Fashion Coordinates Recommender System using Photographs from Fashion Magazines

Tomoharu Iwata **Shinji Watanabe** **Hiroshi Sawada**
NTT Communication Science Laboratories
2-4 Hikaridai, Seika-cho, Soraku-gun, Kyoto, Japan
{iwata.tomoharu,watanabe.shinji,sawada.hiroshi}@lab.ntt.co.jp

Abstract

Fashion magazines contain a number of photographs of fashion models, and their clothing coordinates serve as useful references. In this paper, we propose a recommender system for clothing coordinates using full-body photographs from fashion magazines. The task is that, given a photograph of a fashion item (e.g. tops) as a query, to recommend a photograph of other fashion items (e.g. bottoms) that is appropriate to the query. With the proposed method, we use a probabilistic topic model for learning information about coordinates from visual features in each fashion item region. We demonstrate the effectiveness of the proposed method using real photographs from a fashion magazine and two fashion style sharing services with the task of making top (bottom) recommendations given bottom (top) photographs.

1 Introduction

Many people are interested in their fashion. They regularly check trends in fashion magazines, and refer to them when buying clothes and choosing coordinates. Fashion magazines contain a number of photographs of fashion models, and their clothing coordinates serve as useful references.

In this paper, we propose a recommender system for clothing coordinates using full-body photographs from fashion magazines. The task is that, given a photograph of a fashion item (e.g. tops) as a query, to recommend a photograph of other fashion items (e.g. bottoms) that is appropriate to the query. The proposed method can be helpful when selecting clothes to wear from one's wardrobe. Sometimes, it is difficult for a person to make a quick decision about stylish coordinates. The proposed method can also be used when purchasing clothes that match one's own clothes, and when purchasing matching clothes for the top and bottom half of the body in an online store. If the store employs experts to provide advice, we can ask them about coordinates. However, online stores have no such staff.

With the proposed method, coordinate information is learned from visual features in each fashion item region (e.g. top or bottom half of the body) using full-body photographs from fashion magazines. We use a topic model to learn coordinate information that includes the intrinsic relationship between fashion items. The proposed topic model is a probabilistic generative model of images for each fashion item region. The learned coordinate information is used to recommend coordinates. Intuitively, the proposed method finds reference photographs that are similar to the query photograph, and recommends fashion items that are similar to those in the found reference photograph. By using photographs that are taken from the user's favorite fashion magazines to train the topic model, the proposed method can recommend personalized coordinates that coincide with the user's preference.

The proposed method is novel in the sense that it learns information about coordinates automatically from a collection of reference photographs. Some fashion recommendation methods have been proposed [Nagao *et al.*, 2008; Shen *et al.*, 2007; Yonezawa and Nakatani, 2009]. For example, [Shen *et al.*, 2007] proposed a recommender system for clothing to suit a given situation using item attributes such as brands, types (e.g. jeans) and styles (e.g. formal, trendy and sporty). [Nagao *et al.*, 2008] recommends clothing relevant to the weather and the situation using annotated data. These methods require system developers and users to attach meta-data to clothes, which is a tiresome and time-consuming task. On the other hand, the proposed method does not require the attachment of meta-data; it only requires users to take photographs of their own clothes and system developers to collect reference photographs. We can easily collect reference photographs from fashion magazines, online clothes stores, or fashion photograph sharing services, such as Chicisimo[1] and Weardrobe[2].

The remainder of this paper is organized as follows. In Section 2, we present a method for detecting the top and bottom regions in full-body photographs from a wide variety of photographs. In Section 3, we propose a topic model for learning information about coordinates from the reference full-body photographs. In Section 4, we describe a coordinates recommendation method that uses the topic model. In Section 5, we demonstrate the effectiveness of the proposed method by using real photographs from a fashion magazine and two style sharing services with the task of top (bottom) recommenda-

[1] http://chicisimo.com
[2] http://www.weardrobe.com

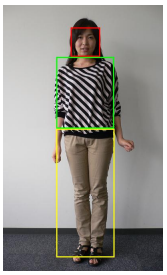


Figure 1: Example of top and bottom region detection using a face detection algorithm.

tions given bottom (top) photographs. Finally, we present concluding remarks and a discussion of future work in Section 6.

2 Top and Bottom Region Detection

The proposed method described in Sections 3 and 4 can be used to recommend arbitrary fashion items, such as tops, bottoms, hats, bags, shoes and accessories. However, this paper focuses on top (bottom) recommendations given a bottom (top) photograph, since it is the most basic and frequently occurring situation. In such cases, the proposed method requires a collection of pairs of top and bottom images to learn coordinates. This section describes a simple method of identifying full-body photographs from a collection of photographs in fashion magazines, and detecting their top and bottom regions.

In addition to full-body photographs of fashion models, fashion magazines include photographs of clothes, bags, accessories, wording and logos. Therefore, we need to select only full-body photographs from all the given photographs. We utilize a face detection algorithm [Rowley *et al.*, 1998; Viola and Jones, 2004] for this task. First, we detect the location of faces in each photograph. Next, we select full-body photographs by using the following two rules: they must contain a face region, and they must have sufficient space for a full body; the space can be calculated from the size and position of the face region. When more than one face region is detected, we use one that is the closest to the vertical center line of the photograph.

We determine the top and bottom regions by assuming that the top region is double the width and 2.5 times the height of the face region, and the bottom region is double the width and 3.5 times the height of the face region. Figure 1 shows an example of top and bottom region detection.

We can detect the top and bottom regions directly by using pose estimation instead of face detection algorithms. Some algorithms for detecting human body parts have been proposed [Ramanan, 2007; Andriluka *et al.*, 2009]. However, we utilize a face detection algorithm for simplicity, because face detection performance is superior to that for other regions, and most of the poses in fashion magazine photographs show the front view of a standing model.

Table 1: Notation

Symbol	Description
$\boldsymbol{D}$	set of reference full-body photographs
$\boldsymbol{R}$	set of regions, e.g. $\boldsymbol{R} = \{\mathrm{top}, \mathrm{bottom}\}$
r	region, $r \in \boldsymbol{R}$
V	number of unique visual features
$N_d^{(r)}$	number of features in region r of photograph d
$x_{dn}^{(r)}$	nth feature in region r of photograph d, $x_{dn}^{(r)} \in \{1, \cdots, V\}$
K	number of latent topics
$z_{dn}^{(r)}$	latent topic of the nth feature in region r of photograph d, $z_{dn}^{(r)} \in \{1, \cdots, K\}$
$\boldsymbol{\theta}_d$	topic proportions for photograph d, $\boldsymbol{\theta}_d = \{\theta_{dk}\}_{k=1}^K$, $\theta_{dk} \geq 0$, $\sum_k \theta_{dk} = 1$
$\boldsymbol{\phi}_k^{(r)}$	multinomial distribution over features for topic k in region r, $\boldsymbol{\phi}_k^{(r)} = \{\phi_{kv}^{(r)}\}_{v=1}^V$, $\phi_{kv}^{(r)} \geq 0$, $\sum_v \phi_{kv}^{(r)} = 1$

3 Topic Model for Coordinates

Let $\boldsymbol{D}$ be a set of reference photographs. Each reference photograph consists of a pair of visual features, one in the top and one in the bottom region $(\boldsymbol{x}_d^{(\mathrm{t})}, \boldsymbol{x}_d^{(\mathrm{b})})$. Here, $\boldsymbol{x}_d^{(r)}$ represents visual features in region r of photograph d. The top and bottom regions can be detected using the method described in Section 2. We can use arbitrary visual features, such as color, texture and local descriptors such as the Scale Invariant Feature Transform (SIFT) [Lowe, 2004]. We assume that $\boldsymbol{x}_d^{(r)}$ is represented by bag-of-features; i.e. $\boldsymbol{x}_d^{(r)} = \{x_{dn}^{(r)}\}_{n=1}^{N_d^{(r)}}$, where $x_{dn}^{(r)}$ is the nth visual feature in region r of photograph d. Our notation is summarized in Table 1.

We use a topic model for learning information about coordinates since the topic model can extract the intrinsic relationship between fashion items from full-body photographs. A topic model is a probabilistic generative model for discrete data such as text documents. Topics in fashion photographs, for example, represent fashion categories, such as casual, formal, elegant, classic and conservative. Topic models are successfully used for a wide variety of applications including information retrieval [Blei *et al.*, 2003; Hofmann, 1999], collaborative filtering [Hofmann, 2003; Iwata *et al.*, 2009a], and modeling images [Blei and Jordan, 2003; Cao and Fei-Fei, 2007; Iwata *et al.*, 2009b]. The proposed model is an extension of latent Dirichlet allocation (LDA) [Blei *et al.*, 2003], which is a representative topic model.

The proposed topic model can learn cooccurrence information about visual features in top and bottom regions. For example, a white top is likely to be coordinated with a dark blue bottom when color information is used for the visual features. Using the learned cooccurrence information, we recommend coordinates as described in the next section. The proposed model assumes that photograph d has its own topic proportions $\boldsymbol{\theta}_d$, which is a low dimensional intrinsic representation of the photograph. Each topic has its own visual feature prob-

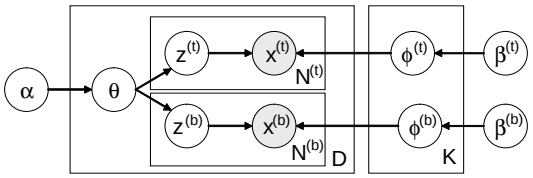


Figure 2: Graphical model representation of the proposed topic model for recommending top and bottom coordinates.

ability depending on region $\boldsymbol{\phi}_k^{(r)}$, which represents cooccurrence information. The visual feature $x_{dn}^{(r)}$ is generated according to the region dependent probabilities $\boldsymbol{\phi}_{z_{dn}^{(r)}}^{(r)}$ after topic $z_{dn}^{(r)}$ is chosen from topic proportions $\boldsymbol{\theta}_d$.

In summary, the proposed model assumes the following generative process for a set of reference photographs $\boldsymbol{X} = \{\boldsymbol{x}_d^{(r)}\}_{r,d}$,

1. For each topic $k = 1, \cdots, K$:
   (a) For each region $r \in \boldsymbol{R}$:
      i. Draw the visual feature probability for region r $\boldsymbol{\phi}_k^{(r)} \sim \mathrm{Dirichlet}(\beta^{(r)})$
2. For each reference photograph $d \in \boldsymbol{D}$:
   (a) Draw topic proportions $\boldsymbol{\theta}_d \sim \mathrm{Dirichlet}(\alpha)$
   (b) For each region $r \in \boldsymbol{R}$:
      i. For each feature in region r, $n = 1, \cdots, N_d^{(r)}$:
      A. Draw topic $z_{dn}^{(r)} \sim \mathrm{Multinomial}(\boldsymbol{\theta}_d)$
      B. Draw feature $x_{dn}^{(r)} \sim \mathrm{Multinomial}(\boldsymbol{\phi}_{z_{dn}^{(r)}}^{(r)})$

where α and $\beta^{(r)}$ are Dirichlet distribution parameters. Figure 2 shows a graphical model representation of the proposed topic model for top and bottom regions, where shaded and unshaded nodes indicate observed and latent variables, respectively. The structure of the model is the same as that of the polylingual topic model [Mimno *et al.*, 2009], which is used for extracting topics from documents in many languages.

The proposed topic model can learn coordinate information about other regions in addition to top and bottom regions, such as hats, bags, shoes and accessories. In this case, a reference photograph consists of a set of visual features with multiple regions $\{\boldsymbol{x}_d^{(r)}\}_{r \in \boldsymbol{R}}$, where e.g. $\boldsymbol{R} = \{\mathrm{top}, \mathrm{bottom}, \mathrm{hat}, \mathrm{bag}\}$.

The joint distribution of visual features and latent topics is described as follows:

$$P(\boldsymbol{X}, \boldsymbol{Z}, |\alpha, \boldsymbol{\beta}) = P(\boldsymbol{Z}|\alpha) P(\boldsymbol{X}|\boldsymbol{Z}, \boldsymbol{\beta}), \tag{1}$$

where $\boldsymbol{Z} = \{\boldsymbol{z}_d^{(r)}\}_{r,d}$ is a set of latent topics, $\boldsymbol{z}_d^{(r)} = \{z_{dn}^{(r)}\}_{n=1}^{N_d^{(r)}}$, and $\boldsymbol{\beta} = \{\beta^{(r)}\}_r$ is a set of Dirichlet parameters. We can integrate out multinomial distribution parameters, $\{\boldsymbol{\theta}_d\}_d$ and $\{\boldsymbol{\phi}_k^{(r)}\}_{k,r}$, because we use Dirichlet distributions for their priors, which are conjugate to multinomial distributions. The first factor on the right hand side of (1) is calculated as follows:

$$P(\boldsymbol{Z}|\alpha) = \left(\frac{\Gamma(K\alpha)}{\Gamma(\alpha)^K}\right)^{|\boldsymbol{D}|} \prod_d \frac{\prod_k \Gamma(N_{dk} + \alpha)}{\Gamma(N_d + K\alpha)}, \tag{2}$$

where $\Gamma(\cdot)$ is the gamma function, $|\boldsymbol{D}|$ is the size of set $\boldsymbol{D}$, N_{dk} is the number of features assigned to topic k in the photograph d and $N_d = \sum_k N_{dk}$. Similarly, the second factor is given as follows:

$$P(\boldsymbol{X}|\boldsymbol{Z}, \boldsymbol{\beta}) = \prod_r \left(\frac{\Gamma(V\beta^{(r)})}{\Gamma(\beta^{(r)})^V}\right)^K \prod_k \frac{\prod_w \Gamma(N_{kv}^{(r)} + \beta^{(r)})}{\Gamma(N_k^{(r)} + V\beta^{(r)})}, \tag{3}$$

where $N_{kv}^{(r)}$ is the number of times feature v has been assigned to topic k in region r, and $N_k^{(r)} = \sum_v N_{kv}^{(r)}$.

The inference of the latent topics $\boldsymbol{Z}$ given reference photographs can be efficiently computed using collapsed Gibbs sampling [Griffiths and Steyvers, 2004]. Given the current state of all but one variable z_j, where $j = (d, r, n)$, the assignment of a latent topic to the nth feature in region r of photograph d is sampled from the following probability:

$$P(z_j = k|\boldsymbol{D}, \boldsymbol{Z}_{\setminus j}) \propto (N_{dk\setminus j} + \alpha) \cdot \frac{N_{kx_j\setminus j}^{(r)} + \beta^{(r)}}{N_{k\setminus j}^{(r)} + V\beta^{(r)}}, \tag{4}$$

where $\setminus j$ represents the count when excluding the jth feature.

The parameters α and $\beta^{(r)}$ can be estimated by maximizing the joint distribution (1) by using the fixed-point iteration method described in [Minka, 2000] as follows:

$$\alpha \leftarrow \alpha \frac{\sum_d \sum_k \Psi(N_{kd} + \alpha) - |\boldsymbol{D}| K \Psi(\alpha)}{K\left(\sum_d \Psi(N_d + K\alpha) - |\boldsymbol{D}|\Psi(K\alpha)\right)}, \tag{5}$$

$$\beta^{(r)} \leftarrow \beta^{(r)} \frac{\sum_k \sum_v \Psi(N_{kv}^{(r)} + \beta^{(r)}) - KV\Psi(\beta^{(r)})}{V\left(\sum_k \Psi(N_k^{(r)} + V\beta^{(r)}) - K\Psi(V\beta^{(r)})\right)}, \tag{6}$$

where $\Psi(\cdot)$ is the digamma function defined by $\Psi(x) = \frac{\partial \log \Gamma(x)}{\partial x}$. By iterating Gibbs sampling with (4) and maximum likelihood estimation with (5) and (6), we can infer latent topics while optimizing the parameters.

We can estimate $\phi_{kv}^{(r)}$ by using the following equation:

$$\hat{\phi}_{kv}^{(r)} = \frac{N_{kv}^{(r)} + \beta^{(r)}}{N_k^{(r)} + V\beta^{(r)}}, \tag{7}$$

which is used for recommending coordinates as described in the next section.

4 Coordinate Recommendation based on Topic Model

In this section, we describe a method for recommending coordinates using the topic model learned with reference photographs. Intuitively speaking, we recommend a bottom (top) that has the closest topic proportions to those of the given top (bottom).

Let $\boldsymbol{C}^{(\mathrm{t})}$ and $\boldsymbol{C}^{(\mathrm{b})}$ respectively be a set of photographs of top and bottom garments that the user owns. The proposed recommendation method finds a bottom (top) photograph $\hat{c} \in \boldsymbol{C}^{(\mathrm{b})}$ ($\hat{c} \in \boldsymbol{C}^{(\mathrm{t})}$) that matches the given top (bottom) photograph $c_* \in \boldsymbol{C}^{(\mathrm{t})}$ ($c_* \in \boldsymbol{C}^{(\mathrm{b})}$). Each photograph is associated with visual features $\boldsymbol{x}_c^{(r)}, c \in \boldsymbol{C}^{(r)}$. Note that photographs in $\boldsymbol{C}^{(r)}$ need not be paired; they can be photographs of a shirt, a blouse or a skirt. We can create those data from full-body photographs of the user by using the region detection method described in Section 2. Here, we aim to recommend coordinates from among the clothes that the user owns. With recommendation in an online store, $\boldsymbol{C}^{(r)}$ can be the store's photographs.

In the following, we explain the case of a bottom recommendation given a top. We can recommend a top given a bottom in the same way. First, we estimate topic proportions $\boldsymbol{\theta}_{c_*}$ for a given top photograph and the user's own bottom photographs $\boldsymbol{C}^{(\mathrm{b})}$ using the topic model learned with reference photographs $\hat{\boldsymbol{\Phi}} = \{\hat{\boldsymbol{\phi}}_k^{(r)}\}_{k,r}$. The assignment of a latent topic z_j, where $j = (c, r, n)$, is sampled as follows:

$$P(z_j = k | \boldsymbol{x}_c^{(r)}, \boldsymbol{z}_{c\backslash j}^{(r)}, \hat{\boldsymbol{\Phi}}) \propto (N_{kc\backslash j} + \alpha) \cdot \hat{\phi}_{kx_j}, \quad (8)$$

where $\boldsymbol{z}_c^{(r)} = \{z_{cn}\}_{n=1}^{N_c^{(r)}}$ represents a set of latent topics in region r of photograph c. Then, we recommend the bottom $\hat{c}$ that has the closest topic proportions to those of the given top photograph as follows:

$$\hat{c} = \arg \max_{c \in \boldsymbol{C}^{(\mathrm{b})}} S(\boldsymbol{\theta}_{c_*}, \boldsymbol{\theta}_c), \quad (9)$$

where $S(\cdot, \cdot)$ represents a similarity function. For the similarity, we use the following negative KL divergence:

$$S(\boldsymbol{\theta}_{c_*}, \boldsymbol{\theta}_c) = -\sum_k \theta_{c_* k} \log \frac{\theta_{c_* k}}{\theta_{ck}}, \quad (10)$$

since topic proportions $\boldsymbol{\theta}_c$ are a probability distribution, and the KL divergence is one of the most commonly used distances for probability distributions. Although the recommendation of one photograph is assumed above, we can also recommend n photographs with the highest $S(\boldsymbol{\theta}_{c_*}, \boldsymbol{\theta}_c)$ values.

Figure 3 shows the framework of the proposed coordinate recommendation method. In the framework, full-body photographs for reference are converted to visual features for each region, and the proposed topic model is learned using the visual features. The query top and the user's own bottom photographs are also converted to visual features, and topic proportions of the query and the user's own photographs are estimated using the learned topic model. Then, a bottom is recommended that has the closest topic proportions to those of the query top.

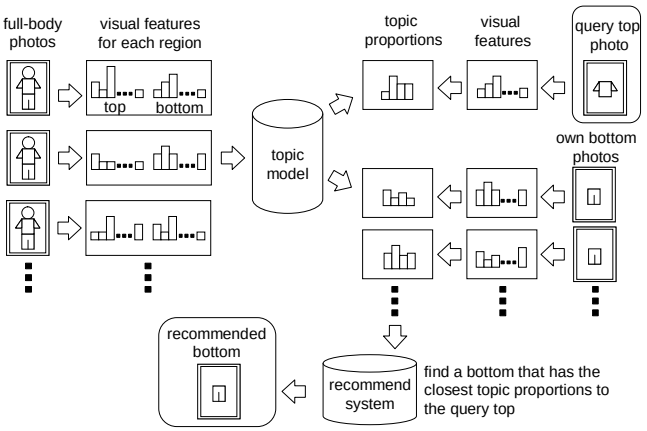


Figure 3: Framework of proposed coordinate recommendation method.

5 Experiments

5.1 Data Set

We evaluated our proposed coordinate recommender system based on the topic model using three real photograph collections: Magazine, Chicisimo and Weardrobe. The Magazine data consists of photographs taken from 32 copies of a women's fashion magazine. There were 14,813 photographs including photographs of clothes, bags, shoes, cosmetics and accessories as well as full-body fashion model pictures. Chicisimo and Weardrobe are fashion community web services for women in which users can share style photographs. The communities have a rule stating that shared photographs should show the full-body of the user. We used 1,059 and 1,410 photographs from Chicisimo and Weardrobe data, respectively.

5.2 Region Detection

In the Magazine data, 2,062 photographs out of 14,813 were estimated as full-body by the proposed method described in Section 2. We checked the results for top and bottom region extraction from the 2,062 photographs by human judgement. The regions were adequately extracted in 1,502 photographs (73%). We also checked the Chicisimo and Weardrobe data, and the regions were adequately extracted from 75%(797/1,059) and 71%(1,004/1,410) of the photographs, respectively. For the Magazine data, we checked randomly sampled 221 photographs that were determined as not being full-body using the proposed method. 81%(180/221) of the photographs were truly not full-body. Although the proposed detection method is simple, it can detect top and bottom regions with the high true positive rate, and the high true negative rate. In the following recommendation experiments, we used 1,502, 797 and 1,004 photographs respectively from the Magazine, Chicisimo and Weardrobe data that were checked by human judgement.

5.3 Baseline Recommendation Method based on Visual Feature Similarities

Before describing our experimental coordinate recommendation results, we describe a baseline method for recommendation based on visual feature similarities and compare it with the proposed topic model based method.

Assume that a photograph of a top c_* is given as a query. We would like to find a photograph of a bottom that matches a given top from a set of the user's own bottom photographs

$\boldsymbol{C}^{(\mathrm{b})}$. First, the baseline method finds the top photograph $\hat{d}$ that is closest to the given top c_* from a collection of reference photographs $\boldsymbol{D}$ as follows:

$$\hat{d} = \arg\max_{d \in \boldsymbol{D}} F(\boldsymbol{x}_{c_*}^{(\mathrm{t})}, \boldsymbol{x}_d^{(\mathrm{t})}), \tag{11}$$

where $F(\cdot, \cdot)$ represents a similarity function, For the similarity function, we used the cosine similarity in the experiment, which is a commonly used similarity function in the bag-of-features representation. Then, the baseline method recommends bottom photograph $\hat{c}$ that is the most similar to the bottom in the found reference photograph $\hat{d}$ from the user's own bottom photographs $\boldsymbol{C}^{(\mathrm{b})}$:

$$\hat{c} = \arg\max_{c \in \boldsymbol{C}^{(\mathrm{b})}} F(\boldsymbol{x}_{\hat{d}}^{(\mathrm{b})}, \boldsymbol{x}_c^{(\mathrm{b})}). \tag{12}$$

In the same way, the baseline method can recommend a matching top given a bottom.

5.4 Recommendation

We evaluated the proposed method for recommending coordinates based on a topic model with a baseline method based on the similarities between visual features.

We used a color histogram for the visual features as described in [Swain and Ballard, 1991]. Specifically, the three color axes used for the histogram are defined as follows: $rg = r - g, by = 2b - r - g, wb = r + g + b$, where r, g and b represent red, green and blue signals, respectively. The wb axis, which indicates intensity, was divided into eight sections, and the rg and by axes were divided into 16 sections, and therefore the dimension of the visual features was $V = 2,048$.

We generated 100 sets of training and test data, in which 10 % of the samples were randomly selected as test data. With the proposed method, the training data set is assumed to be a collection of reference photographs $\boldsymbol{D}$, and the parameters of the topic model are inferred using the training data. The test data are assumed to be sets of photographs $\boldsymbol{C}^{(\mathrm{t})}$ and $\boldsymbol{C}^{(\mathrm{b})}$ owned by a user. In the evaluation, each of the top (bottom) photographs in the test data is given as the input, and a matching bottom (top) photograph is recommended, while the bottom (top) photographs are held out. We assumed that the correct answer for a recommendation given a top (bottom) photograph was the bottom (top) photograph that was paired with the given photograph.

For the evaluation, we used the following two measurements: n-best accuracy and similarity. The n-best accuracy represents the rate of recommending the correct bottom (top) photograph with n recommendations given a test top (bottom) photograph. The similarity evaluation measurement is the average similarity between the recommended clothing and the held-out paired clothing.

Figure 4 shows the average accuracies and the standard deviations over 100 experiments on the three data sets. The Proposed plot represents results for the proposed method where the number of topics $K = 20$, the Baseline plot represents results for the baseline method based on the similarities between visual features, and the Random plot represents results for a method that recommends clothes randomly. Figure 5 shows the average similarities and the standard deviations with different numbers of topics. The proposed method achieved higher accuracies and similarities for all three data sets compared with the baseline and random methods. This result indicates that the proposed method can learn information about coordinates adequately using the topic model. Since the proposed method uses the similarities of extracted low dimensional intrinsic representations, it is more robust than the baseline method, which naively uses similarities between high dimensional visual features.

When we analyze the extracted topics, photographs with similar coordinates form clusters. For example, a topic is frequently assigned to photographs showing white one-piece suits, and another topic is frequently assigned to photographs showing white t-shirts and blue jeans, which correspond to formal and casual topics, respectively.

The computational time of a query response was 0.04 second, which includes estimating topic proportions of the query and sorting photographs by their similarities.

6 Conclusion

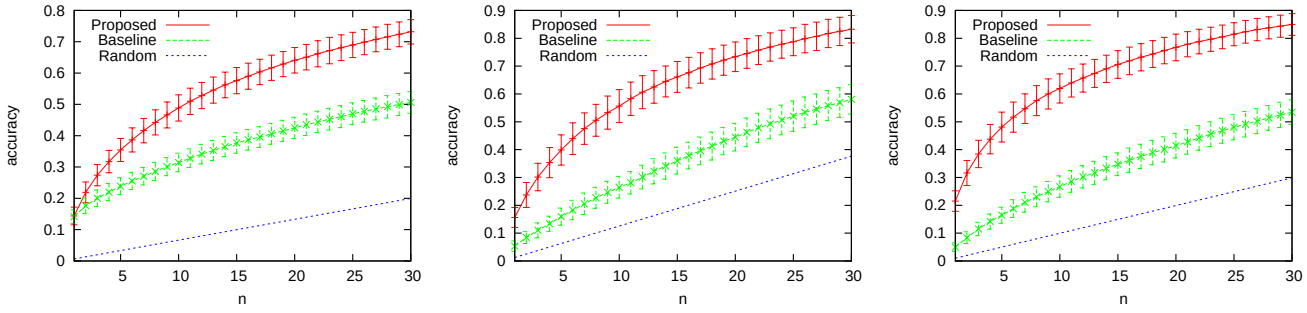
We have proposed a topic model for recommending fashion coordinates using a collection of photographs from fashion magazines. We have confirmed experimentally that the proposed method can learn information about coordinates appropriately, and can be used for recommending fashion coordinates.

Although our results have been encouraging to date, our approach can be further improved in a number of ways. First, we would like to improve the region detection method. We utilized a face detection algorithm for simplicity to detect top and bottom regions. However, top and bottom regions may not appear immediately below the face region when the fashion model leans, sits or lies down; the border of top and bottom regions differs depending on the clothes, the fashion model's style, and the angle at which the photographs were taken. By detecting regions directly using pose estimation algorithms, the method becomes applicable to photographs with a wide variety of poses. Second, the proposed topic model can be used to recommend coordinates for fashion items other than tops and bottoms, such as shoes, bags, hats and accessories. Third, we would like to investigate appropriate visual features for recommending coordinates. In our experiments, we used color information for the features. However, the texture and the shape of items may also constitute important information. Finally, it is valuable to recommend coordinates that match the user's preference and situation, such as business, dating or formal. This personalized recommendation can be achieved by using the history of the clothes the user has worn and the context.

References

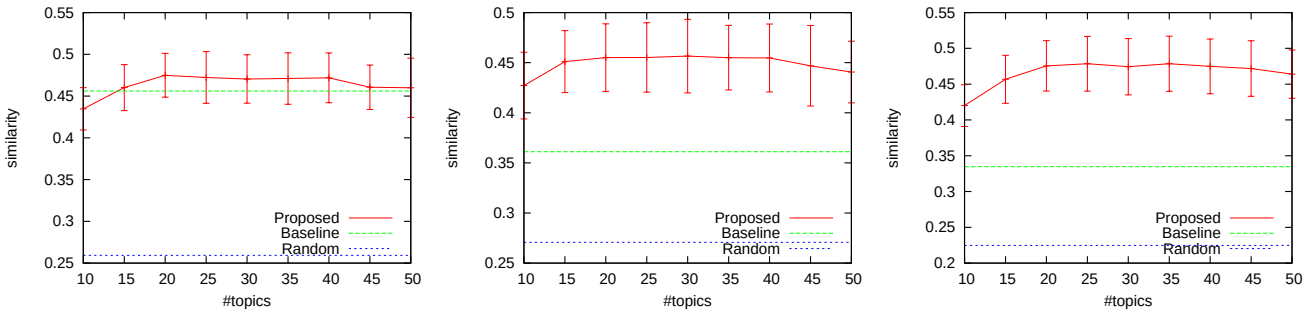
[Andriluka *et al.*, 2009] Mykhaylo Andriluka, Stefan Roth, and Bernt Schiele. Pictorial structures revisited: People detection and articulated pose estimation. In *CVPR*, pages 1014–1021, 2009.

[Blei and Jordan, 2003] David M. Blei and Michael I. Jordan. Modeling annotated data. In *SIGIR '03: Proceedings of the 26th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 127–134, 2003.



(a) Magazine (b) Chicisimo (c) Weardrobe

Figure 4: n-best accuracies with different numbers of recommendations.



(a) Magazine (b) Chicisimo (c) Weardrobe

Figure 5: Similarities with different numbers of topics.

[Blei *et al.*, 2003] David M. Blei, Andrew Y. Ng, and Michael I. Jordan. Latent Dirichlet allocation. *JMLR*, 3:993–1022, 2003.

[Cao and Fei-Fei, 2007] L. Cao and L. Fei-Fei. Spatially coherent latent topic model for concurrent object segmentation and classification. In *Proceedings of IEEE Intern. Conf. in Computer Vision (ICCV)*, 2007.

[Griffiths and Steyvers, 2004] Thomas L. Griffiths and Mark Steyvers. Finding scientific topics. *Proceedings of the National Academy of Sciences*, 101 Suppl 1:5228–5235, 2004.

[Hofmann, 1999] Thomas Hofmann. Probabilistic latent semantic analysis. In *UAI '99: Proceedings of 15th Conference on Uncertainty in Artificial Intelligence*, pages 289–296, 1999.

[Hofmann, 2003] Thomas Hofmann. Collaborative filtering via Gaussian probabilistic latent semantic analysis. In *Proceedings of the 26th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 259–266, 2003.

[Iwata *et al.*, 2009a] Tomoharu Iwata, Shinji Watanabe, Takeshi Yamada, and Naonori Ueda. Topic tracking model for analyzing consumer purchase behavior. In *IJCAI '09: Proceedings of 21st International Joint Conference on Artificial Intelligence*, pages 1427–1432, 2009.

[Iwata *et al.*, 2009b] Tomoharu Iwata, Takeshi Yamada, and Naonori Ueda. Modeling social annotation data with content relevance using a topic model. In *NIPS '09*, pages 835–843, 2009.

[Lowe, 2004] David G. Lowe. Distinctive image features from scale-invariant keypoints. *International Journal of Computer Vision*, 60(2):91–110, 2004.

[Mimno *et al.*, 2009] David Mimno, Hanna M. Wallach, Jason Naradowsky, David A. Smith, and Andrew McCallum. Polylingual topic models. In *EMNLP '09*, pages 880–889, 2009.

[Minka, 2000] Thomas Minka. Estimating a Dirichlet distribution. Technical report, M.I.T., 2000.

[Nagao *et al.*, 2008] S. Nagao, S. Takahashi, and J. Tanaka. Mirror appliance: Recommendation of clothes coordination in daily life. In *HFT '08*, pages 367–374, 2008.

[Ramanan, 2007] Deva Ramanan. Learning to parse images of articulated bodies. In *Advances in Neural Information Processing Systems*, volume 19, 2007.

[Rowley *et al.*, 1998] Henry A. Rowley, Shumeet Baluja, and Takeo Kanade. Neural network-based face detection. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 20(1):23–38, 1998.

[Shen *et al.*, 2007] Edward Shen, Henry Lieberman, and Francis Lam. What am I gonna wear?: scenario-oriented recommendation. In *IUI '07*, pages 365–368, 2007.

[Swain and Ballard, 1991] Michael J. Swain and Dana H. Ballard. Color indexing. *Int. J. Comput. Vision*, 7(1):11–32, 1991.

[Viola and Jones, 2004] Paul Viola and Michael J. Jones. Robust real-time face detection. *Int. J. Comput. Vision*, 57:137–154, May 2004.

[Yonezawa and Nakatani, 2009] Yuri Yonezawa and Yoshio Nakatani. Fashion support from clothes with characteristics. In *HCI International '09*, pages 323–330, 2009.