r,a}$.

Accuracy [30] between two sets of annotated image pairs $L_{e,a}$ and $L_{e,e}$ is given by:

$$acc(L_{e,a}, L_{e,e}) = \frac{|\{\forall i, L_{e,a}\{i\} = L_{e,e}\{i\}\}|}{|L_e|} \quad (1)$$

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 *acc* value of 0.5 corresponds to a random score and 1 to identical lists.

Cohen's κ [31] coefficient is a statistical measure that is used to qualitatively evaluate the agreement amongst annotators [3]. 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. The values for κ 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 τ [32] coefficient is used to assess the correlation between ranked lists [5]. It is based on a non-parametric statistic to measure the level of association between two rankings. The correlation score varies between 1 for identical rankings, 0 for independent ones to -1 for perfect inverse of two rankings. Note that in our approach a ranking has to be generated from pairwise comparisons in order to use Kendall's τ measure.

Annotator validation is based on the the agreement scores from all three measures. If the values for an annotator a are above a threshold, then the annotated list L_a is considered valid. We use the expert $L_{e,a}$ and the repeated $L_{r,a}$ lists to evaluate annotation agreements. Based on this validation we can identify a group of strong and weak annotators.

3.2. Rankings based on pairwise comparisons

To generate rankings from $L_{r,a}$ and $L_{e,a}$ we order all the pairwise decisions into overall preference orders. This problem has been widely studied in the domain of electoral voting for which the Kemeny-Young [20, 21] method was developed. As a voting algorithm it not only computes the winner but also

determines an entire ranked list. This Condorcet method orders M candidates such that the winners are ranked at the top N. It is similar to Kendall's τ since it makes use of relative ordering and minimizes the disagreements amongst the voters in their pairwise preferences between all the candidates. However, it is known that this ranking problem is NP hard even for a single minimum discrepancy ordering [33]. Real-application use of this approach can only be heuristic and we make use of one such implementation [33]. It consists of $Ntry$ 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. In our application, 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. We have tried a range of values for $Ntry > 10$ and observed little variations between the resulting rankings, we therefore use $Ntry = 10$ in all our experiments.

Several ranked lists can be produced from the annotated image pairs using the Kemeny-Young method. 1) A ranking of images within each annotator list L_a combined from the three subsets $L_{e,a}$, $L_{r,a}$ and $L_{u,a}$, 2) A ranking of all images combined from several or all annotators. 3) Note that each image is an example of a configuration of clothing and body shape we therefore can compress each ranking of images into a ranking of 120 configurations (cf. Section 2.1). For the image pairs from the repeated list $L_{r,a}$ that are scored several times by different annotators we select the more frequent score.

4. Rankings based on attribute recognition

We recognise the attributes with the bag-of-features approach based on SIFT features used to train SVM classifiers [17]. There are 15 different attributes including 11 clothing and 4 for body shapes. The training and testing of the classifiers is done using images from the dataset that are presented in Section 2.1 and shown in Fig. 2 (Left). The classification decision can be based on hard threshold of the confidence score that is output by the classifier or by using the

label of the classifier with the highest score. We consider both techniques in our experimental evaluation.

Automatic assessment: The validated annotations are represented in the form of objective rankings using the Kemeny-Young method outlined in Section 3.2 ($Ntry = 2000$). A baseline approach is to use the ranking 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 query 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 a certain error associated with the 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 for the lookup ranking. This will serve as a baseline when evaluating the representation model described in Section 5.

5. Image model: learning and recognition

We first present our graph based representation and then describe how the global ranking is utilized to model various attributes and rating criteria. Finally, we present specifics of the model for application to fashion assessment.

5.1. Graph based model

The ranking of images based on the aesthetic impression they make can be simplified to producing an absolute rating where the approach is presented with a single image and generates a score within a normalised 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. We propose to model the attributes and their relations with graphical modelling, which is well suited to represent the attributes as states of nodes of a graph as well as relations between the various attributes represented by edges between the

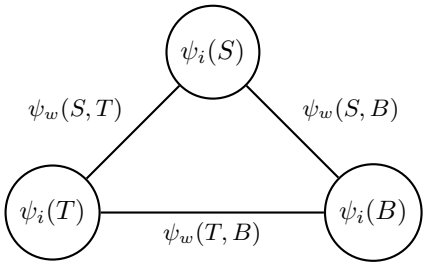

Figure 3: Object representation model for modelling the ranked lists with 3-nodes each at a range of states. In particular, this figure depicts two nodes for the clothing attributes of top (T), bottom (B) clothing and another for the body shape attribute (S) with the associated node ψ_i and edge ψ_w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

graph nodes. The states of each node and the relations between the states have certain potentials with which they contribute to the overall score of aesthetic appearance. In our fashion assessment application the nodes correspond to body parts and the states of the nodes correspond to clothing and body attributes. Fig. 3 illustrates the model we adopt, where edges between the states of the nodes represent relations between the attributes.

5.2. Learning image ranking

To facilitate the modelling and rating images we consider the position in the ranking 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. We consider the probabilistic scenario where the dependencies within the graph involve N nodes. 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 ψ as:

$$p(y_1, y_2, ..., y_N) = \frac{1}{Z} \prod_{i=1}^{N} \psi_i(y_i) \prod_{w=1}^{W} \psi_w(y_q, y_v), \tag{2}$$

where potential function ψ_i is associated with node i and ψ_w is associated with edge w connecting nodes y_q and y_v. This distribution is normalised with

constant Z given by:

$$Z = \sum_{y_1} \sum_{y_2} \cdots \sum_{y_N} \prod_{i=1}^{N} \psi_i(y_i) \prod_{w=1}^{W} \psi_w(y_q, y_v) \tag{3}$$

Learning attributes potentials: Learning the model requires estimating all node and edge potentials from a training data. The training data is in the form of a ranked list of images that is obtained by manual annotation. Providing objective ranking by manual annotation, in particular when there can be hundreds of possible configurations, is not straightforward. We discussed the process of obtaining such ranking in Section 3 and below we discuss the estimation of the potentials.

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

$$p(y_1, y_2, ..., y_N) = \frac{1}{E} \sum_{i=1}^{E} p(y_{1_i}, y_{2_i}, ..., y_{N_i}) \tag{4}$$

where $p(y_{1_i}, y_{2_i}, ..., y_{N_i})$ 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 we obtain this estimate for each unique configuration of node states, we can use it to learn the node potentials ψ_i and edge potentials ψ_w. For $\psi(y_1)$ we average over all configurations that include state $y_1 = z_1$ of node 1 as follows:

$$\psi(y_1 = z_1) = \sum_{y_2} \sum_{y_3} \cdots \sum_{y_N} p(y_1 = z_1, y_2, y_3, ..., y_N) \tag{5}$$

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

$$\psi(y_1, y_2) = \sum_{y_3} \sum_{y_4} \cdots \sum_{y_N} p(y_1 = z_1, y_2 = z_2, y_3, y_4, ..., y_N) \tag{6}$$

Fashion aesthetics: In our application of aesthetic assessment we consider a 3-node model with 4 states for the body shape, 5 for top and 6 for bottom clothing attributes. The attributes are listed in Table 3. One could add more nodes to represent shoes, jewellery, purse and other accessories as well as more states such

as colour and texture but this requires a large training set where each state is included in various configurations of attributes. Furthermore, the study from [1] shows that colour has little impact on the overall dressing attractiveness. For example, in our specific case, the potential $\psi(S_s)$ for body shape at state s is:

$$\psi(S_s) = \sum_t \sum_b p(S = s, T_t, B_b) \tag{7}$$

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

$$\psi(S_s, B_b) = \sum_t p(S = s, T_t, B = b) \tag{8}$$

6. Experimental results

In this section the performance of our approach is evaluated. We first validate the annotations obtained from the pairwise comparisons to determine stronger and weaker annotators. Next, performance of the attribute recognition is discussed which is used to train a lookup model based on annotations. Finally, we investigate the performance of the representation model.

6.1. Image ranking using crowdsourcing

We first determine the criteria that are used for computing the measures. Next, we analyse the agreement and correlation between the rankings generated from the expert and other annotators. Finally, we perform the inter-annotator analyses by comparing rankings from the annotators.

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 we also perform an experiment on a simulated dataset. This dataset is obtained for 57400 image-pairs by randomly generating the comparison score for the 10 annotation sets.

Measures: The evaluation is performed using the procedure and measures of agreement and correlation outlined previously in Section 3.1.

6.1.1. *Criteria for agreement*

In this section we analyse the relations between the accuracy, Cohen's κ and Kendall's τ as well as determine the thresholds that will be used to validate the annotators. This is performed using two sets of 700 image-pairs with simulated binary scores such that the agreement between the two sets increases from 0% to 100%. Agreement results based on the accuracy and Cohen's κ can be calculated from the binary scores in contrast to Kendall's τ which compares full rankings of images generated with Kemeny-Young method. Fig. 4 (Left) displays the values for accuracy, Kendall's τ and Cohen's κ when increasing agreement between two sets. The results for Cohen's κ are closely correlated with the accuracy. There is a non linear increase of Kendall's τ 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 τ values at high levels of agreements. A positive agreement and correlation in which the agreement exceeds 50% 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.

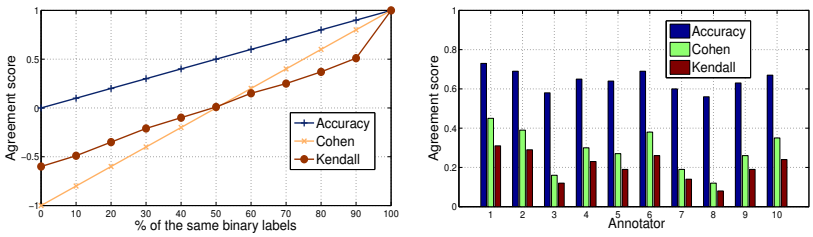

Figure 4: (Left) Accuracy, Cohen's κ and Kendall's τ for increasing agreement in the annotation decisions. (Right) Accuracy, Cohen's κ and Kendall's τ between the expert and the 10 annotators.

6.1.2. *Annotator vs. expert agreement*

This section investigates the quality of the annotations by measuring the agreement between the annotators and the expert. We also measure 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.

The results for the three measures are presented in Fig. 4 (Right). The highest agreement is observed for annotator 1 where 511 out of the 700 pairwise comparisons were in agreement. Note that the accuracy value of a random agreement is 0.5 which corresponds to 0 of Cohen's κ and Kendall's τ. 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.

We use Kemeny-Young method to obtain ranking lists of 711 images which formed the set of 700 image-pairs. The results obtained for full rankings and Kendall's τ measure are consistent with the above observations. Similarly, for annotator 1, Kendall's $\tau = 0.31$ which in Fig. 4 (Left) corresponds to an agreement of nearly 80%. The smallest correlation is seen for annotator 8 with Kendall's τ of 0.08 which corresponds to approximately 50% agreement. 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 were honoured in the full ranking. This is due to fewer conflicting decisions within each annotated set than in the case of the random scores. The effectiveness of Kemeny-Young in converting binary decisions into a ranking list is thus demonstrated.

6.1.3. Inter-annotator agreement

After comparing the annotations with the expert, the agreement between the annotators is evaluated. This can be used to find which annotators apply similar criteria in scoring fashion images. The values of the accuracy and Cohen's κ are shown in Table 1. The highest agreement is observed between annotators 1 and 2 with an accuracy of 0.70 and Cohen's κ of 0.40. This corresponds to an agreement of 490 image-pairs out of the 700 in this set. It was found from the previous testing that annotators $1, 2, 4, 6, 10$ had the most agreement and correlation with expert annotations. On further inspection of Table 1, 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

	E	A1	A2	A3	A4	A5	A6	A7	A8	A9	A10
E		0.73	0.69	0.58	0.65	0.64	0.69	0.60	0.56	0.63	0.67
A1	0.45		0.70	0.54	0.62	0.61	0.69	0.56	0.60	0.61	0.68
A2	0.39	0.40		0.57	0.62	0.57	0.66	0.61	0.61	0.66	0.69
A3	0.16	0.08	0.12		0.60	0.57	0.54	0.64	0.61	0.63	0.61
A4	0.30	0.25	0.23	0.19		0.57	0.63	0.62	0.60	0.61	0.62
A5	0.27	0.22	0.15	0.15	0.14		0.60	0.54	0.65	0.53	0.56
A6	0.38	0.37	0.33	0.08	0.26	0.20		0.60	0.57	0.61	0.65
A7	0.19	0.11	0.22	0.28	0.25	0.07	0.21		0.60	0.64	0.67
A8	0.12	0.20	0.21	0.22	0.21	0.30	0.14	0.20		0.56	0.64
A9	0.26	0.22	0.33	0.25	0.21	0.06	0.22	0.28	0.12		0.66
A10	0.35	0.36	0.39	0.22	0.25	0.11	0.30	0.34	0.28	0.33	

Table 1: The results for inter-annotator agreement. Accuracy is shown in the upper triangle and Cohen's κ is given in the lower triangle. Highlighted values represent the pairwise annotations that met the criteria when compared to the expert (cf. Section 6.1.1). A1-A2 is highlighted because both A1 and A2 meet the criteria compared to the expert annotations.

additionally with annotator 9. Overall, accuracy and Cohen's κ follow a pattern where the annotators who are in agreement with the expert annotations also agree amongst each other.

For the full rankings the inter-annotator correlation is measured using Kendall's τ and presented in Table 2. This evaluation set includes 741 independent images for the 700 image-pairs. Inter-annotator assessments are coherent with higher correlation coefficients observed for annotators $1, 2, 4, 6, 10$. 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. Furthermore, these results indicate that there is a significant consistency in assessing images based on aesthetic qualities, and it should be possible to automatically learn these criteria from the provided rankings.

6.2. Attribute recognition

For an automatic ranking of images we recognise the attributes using bags-of-features approach [17]. We report the performance for the clothing and body shapes with recall and precision measures. For each category, we split the data randomly into training and test sets. The positive training images for one category are used as the negative examples for all the other categories. Similarly

	A1	A2	A3	A4	A5	A6	A7	A8	A9	A10
E	0.31	0.29	0.12	0.23	0.19	0.26	0.14	0.08	0.19	0.24
A1		0.26	0.06	0.18	0.16	0.29	0.09	0.12	0.17	0.26
A2			0.09	0.17	0.09	0.20	0.17	0.14	0.23	0.27
A3				0.13	0.09	0.07	0.21	0.14	0.15	0.16
A4					0.09	0.16	0.19	0.14	0.15	0.17
A5						0.14	0.06	0.18	0.05	0.09
A6							0.16	0.10	0.14	0.23
A7								0.12	0.19	0.22
A8									0.08	0.20
A9										0.23

Table 2: Inter-annotator correlation between rankings using Kendall's τ. A1-A2 with a value of 0.26 is highlighted because both A1 and A2 meet the criteria when compared to the expert. τ of 0.26 corresponds to a large overlap of 70% for the decisions of the pairwise comparisons.

the positive test images for one category are used as the negative test images for all the other categories. The results reported in Table 3 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 (cf. Section 4). 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 $pmax^*$. 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. The reason for this performance decline comes from the very subtle differences in features extracted from different body shapes. In addition, the quantisation of SIFT features and spatial bins of the pyramid do not allow to capture these variations. We experimented with different variants of feature extractors but no significant improvement was observed. General shape descriptors 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. The best performance of 0.43 was observed for apple body where the shape differences are the largest compared to column or hourglass. Overall, our classification error is below 10% for most of the other attributes which is realistic for state-of-the-art recognition.

Category		Perf	$0.5th^*$	$0.7th^*$	$1th^*$	$pmax^*$
top	fitted	rec	0.91	0.90	0.89	0.97
		pre	0.98	1.00	1.00	0.97
	loose	rec	0.73	0.64	0.53	0.92
		pre	0.99	0.99	1.00	0.92
	ruffled	rec	0.75	0.74	0.74	0.90
		pre	0.98	0.99	1.00	0.90
jkt	fitted	rec	0.90	0.87	0.81	0.97
		pre	1.00	1.00	1.00	0.97
	loose	rec	0.83	0.80	0.76	0.94
		pre	1.00	1.00	1.00	0.94
trous	flared	rec	0.83	0.82	0.79	0.93
		pre	0.98	0.98	1.00	0.93
	fitted	rec	0.82	0.77	0.72	0.96
		pre	1.00	1.00	1.00	0.96
	straight	rec	0.72	0.66	0.58	0.92
		pre	0.98	0.98	1.00	0.92
skirt	flared	rec	0.80	0.78	0.73	0.95
		pre	0.98	0.99	1.00	0.95
	fitted	rec	0.84	0.81	0.77	0.97
		pre	0.99	0.99	1.00	0.97
	straight	rec	0.67	0.66	0.63	0.85
		pre	0.96	0.97	1.00	0.85
bshape	all	rec				0.30
		pre				0.30

Table 3: Recall and precision averaged over five runs of random splits for the clothing and body shape attributes where th^* is the threshold estimate for each individual category and $pmax^*$ is the maximum prediction estimate over the categories that are part of the same region (top, bottom).

6.3. Ranking with attribute recognition

To evaluate the impact of error due to attribute recognition, a 10% error (performance from Section 6.2) in the attributes is incorporated within a ranking for generating the lookup model as described in Section 4. The stronger annotations $1, 2, 4, 6, 10$ consisting of 29400 comparisons which includes 1400 control image-pairs serve as the ranking for this model (cf. Section 6.1). In addition to rankings from strong, weak, and all annotators, we also use a simulated dataset with random annotations for evaluating the accuracy. This gives a random ranking for the 10 annotators which will provide a baseline for the performance.

Automatic assessment: Accuracy of $0.94, 0.83, 0.71, 0.54$ is obtained when testing using stronger, all, weaker and random rankings for automatic comparison. As expected the highest score is obtained for the stronger group. Random

ranking has a much lower value as it is not correlated with any of the annotations. A reasonably high score is also seen when all the annotations are utilized. 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 the stronger group and all of the annotations. This is most likely due to these annotators applying different criteria when making fashion judgements.

6.4. Graph based model

In order to generate training and test rankings for evaluating the proposed approach we split the 10 annotators into four groups: two strong and two weak ones, with two or three annotators in each group (strong-2, strong-3, weak-2, weak-3). We first assess the potentials obtained from the representation model trained from the rankings. Then, accuracy of rankings from the model assuming that attributes in an image are known is discussed. Finally, we investigate the accuracy of produced rankings when attribute recognition is incorporated in the representation model.

6.4.1. Attributes potentials

We use the rankings generated with Kemeny-Young method [33] to learn the node and edge potentials, that is the attribute and relations potentials in the graph model as discussed in Section 5.2. In order to better visualise the learnt potentials we subtract from each estimated potential the corresponding potential learnt from a random ranking. Thus negative potentials in Fig. 5 indicate lower than random influence of a body shape or a cloth part on the overall rating of the image. For example, apple body shape, loose jacket and straight skirt have the lowest potentials. In addition, we observe that some potentials differ for strong and weak annotators groups e.g. loose top, which indicates slightly different criteria used by these groups. The overall rating consists of individual node potentials and edge potentials that correspond to the relations between certain clothing and body shapes. The relation potentials are illustrated in Fig. 6. Some relations are particularly strong in both negative and positive impact on the rating e.g. loose jackets or tops in combination with

apple shape in contrast to fitted jacket with column shape. These observations have been validated by expert annotator.

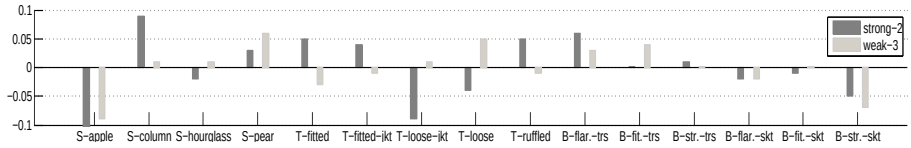

Figure 5: Attributes potentials learnt from rankings: strong-2 w.r.t. rand-s-2 and weak-3 w.r.t. rand-s-2.

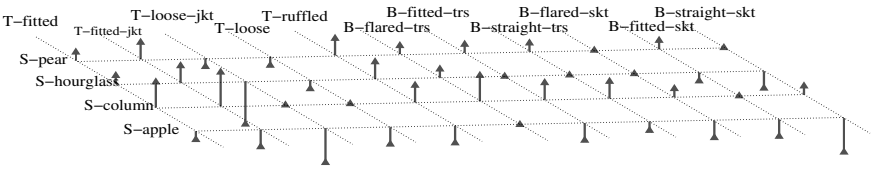

Figure 6: Attribute relations potentials between body shape and cloth, learnt from rankings: strong-2 w.r.t rand-s-2.

6.4.2. Ranking based on learnt attribute potentials

In order to validate the model we train it on image ranking that resulted from one group of annotators and test it on another one. This is measured with the average accuracy of pairwise preference ratings of images. In this experiment we assume that all the states of the nodes, that is the attributes present in the image are known. 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 rather than the error of the attribute recognition. Note that not all pairwise constraints given by the annotators can be satisfied in one global ranking as 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 4 shows the percentage of pairs correctly ranked for different training and testing sets. The highest score is obtained when trained and tested on rankings provided 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. We observe

that the results gradually decrease when training and testing on weak sets with the lowest results for randomly generated rankings. The random chance score for all train/test combinations is 0.5.

test \train	strong-2	strong-3	weak-2	weak-3	rand-s-2
strong-2	0.91	-	-	-	-
strong-3	0.76	0.87	-	-	-
weak-2	0.75	0.76	0.88	-	-
weak-3	0.62	0.66	0.66	0.84	-
rand-s-2	0.58	0.59	0.58	0.55	0.72
rand-w-2	0.55	0.58	0.54	0.52	0.56

Table 4: Accuracy of ranked pairs of images using the representation model and assuming the node states are known.

6.4.3. Image ranking with attribute recognition

The attribute recognition error has an impact on the performance of the entire ranking system. To assess this impact we carry out a controlled experiment where the percentage of the misclassified attributes in the individual classifiers is increased by a constant value for every consequent test. Previous misclassification are kept and used with added error for the next test. For every test, we estimate the performance by comparing the training and testing paired-configurations as in Section 6.4.2. The results are presented in Table 5. We make several observations 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. This compares well against the baseline lookup model presented in Section 6.3 in which an accuracy of 0.94 was obtained for stronger annotators. Moreover, our 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.

7. Conclusion and future work

We have presented an approach for obtaining an objective ranking using the judgements made from visual assessments of static images. It uses the knowl-

train/test	0%	10%	40%	70%
strong-2/strong-3	0.76	0.73	0.63	0.54
weak-2/weak-3	0.66	0.64	0.58	0.53
strong-2/rand-s-2	0.58	0.57	0.53	0.51
weak-2/rand-w-2	0.54	0.53	0.51	0.50

Table 5: Accuracy of ranked pairs of images using the representation model and attribute recognition with increasing % error for each of the 15 influencing categories.

edge of crowdsourcing annotators captured in pairwise comparisons. We introduced a dataset comprised of 1064 images fully labelled with various attributes of clothing and body shapes which will allow further development of similar approaches. We also presented an effective approach for learning the ranking of images using qualitative assessments of visual aesthetics. We proposed a graph based representation where node and edge potentials capture the importance of visual attributes and their relations. We have shown effectiveness of our approach on a collection of fashion images that include different combinations of clothing and body shapes. The results show that on using a sufficiently large number of comparisons, noisy assessments made by non-expert annotators can be filtered out. The method can also be applied to learn specific/individual preferences, or fashion rules in various cultural groups. A possible direction for future work would be to investigate alternative ranking methods of varying criteria and comparing them in order to obtain optimal rankings. We also aim to develop a more reliable body shape classifier and extend the model with other attributes such as shoes, colour, and different accessories. A comparison to [25, 2] will also be interesting once the dataset is released.

Acknowledgement. This work was supported by EU Chist-Era EPSRC EP/K01904X/1.

References

[1] T. V. Nguyen, S. Liu, B. Ni, J. Tan, Y. Rui, S. Yan, Sense beauty via face, dressing, and/or voice, in: ACM Multimedia, 2012.

[2] S. Liu, Z. Song, G. Liu, C. Xu, H. Lu, S. Yan, Street-to-shop: Cross-scenario clothing retrieval via parts alignment and auxiliary set, in: CVPR, 2012.

[3] S. Nowak, S. Rüger, How reliable are annotations via crowdsourcing, in: MIR, 2010.

[4] P. Ye, D. Doermann, Combining preference and absolute judgements in a crowd-sourced setting, in ICML'13 workshop: Machine Learning Meets Crowdsourcing (2013).

[5] A. Mao, A. D. Procaccia, Y. Chen, Better human computation through principled voting, in: AAAI Conference on Artificial Intelligence, 2013.

[6] G. Goel, A. Nikzad, A. Singla, Matching workers expertise with tasks: Incentives in heterogeneous crowdsourcing markets, in NIPS'13 Workshop on Crowdsourcing: Theory, Algorithms and Applications (2013).

[7] O. Wu, W. Hu, X. Li, J. Gao, Learning to predict the perceived visual quality of photos, in: ICCV, 2011.

[8] P. Ye, J. Kumar, L. Kang, D. Doermann, Real-time no-reference image quality assessment based on filter learning, in: CVPR, 2013.

[9] X. Kong, K. Li, Q. Yang, L. Wenyin, M.-H. Yang, A new image quality metric for image auto-denoising, in: ICCV, 2013.

[10] W. Luo, X. Wang, X. Tang, Content-based photo quality assessment, in: ICCV, 2011.

[11] L. Marchesotti, F. Perronnin, D. Larlus, G. Csurka, Assessing the aesthetic quality of photographs using generic image descriptors, in: ICCV, 2011.

[12] M. Nishiyama, T. Okabe, I. Sato, Y. Sato, Aesthetic quality classification of photographs based on color harmony, in: CVPR, 2011.

[13] J. He, J. Feng, X. Liu, T. Cheng, T.-H. Lin, H. Chung, S.-F. Chang, Mobile product search with bag of hash bits and boundary reranking, in: CVPR, 2012.

[14] M. Rastegari, C. Fang, L. Torresani, Scalable object-class retrieval with approximate and top-k ranking, in: ICCV, 2011.

[15] W. Liu, Y. G. Jiang, J. Luo, S. F. Chang, Noise resistant graph ranking for improved web image search, in: CVPR, 2011.

[16] K. Liu, X. Wang, Query-specific visual semantic spaces for web image re-ranking, in: CVPR, 2011.

[17] P. Koniusz, F. Yan, K. Mikolajczyk, Comparison of mid-level feature coding approaches and pooling strategies in visual concept detection, CVIU 117 (5) (2013) 479–492.

[18] P. Yadollahpour, D. Batra, G. Shakhnarovich, Discriminative re-ranking of diverse segmentations, in: CVPR, 2013.

[19] C. Yang, L. Zhang, H. Lu, X. Ruan, M. H. Yang, Saliency detection via graph-based manifold ranking, in: CVPR, 2013.

[20] J. Kemeny, Mathematics without numbers, Daedalus 88 (1959) 577–591.

[21] H. P. Young, A. Levenglick, A consistent extension of condorcet's election principle, SIAM Journal on Applied Mathematics 35 (2) (1978) 285–300.

[22] A. Gaur, K. Mikolajczyk, Ranking images based on aesthetic qualities, in: ICPR, 2014.

[23] E. Shen, H. Lieberman, F. Lam, What am i gonna wear?: Scenario-oriented recommendation, in: IUI, 2007.

[24] T. Iwata, S. Watanabe, H. Sawada, Fashion coordinates recommender system using photographs from fashion magazines, in: IJCAI, 2011.

[25] S. Liu, J. Feng, Z. Song, T. Zhang, H. Lu, C. Xu, S. Yan, Hi, magic closet, tell me what to wear!, in: ACM Multimedia, 2012.

[26] D. Gray, K. Yu, W. Xu, Y. Gong, Predicting facial beauty without landmarks, in: ECCV, 2010.

[27] K. Yamaguchi, M. H. Kiapour, L. E. Ortiz, T. L. Berg, Parsing clothing in fashion photographs, in: CVPR, 2012.

[28] R. Likert, A technique for the measurement of attitudes, Archives of Psychology 22 (140) (1932) 5–55.

[29] A. Oliva, A. Torralba, Modelling the shape of the scene: A holistic representation of the spatial envelope, IJCV 42 (2001) 145–175.

[30] T. Brants, Inter-annotator agreement for a German newspaper corpus, in: International Conference on Language Resources and Evaluation, 2000.

[31] J. Cohen, A coefficient of agreement for nominal scales, Educational and psychological measurement 20 (1) (1960) 37–46.

[32] M. Kendall, A new measure of rank correlation, Biometrika 30 (1938) 81–89.

[33] W. H. Press, C++ program for kemeny-young preference aggregation, `http://www.nr.com/whp/ky/kemenyyoung.html`, [Online] (August 2012).