Stony Brook University

Understanding Online Fashion Networks

A Dissertation, Presented

by

Kota Yamaguchi

to

The Graduate School

in Partial Fulfillment of the Requirements

for the Degree of

Doctor of Philosophy

in

Computer Science

Stony Brook University

May 2014

Stony Brook University

The Graduate School

Kota Yamaguchi

We, the dissertation committee for the above candidate for the Doctor of Philosophy degree, hereby recommend acceptance of this dissertation.

Tamara L. Berg – Dissertation Advisor
Assistant Professor, Department of Computer Science, UNC at Chapel Hill

Dimitris Samaras – Chairperson of Defense
Associate Professor, Department of Computer Science, Stony Brook University

Luis E. Ortiz
Assistant Professor, Department of Computer Science, Stony Brook University

Serge J. Belongie
External Member
Professor, Department of Computer Science, Cornell University

This dissertation is accepted by the Graduate School.

Charles Taber
Dean of the Graduate School

Abstract of the Dissertation

Understanding Online Fashion Networks

by

Kota Yamaguchi

Doctor of Philosophy

in

Computer Science

Stony Brook University

2014

Emergence of online social networks has transformed how people interact with digital media. Any user is a consumer and a publisher of media content, such as texts, images, or videos, in the online community. In such an environment, it is crucial to develop technology to help people by organizing and utilizing plethora of media content to meet the community demands. To this goal, this dissertation studies and tries to establish computational approaches to understand networked content, using large-scale data from a real-world online fashion network, Chictopia. This study focuses on two major challenges of fashion networks: understanding of visual content, and understanding of user behavior.

The first component is the understanding of fashion pictures. This dissertation studies computer vision techniques to recognize garment items in a picture. This dissertation proposes clothing-parsing algorithms, which assign one of clothing category to every pixel. The algorithm takes advantage of the unique characteristics of fashion pictures that human pose gives a strong contextual cue in clothing parsing. The proposed approach considers two scenarios in clothing parsing. The first is to identify the location of items given an item list (localization scenario). This localization problem is formulated as a joint label assignment with respect to a probability distribution. The second scenario is to identify both kind of items and their locations (detection scenario). This dissertation proposes a data-driven approach to solve the difficulty in identifying clothing items. The empirical results show promising recognition performance in both scenarios, as well as the benefits of clothing parsing in human pose estimation.

The second focus of this dissertation is the analysis of user behavior in fashion networks. Specifically, this research studies the effects of visual, textual, and social factors on content popularity. The analysis makes use of the clothing parsing techniques, as well as network and text information to predict picture popularity in both in-network and out-of-network cases. The experiments find significant statistical evidence that social factors dominate the in-network scenario, but a combination of content and social factors

can help predicting popularity outside of the network.

Contents

List of Figures

List of Tables

Acknowledgements

First and foremost, I would like to thank my advisor Prof. Tamara Berg for supporting me all the time towards the completion of the dissertation research. I have learned a lot from Tamara on the importance of defining a good research question. Tamara's advising always starts from drawing a big picture that nobody in the field has ever thought about. Taking an example, it was all from this short comment that brought me to the exploration on the clothing recognition problem: "Clothing is ripe." What this really meant was her view on what interesting idea comes next, given the state-of-the-art techniques in semantic segmentation and pose estimation. Tamara lets students not necessarily think about advancing the state-of-the-art, but think *a*head of the state-of-the-art with a different view. Also Tamara always motivates students with her enthusiasm and infinite ideas to approach a problem. I believe such attitudes to research is certainly shaping my basis in thinking about a research problem. I am grateful of the countless hours Tamara spent with me for brainstorming and discussion, in person and over the Internet.

I would like to thank Prof. Luis Ortiz for co-advising me towards the degree. I still remember Luis' simple yet complete explanation on probabilistic graphical models in his course: "potential defined over cliques". I have never

heard of any better definition on graphical models, and it was one of the most exciting classes I have taken in my life. Luis always has had a different perspective on the problem we worked on, and that always reminded me of the importance of formally and clearly describing a problem so that no other scientist makes a confusion, from the small notational difference in math to the interpretation of statistics. I can confidently say that Luis taught me the good formalism in research.

I would like to thank Prof. Dimitris Samaras for serving as a chairperson of the RPE and on my dissertation committees, and also motivating students through occasional social events. I would like to thank Prof. Serge Belongie for serving on my dissertation committee. I would also like to thank Prof. Alexander Berg for giving me valuable opinions. It was an exciting experience to occasionally work with Alex.

Lastly, I would like to thank all my fellow students at Stony Brook and at Chapel Hill. Vicente Ordonez always helped me with his great sense, intuition, and knowledge about research and life. It was enjoyable to stay with the companies including Xufeng Han, Girish Kulkarni, Yifan Peng, Kiwon Yun, Hadi Kiapour, Sirion Vittayakorn, and Wei Liu. I will cherish the memories we spent together exploring eastern Long Island, climbing the mountain in Colorado Springs, or snorkeling in the Great Barrier Reef. I am looking forward to seeing the beloved fellows in New York and in a future conference.

Chapter 1

Introduction

The growth of online visual social networks, such as Facebook, Google+, Instagram, and Pinterest, to name a few, is transforming how people interact with visual media. People take a picture of their interest on a mobile device and instantly upload it to the Internet so that they can share it with friends, families, or even with the general public. In consequence, there is now a huge collection of user-generated photos and videos quickly emerging on the Web; To list a few, as of July 2013, Facebook users upload 350 million photos daily and the total number of hosted photos reaches 240 billion. Instagram users upload 40 million photos daily, 16 billion in total. Flickr hosts 8 billion photos in total. Lastly, Pinterest has gained 70 million users.

The key functionality of online visual networks is that people share a picture they find interesting by taking a certain action. For example, in Facebook, people *like* an image to implicitly show their interest to others, or explicitly *share*. In Pinterest, the action is to *pin* any image they find interesting on the Web, or even *repin* an image that somebody on their network already *pinned*.

In other words, social interaction happens because of visual media.

In face of such new visual-social interaction on the Web, it is natural to ask a fundamental question; "How can we model online visual networks?". More precisely, there arise interests in understanding what kind of images or videos people upload on the network, how people react to shared visual media, and what we can learn from such interaction. These questions further break down into what methodology we can apply to computationally and quantitatively analyze online visual networks.

This dissertation specifically focuses on *online fashion networks* as a case study of visual networks, an instance of online visual network specializing in fashion pictures, such as Chictopia, Lookbook, and Chicisimo. There are advantages in analyzing fashion networks:

Data availability It is relatively easy to collect fashion images on the Web, since fashion pictures are by nature to be shown to the public.

Data quality Fashion pictures are consistent in data quality, compared to general categories. Pictures on a fashion network mostly show a standing person with a visible full body, which is suitable for analysis.

Analysis technology Recent advancement in computer vision enabled reliable human detection, which greatly helps analyzing fashion pictures.

The broad goal of this dissertation is to establish a computational methodology to understand online visual networks. These include two different aspects: content understanding and behavior understanding. The former is a

(a) Input (b) Parsing result

Figure 1.1: Clothing-parsing problem. The goal is to detect clothing items at pixel level.

class of computer vision problem that try to recognize the content of an image. The latter is an attempt to model the behavior of users in online visual networks.

As a content understanding problem, this dissertation tackles the clothing parsing problem, where the goal is to assign a pixel-wise label of clothing items. Figure 1.1 illustrates the input and output of the parsing process. This dissertation proposes an approach to solve this challenging problem under two scenarios. The first is a localization scenario, where the goal is to locate items in a picture given what kind of items are possibly present. The other is a more challenging detection scenario, where we do not know what items are shown in the picture beforehand.

For behavior understanding, this dissertation tries to statistically analyze popularity of a picture. Using Web-crawled data from `chictopia.com`, the proposed analysis quantifies both content and social factors found on a user-

uploaded picture and try to analyze the relationships between a picture and its resulting popularity through statistical analysis. On content representation, the analysis makes use of the method taken from the clothing parsing.

This dissertation begins by reviewing relevant work in clothing recognition and social popularity analysis. The proposed clothing-parsing approach in this dissertation relies on some of the state-of-the-art computer vision methods, and Chapter 2 will go through recent efforts in the research community on clothing recognition and its applications, fundamentals of semantic image segmentation, as well as human pose estimation. Also this chapter will review the literature on computational analysis of social multimedia networks and discuss the insights and hypotheses previously proposed about popularity in networks.

Chapter 3 explores the first clothing parsing attempt in a localization scenario. The approach is probabilistically formulated as a joint labeling problem over image regions called superpixels. The key idea in this formulation is to incorporate the output of pose estimation as a condition to the parsing problem. The results show that inclusion of pose estimation contributes to the resulting parsing quality.

Chapter 4 discusses clothing parsing in a more challenging detection scenario. This dissertation proposes a data-driven approach, which is named *Paper Doll Parsing*, to overcome the difficulty of dealing with large item categories. This approach first creates a large collection of fashion images from chictopia.com. This large dataset is both used for narrowing down item categories in a picture as well as helping item localization. The results show that this data-driven approach performs significantly better in detection scenarios.

In an effort to understand user behaviors, the popularity of a fashion picture is analyzed in Chapter 5. The statistical study shows that a social network has a dominant influence on photo popularity, but not necessarily when a photo is not shared in the network. A closer look reveals that social factors are also playing some role in predicting unbiased content popularity.

Finally, Chapter 6 concludes this dissertation.

Chapter 2

Background

This chapter will review the previous work relevant to understanding of online visual networks, both in terms of content understanding and behavior understanding.

2.1 Content Understanding

Understanding visual content is the ultimate goal of computer vision research. This section will first review recent efforts in clothing recognition and related applications, fundamentals of image parsing, and pose estimation.

2.1.1 Clothing recognition

Clothing retrieval

There is a growing interest in clothing recognition, perhaps mostly due to its potential benefit in e-commerce applications. Automatic clothing recognition enables a natural and semantic image search to users of online fashion shops.

This is reflected in the increasing number of recent work in clothing recognition considering retrieval or recommendation applications [67, 66, 48, 29, 70, 25, 109].

Most notably, the work of Liu et al. [67] proposes a visual search approach to match a fashion picture taken on the street to clothing images in online shopping sites. They consider a mapping between street and shopping images with a sparsely coded transfer matrix so that the difference between these two distributions does not affect the quality of retrieval. Kalantidis et al. report in [48] a similar cross- scenario retrieval approach, where they utilize clothing parsing to explicitly represent each item. Also, Cushen et al. propose a visual search approach with efficiency in mind in a mobile scenario [25]. Retrieval of similar clothing can also help finding similar pose or person [38] in digital media. Alongside the retrieval applications, efforts are made to a create fashion- related datasets for further study, such as [111] and [68].

Recognition of clothing items constitutes the basis for analyzing or describing a fashion image. The goal in this research is to provide a fundamental technique and insight in clothing recognition to enable such applications.

Attribute recognition

When designing a retrieval system, the requirement is often not only the index of item categories but also attributes of items, such as color, pattern, or shape. Automatic identification of such attributes, namely, the attribute-recognition problem, has been the focus of several clothing recognition work [13, 22, 11, 29]. Such attribute analysis is built upon detection and localization of items in a picture, which is the result of clothing item recognition.

The idea of clothing attribute recognition dates back to the early work by Borras et al. [10], which discusses the detection of certain composition of clothing in upper-front body. More recent work of Berg et al. proposes automatic attribute discovery from shopping images using associated text description [6]. Also, Bossard et al. discusses attribute classification in noisy Web images [11]. The work of Bourdev et al. [13] reports the use of *poselet*, a discriminative image patch to capture small visual pattern in a picture, to detect clothing attributes. The approach by Chen et al. considers dependency between attributes using conditional random fields (CRF) [22]. Finally, Di et al. [29] proposes a retrieval system based on fine-grained attribute detection.

Attributes are often difficult to quantize, because there is typically no single absolute measure for them [81]. In an effort to better represent attributes, some papers propose a human-in-the-loop approach to improve supervision [23, 54].

Clothing and person identification

Another important application of clothing recognition is the identification of person by clothing. Because what he/she wears give a strong context to identify that person in a picture, several work consider person identification using clothing cue in personal photo collection [2, 39, 108], repeated shots [96], or in surveillance scenario [110, 112].

Fashion style may identify not only a person, but also the social status, occupation, and occasion. In this direction, work by [98] and [92] attempt to recognize occupation of people by clothing. Also attempts are made to identify social group based on fashion style [80, 55]. In the other way, Liu et al. proposes a system to recommend a fashion coordination by occasion [66].

Clothing parsing

The key idea in clothing parsing is to consider pose estimation and image parsing together. Human pose gives a strong context for the recognition of garment items; for example, locating shoes or boots becomes easy if we already know the location of a foot because we know that people wear shoes on foot. In clothing parsing, such pose context is explicitly taken into account in an image parsing approach. Clothing parsing has not been studied much until recently, perhaps because of the lack of some of the important algorithms to reliably recognize deforming clothing items.

There is an early work on clothing representation [21], where clothing is modeled by a grammar of sketch templates. Another attempt in clothing representation is made in [18, 43], where clothing deformation is described by a subspace approach. A probabilistic approach is made in [44], which considers a shape prior for *jacket* recognition. There is work that takes a robotics perspective [76]. These attempts are not necessarily made for general clothing parsing where the goal is to recognize large categories of items, but they give an insight into how a shape deformation should be modeled in clothing recognition.

A general clothing parsing with large clothing categories first appears in [111], where parsing is formulated as a MAP estimation of image-region labels in conditional random field (CRF) given pose estimation. Following this, Dong et al. [30] proposes clothing parsing as an inference problem over *parselet*, which is a basis group of image regions that constitute clothing items. Also, Liu et al. proposes a method to eliminate a pixel-level supervision in learning

using image-level color tags [65].

The clothing parsing framework proposed in this research differs from previous attempts in that 1) the method aims at recognition of fine-grained clothing categories without any prior information about an image, 2) the approach does not rely on an oversegmentation algorithm thus overcoming the limitation imposed by the uniformity assumption in image regions, and 3) the approach takes advantage of large, weakly annotated noisy Web images with small annotation effort by human.

2.1.2 Image parsing

Image parsing is one form of image recognition problem, where the goal is to give a semantic label to each pixel in a given image. There has been numerous attempts in solving image parsing problems in computer vision [93, 62, 42, 87, 101, 31, 31, 102, 95]. Clothing parsing is considered a domain-specific formulation of image parsing.

Image parsing is generally formulated as a joint labeling problem over pixels or local groups of pixels, which are often called superpixels, segments, or regions. In case of superpixel labeling, it is assumed that all pixels inside the same superpixel have the same semantic label. This view treats an image as a graph, where nodes correspond to a pixel or a superpixel, and edges correspond to spatial connection between them. Let us denote an image with a vector of pixels $\boldsymbol{x} \equiv [x_0, x_1, ..., x_N]$, where x_i is a pixel or a superpixel at location i. The goal is to assign corresponding semantic labels $\boldsymbol{y} \equiv [y_0, y_1, ..., y_n]$ at each location, where y_i takes a semantic label, such as *t-shirt* or *pants* in

clothing parsing, or *tree* or *sky* in case of scene parsing. The parsing problem is typically formulated as a joint *maximum a posteriori* (MAP) inference problem in a Markov random field $P(\boldsymbol{y}|\boldsymbol{x})$, which is a joint probability distribution holding a Markov property:

$$\hat{\boldsymbol{y}} \in \arg\max_{\boldsymbol{y}} P(\boldsymbol{y}|\boldsymbol{x}). \tag{2.1}$$

Often, the model $P(\boldsymbol{y}|\boldsymbol{x})$ is applied a negative-log transform to avoid numerical issue in solving the problem. The transformed model $E(\boldsymbol{y}|\boldsymbol{x}) \equiv -\ln P(\boldsymbol{y}|\boldsymbol{x})$ is called an energy function, and the parsing problem is treated as an energy minimization problem after the transform:

$$\hat{\boldsymbol{y}} \in \arg\min_{\boldsymbol{y}} E(\boldsymbol{y}|\boldsymbol{x}). \tag{2.2}$$

In practice, a second-order MRF is often employed to encourage smoothness in the final labeling, with a few exceptions utilizing higher-order potentials [83, 52]. The intuition behind this approach is that a spatially neighboring regions with similar appearance are likely to have the same semantic label; For example, two neighboring pixels that look green are both likely to be a part of *tree* but not a boundary of *sky* and *tree* in scene parsing. In a second-order model, the energy function can be expressed in the following form:

$$E(\boldsymbol{y}|\boldsymbol{x}) \equiv \sum_{i} \Phi(y_i|x_i) + \sum_{(i,j)\in V} \Psi(y_i, y_j|x_i, x_j), \tag{2.3}$$

where Φ and Ψ are potential functions and V is a set of connections between node i and j in an image. Usually, V is chosen for neighboring spatial regions

so that labeling of location i is only influenced by the labeling of its neighbor.

The first-order term (unary potential) Φ models a likelihood of assigning a certain semantic label only given its associated pixel x_i, and is the most important term in modeling the energy function. In previous work, a number of nonparametric (data-driven) approaches have been employed in an effort to better represent the likelihood function [62, 87, 101, 63, 31, 31, 102, 95]. Chapter 4 proposes a data-driven approach in clothing parsing.

The second-order term (pairwise potential) Ψ considers a likelihood of assigning two labels at once. A common form of this term enforces a consistent labeling between two adjacent pixels when they share similar appearance (smoothing prior) [86].

2.1.3 Pose estimation

Human pose estimation is another important computer vision problem, where the goal is to identify the configuration of human body in a picture. Recent progress in human pose estimation [79, 85, 84, 37, 38, 14, 113, 116, 41, 82, 27, 1, 56] is a major component in establishing the clothing parsing framework in this dissertation, since the appearance of deformable items is strongly affected by human pose.

The basic formulation of the modern pose estimation approach is as a joint labeling problem over body parts. The formulation is similar to image parsing in that this is also a joint labeling problem, but the labeling is over the location (image coordinates) of body parts or joints, such as *head* or *right arm* [84, 37, 38, 14, 113, 82]. Let us denote pose configuration with $Y \equiv$

$[y_{\text{head}}, y_{\text{neck}}, ..., y_{\text{right ankle}}]$, where $y_i \in \mathbb{R}^2$, i.e., image coordinate. Practically, y_i takes a discrete number due to the difficulty in dealing with continuous space. The pose estimation problem is then described as a MAP inference over Y (Eqn 2.1), and through the negative-log transform (Eqn 2.2), the problem is formulated as an energy minimization problem.

Like the image parsing problem, it is common to use a second-order model for pose estimation (Eqn 2.3). Such model has been sometimes called a pictorial structure [36, 34, 72], which later has been extended to a generic object detector [35] for deformable objects. In pose estimation, the unary potential considers a likelihood of placing a body part i to a certain location in a picture, and typically makes use of appearance of that location to model the likelihood. The second-order term considers a likelihood of relative displacement of two body parts, such as distance between a shoulder and an elbow. This term models kinematic constraints of human body, but connection of the body joints differs depending on the approach taken [38, 113]. Some work explicitly considers clothing in pose estimation [18, 45]. Efforts are also made to take advantage of image segmentation in pose estimation [51, 1, 56].

2.2 Behavior Understanding

There has been research interests in understanding human response to visual content in psychology [46]. This dissertation aims at understanding the behavioral model specific in an online network of visual media from a more statistical or sociological aspect. More precisely, the goal is to understand how people respond to fashion pictures that other people share on the Internet.

Social popularity

In social networks, it is common to observe the so-called *rich-get-richer* phenomenon, which suggests that the growth of connections in a social network is proportional to the current connections [5]. As a result, distribution of connections at each node in a social network follows a power-law distribution. The same setting applies to an online visual network; the popularity of visual content follows a power-law distribution. Consequently, it is hypothesized that the popularity of the content in an online visual network is largely influenced by the structure of the network but does not depend much on the content itself.

In the past, Salganik et al [88] performed controlled experiments on an artificial music market and found evidence that social influence is indeed the primary factor driving the eventual popularity of an "average" song, and popularity prediction of the "average" song based solely on content is essentially impossible in general. Chapter 5 investigates a similar social-popularity hypothesis in a visual content. Yet, this dissertation aims to *quantitatively* understand how content data contributes to the resulting popularity under social influence in a real-world social-network setting. For this purpose, this dissertation incorporates both social and content factors in the analysis based on data collected in the *wild* from an existing social network in addition to data collected under more controlled conditions.

There has also been a number of attempts to address the problem of popularity prediction of online content in social, economic, and engineering contexts. Timely prediction of content popularity would be useful for both strate-

gic marketing and social media infrastructure purposes. Work in this direction has looked into early social reaction to content and the prediction of popularity growth in videos [19, 94, 100, 78, 9, 17], news [60, 104], music [97], and discussion forums [58]. Recent work evaluates how presentation of content influences its popularity [57]. Some of the very recent work also looks into visual influence to popularity or behavior [7, 20, 4], or towards the other direction, attempts to categorize visual content by social information [99, 75, 50]. The social-popularity hypothesis is also consistent with studies of browsing behavior on Flickr [61, 103] that show social factors strongly influence which pictures are visited. This dissertation quantitatively studies the relationship between social influence, content, and popularity and additionally explore the use of computer-vision algorithms for extracting useful content-based features.

There have been also attempts to utilize user comments to recognize sentiment [90, 91]. While this dissertation focuses solely on popularity measure, it would be interesting to extend the analysis to this direction.

Chapter 5 will explore quantitative approach to popularity analysis that builds on computer vision that seeks to predict the aesthetic quality of images [28], but in addition explicitly considers social influence in the analysis and focuses on a specific fashion-social-network setting.

Visual analysis for perceptual tasks

In related tasks, some recent work has made use of visual content analysis in efforts to model subjective human perceptions of images. Applications include computational methods to model the perceptual quality of images [69] or webpages [115]. Generally these techniques extract low-level features such

as color, texture, contrast, or composition [89, 115] or higher-level attribute-based features related to perceived aesthetic quality (e.g. "follows the rule of thirds", or "contains opposing colors") [28]. In this dissertation, because the focus is on a specific type of image content – photos of people – it is possible to additionally make use of highly effective algorithms developed over many years by the vision community for localizing people and their body parts. Additionally, this dissertation proposes a high-level description of outfits specific to the fashion social-network setting.

As demonstrated by these and other research related to analyzing visual information in a social context [105, 24, 114, 47, 40], there is growing interest toward leveraging multimodal information in social network applications. Research in this domain could eventually benefit the social and behavioral sciences, and other related fields. The multimodal approach proposed in this dissertation to quantify online multimedia content is a starting point for further research.

Chapter 3

Clothing Parsing: Localization Approach

This chapter describes a clothing-parsing approach. As a starting point, this chapter considers a *localization* scenario; when we already know what kind of items can appear but not yet about where these items are in a picture. While this localization scenario is only useful for images with associated item lists, such pictures often appear in online photo-sharing websites.

The goal of localization is to find the location of all possible items by assigning one of item labels pixel by pixel. In case of fashion images, the location of clothing items is strongly correlated with human pose. Therefore, the key idea here is to consider an image-parsing problem together with a pose-estimation problem. Specifically, this chapter considers two related problems;

1. predicting a clothing parse given estimates for pose, and

2. predicting pose given estimates for clothing.

In the proposed approach, clothing-parsing is modeled as a label-assignment problem to regions (superpixels) in an image. The task is to assign a clothing label for every region, formulated as a *maximum a posteriori* (MAP) inference problem in a joint probability distribution for region labels. The method incorporates pose estimation as a condition to the probabilistic model.

Pose estimation is also a joint inference problem over a probability distribution, where the variables are locations of body joints instead of region labels. This chapter also discusses an extension to the state-of-the-art work on pose estimation [113], by considering clothing as a contextual input to pose estimation.

3.1 Fashionista Dataset

For evaluation of the clothing-parsing framework, this chapter introduces a novel dataset useful for training and testing clothing estimation techniques. This dataset consists of photographs collected from `chictopia.com`. On this website, fashionistas upload "outfit of the day" type pictures, designed to draw attention to their fashion choices or as a form of social interaction with peers. Because these are people who particularly care about their clothes they tend to display a wide range of styles, accessories, and garments. However, pictures are also often depicted in relatively simple poses (mostly standing), against relatively clean backgrounds, and without many other people in the picture. This makes it an ideal scenario to study clothing.

As a training and evaluation set, 685 photos were selected with good visibility of the full body and covering a variety of clothing items. This carefully

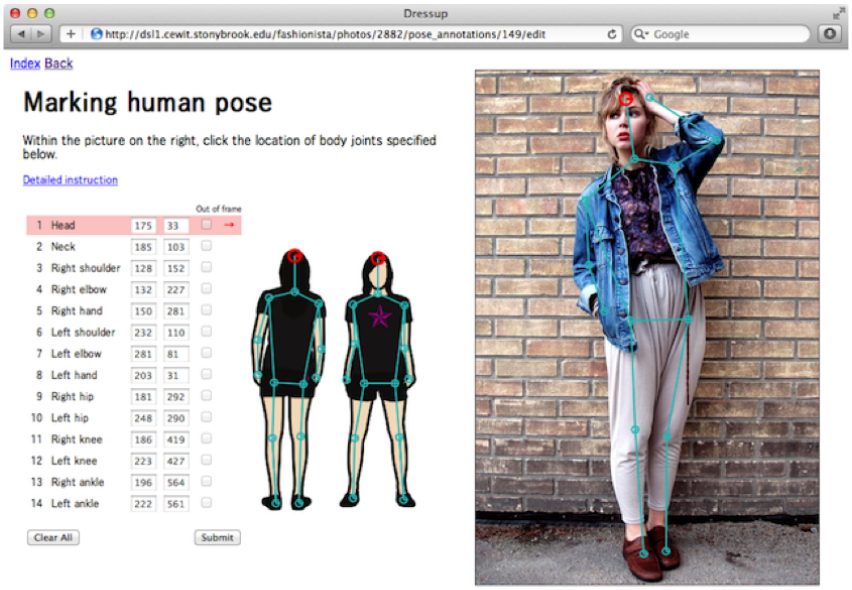


Figure 3.1: Pose-annotation tool.

selected subset were then annotated with two kinds of data in the crowd-sourcing service, Amazon Mechanical Turk. The first Turk annotation gathers ground-truth pose annotations for the usual 14 body parts [113]. The second Turk annotation gathers ground-truth clothing labels on superpixel regions. All annotations are verified and corrected if necessary to obtain high quality annotations. Figure 3.1 and 3.2 show the user interface of the annotation tools.

In this ground-truth data set, there are 53 different clothing items, of which 43 items have at least 50 image regions. Adding additional labels for *hair*, *skin*, and *null* (background), gives a total of 56 different possible clothing labels – a *much larger* number than considered in any previous approach [13, 10, 21, 112, 39, 96, 110]. On average, photos include 291.5 regions and 8.1 different clothing labels. Many common garment items have a large number of occurrences in the data set (number of regions with each label denoted in parenthesis),

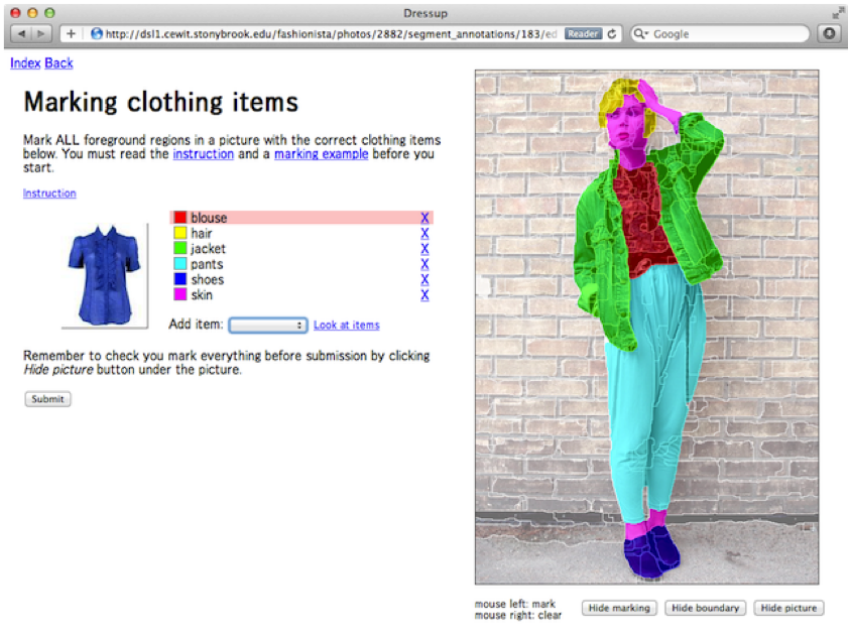


Figure 3.2: Clothing-annotation tool.

including dress (6565), bag (4431), blouse (2946), jacket (2455), skirt (2472), cardigan (1866), t-shirt (1395), boots (1348), jeans (1136), sweater (1027), etc. Table 3.1 shows 20 common clothing items and their frequency (number of segments with that label in the annotated dataset). Frequently observed items include *null* (122068), *skin* (17328), *hair* (9920), *dress* (6565), *bag* (4431), *blouse* (2946), *shoes* (2701), *top* (2543), *skirt* (2472), *jacket* (2455), and so on. However, even items probably unheard of by the fashion non-initiate, also have many occurrences – *leggings* (545), *vest* (955), *cape* (137), *jumper* (758), *wedges* (518), and *romper* (164), for example.

Label	#Region	Label	#Region
null	122068	coat	2343
skin	17328	shirt	1935
hair	9920	cardigan	1866
dress	6565	blazer	1727
bag	4431	t-shirt	1395
blouse	2946	boots	1348
shoes	2701	shorts	1149
top	2543	jeans	1136
skirt	2472	pants	1116
jacket	2455	sweater	1027

Table 3.1: 20 frequent clothing labels.

3.2 Problem Formulation

The goal of parsing is to assign a label of a clothing or null (background) to every pixel. However, the proposed approach approximates pixel labels with region labels and assumes that all pixels in the same region share the same label. This approximation reduces the computational expense of handling large number of variables associated to every pixel.

Let I denote an image showing a person. This section denotes the set of clothing labels by $L \equiv \{l_i\}$, where $i \in U$ is a region index within a set of image regions U in I, and l_i is a clothing label for region indexed by i (e.g., $l_i =$ *t-shirt* or *pants*). Also let $\boldsymbol{s}_i$ denote the set of pixels in the i-th region.

The proposed framework takes a probabilistic approach to the clothing-parsing problem. The parsing problem can be modeled as a MAP assignment of clothing labels to a probability distribution given an image $P(L|I)$. However, to take advantage of clothing and human body relationship, the framework introduces another variable, human pose configuration, and considers the distribution in terms of interactions between clothing items, human pose,

and image appearance. This section denotes a human pose configuration by $X \equiv \{\boldsymbol{x}_p\}$, which is a set of image coordinates $\boldsymbol{x}_p$ for body joints p, e.g., *head* or *right elbow*. Then, the clothing-parsing problem is formulated as two consecutive MAP inference problems: $\arg\max_X P(X|I)$ and $\arg\max_L P(L|X,I)$. Ideally, one would then like to find the joint MAP assignment over both clothing and pose labels with respect to the joint probability distribution $P(X,L|I)$ simultaneously. However, such MAP assignment problems are often computationally intractable because of the large search space and the complex structure of the probabilistic model.

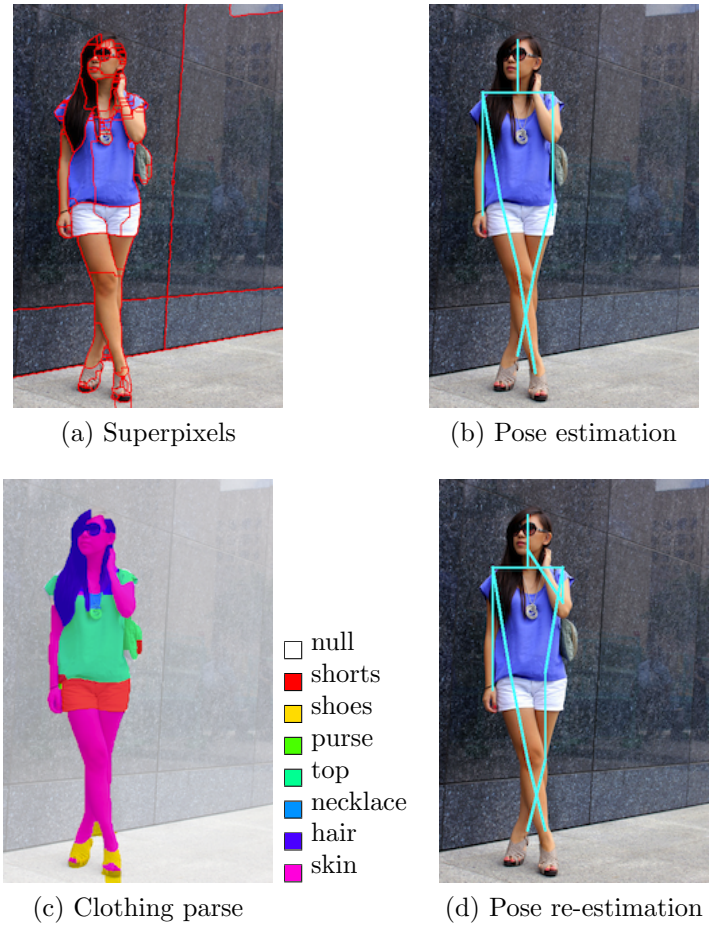
In summary, the proposed clothing-parsing pipeline proceeds as follows:

1. Obtain superpixels $\{\boldsymbol{s}_i\}$ from an image I
2. Estimate pose configuration X using $P(X|I)$
3. Assign the best clothing items L using $P(L|X,I)$
4. Optionally, re-estimate pose configuration X using model $P(X|L,I)$

Figure 3.3 shows an example of this pipeline. The following sections briefly describe each step and formally define the proposed probabilistic model.

3.2.1 Superpixels

This chapter uses the image segmentation algorithm by Arbelaendez et al. [3] to obtain superpixels. The algorithm provides a hierarchical segmentation, but experiments in this chapter set the threshold value to 0.05 to obtain a single over-segmentation for each image. This process typically yields between a few hundred to one thousand regions per image, depending on the complexity of



(a) Superpixels (b) Pose estimation

(c) Clothing parse (d) Pose re-estimation

Figure 3.3: Clothing-parsing pipeline: **(a)** Parsing the image into superpixels [3], **(b)** Original pose estimation using state of the art flexible mixtures-of-parts model [113]. **(c)** Precise clothing parse output by the proposed clothing-estimation model (note the accurate labeling of items as small as the wearer's necklace, or as intricate as her open-toed shoes). **(d)** Optional re-estimate of pose using clothing estimates (note the improvement in her left-arm prediction, compared to the original incorrect estimate down along the side of her body).

the person and background appearance (Fig 3.3(a) shows an example). This size is considerably smaller than a typical $600 \times 400 = 240,000$ pixels in the Fashionista dataset, and greatly reduces the computational expense in parsing.

3.2.2 Pose estimation

The proposed pipeline begins by estimating pose $\hat{X}$ using $P(X|I)$:

$$\hat{X} \in \arg\max_{X} P(X|I) \; . \tag{3.1}$$

For the initial pose estimate, this chapter uses the pose estimator proposed by Yang et al. [113]. In the estimation algorithm, in addition to the above terms, this estimation model includes an additional hidden variable representing a type label for pose mixture components, $T \equiv \{t_p\}$ for each body joint p. This extra variable contains information about the types of arrangements possible for a joint. Hence, the estimation problem is written as $(\hat{X}, \hat{T}) \in \arg\max_{X,T} P(X,T|I)$. The likelihood function used to evaluate pose [113] is:

$$\begin{aligned} \ln P(X,T|I) \;\; \equiv \;\; & \sum_{p} \boldsymbol{w}_p(t_p)^{\mathrm{T}} \boldsymbol{\phi}(x_p|I) \\ & + \sum_{p,q} \boldsymbol{w}_{p,q}(t_p,t_q)^{\mathrm{T}} \boldsymbol{\psi}(x_p - x_q) \\ & - \ln Z, \end{aligned} \tag{3.2}$$

where, $\boldsymbol{w}$ are the model parameters, $\boldsymbol{\phi}$ and $\boldsymbol{\psi}$ are feature functions, and Z is the partition function. Since the estimated mixture components are not used

in parsing, $\hat{T}$ is discarded and only $\hat{X}$ is kept.

3.2.3 Clothing labeling

Once the initial pose estimate $\hat{X}$ is obtained, the pipeline proceeds to the estimation of clothing labels:

$$\hat{L} \in \arg\max_{L} P(L|\hat{X}, I) \ . \tag{3.3}$$

This chapter proposes to model the probability distribution $P(L|X, I)$ with a second-order conditional random field (CRF):

$$\begin{aligned} \ln P(L|X,I) \quad \equiv \quad & \sum_{i \in U} \Phi(l_i|X,I) + \sum_{(i,j) \in V} \lambda_1 \Psi_1(l_i, l_j) + \\ & \sum_{(i,j) \in V} \lambda_2 \Psi_2(l_i, l_j|X,I) - \ln Z, \end{aligned} \tag{3.4}$$

where V is a set of neighboring pairs of image regions, λ_1 and λ_2 are model parameters, and Z is the partition function. Figure 3.4 illustrates the graphical model of the distribution.

The unary potential function Φ is modeled using the probability of a label assignment, given the feature representation of the image region $\boldsymbol{s}_i$:

$$\Phi(l_i|X, I) \equiv \ln P(l_i|\phi(\boldsymbol{s}_i, X)). \tag{3.5}$$

This chapter defines the feature vector ϕ as the concatenation of (1) normalized histograms of RGB color, and (2) normalized histogram of CIE L*a*b* color, (3) histogram of Gabor filter responses, (4) normalized 2D coordinates within

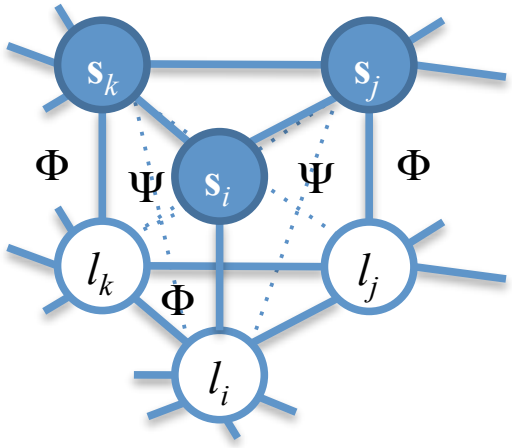


Figure 3.4: Graphical model of the clothing labeling. Each superpixel region has label l_i and pixels $\boldsymbol{s}_i$. The vertical edges represent a unary potential Φ while each 4-node vertical cycle forms a pairwise potential Ψ.

the image frame, and (5) normalized 2D coordinates with respect to each body joint location $\boldsymbol{x}_p$. The following experiments in this chapter use 11 bins for each feature type. Using a 14-joint pose estimator, this results in a 440 dimensional sparse representation for each image region. Logistic regression is experimentally chosen for the specific marginal probability model $P(l_i|\phi(\boldsymbol{s}, X))$ after evaluating a few distributions for the Fashionista dataset.

The pairwise potential function Ψ_1 is a log empirical distribution over pairs of region-labels in a single image:

$$\Psi_1(l_i, l_j) \equiv \ln \tilde{P}(l_i, l_j). \tag{3.6}$$

This term serves as a prior distribution over the pairwise co-occurrence of clothing labels (e.g. shirts are near blazers, but not shoes) in neighboring regions within an image. The function is computed by normalizing average

frequency of neighboring label pairs in training samples.

The second pairwise potential in (3.4) estimates the probability of neighboring pairs having the same label (i.e. label smoothing), given their features, ψ:

$$\Psi_2(l_i, l_j|X, I) \equiv \ln P(l_i = l_j|\psi(\boldsymbol{s}_i, \boldsymbol{s}_j, X)). \tag{3.7}$$

This chapter defines the feature transformation to be

$$\psi(\boldsymbol{s}_i, \boldsymbol{s}_j) \equiv \left[\frac{\phi(\boldsymbol{s}_i) + \phi(\boldsymbol{s}_j)}{2}, |\phi(\boldsymbol{s}_i) - \phi(\boldsymbol{s}_j)|\right]. \tag{3.8}$$

As with the unary potential, this pairwise potential uses logistic regression.

Because of the loopy structure of the graphical model, it is computationally intractable to solve (3.3) for the exact solution. Therefore, this chapter uses belief propagation to obtain an approximate MAP assignment, using the libDAI [77] implementation.

In practice, regions outside of the bounding box around pose estimation are always background. Therefore, the following experiments fixes these outside regions to *null* and runs inference only within the foreground regions.

3.2.4 Pose re-estimation

Clothing-parsing can also be useful in pose estimation, because clothing and pose are tightly coupled. The next section will have a preliminary study to answer if the original pose estimations may be improved by estimated clothing. Given the predicted clothing labels $\hat{L}$, this re-estimation model tries to

improve the prior MAP pose assignment $\hat{X}$ by computing the posterior MAP conditioned on $\hat{L}$ in (3.1):

$$\hat{X} \in \arg\max_{X} P(X|\hat{L}, I) \ . \tag{3.9}$$

The model (3.1) is modified to incorporate clothing predictions in the pose estimation process here. To do this, the appearance feature $\boldsymbol{\phi}(x_p|I)$ in (3.1) is updated to $\boldsymbol{\phi}(x_p|L, I)$, where the new appearance feature includes HOG as well as normalized histograms of clothing labels computed at the location $\boldsymbol{x}_p$.

3.2.5 Learning a model

Training of the proposed clothing parser includes parameter learning of the pose estimator $P(X|I)$ and $P(X|L, I)$, learning of potential functions in $P(L|X, I)$, and learning of CRF parameters in (3.4).

Pose estimator The training procedure of [113] uses separate negative examples, sampled from scene images to use the pose estimator as a detector. Since a localization scenario assumes that a person is always shown in a picture, the following experiments do not use a scene based negative set, but rather mine hard negative examples using false detections in the Fashionista dataset. A detection is treated as negative if less than 30% of the body parts overlap with their true locations with ratio more than 60%.

Clothing parser The probability distributions $P(l_i|\phi)$ and $P(l_i = l_j|\psi)$ in (3.5) and (3.7) are learned using logistic regression with L2 regularization [32].

The distribution given its regional features, $P(l_i|\phi)$ is learned for each possible clothing item, e.g. *shirt* or *boots*. This model is learned using a one-versus-all approach. In learning, the cost parameter is weighted by the ratio of positive to negative samples so that the resulting model does not over-fit to the prior distribution of clothing labels.

The proposed parsing-model (3.4) has two parameters λ_1 and λ_2. The best parameters are determined by maximizing cross-validation accuracy over pixels in the training data using line search and a variant of the simplex method (`fminsearch` in Matlab). In the following experiments, typically both λ_1 and λ_2 preferred small values (e.g., 0.01-0.1).

3.3 Experimental Results

The performance of the proposed approach is evaluated using 685 annotated samples from the Fashionista dataset. All measurements use 10-fold cross validation (9 folds used for training, and the remaining for testing). Since the pose estimator contains some random components, this cross validation protocol is repeated 10 times to draw an average and standard deviation.

The remainder of this section discusses quantitative (Sec 3.3.1) and qualitative (Sec 3.3.2) evaluations of the proposed clothing-parsing model, and demonstrates intriguing initial results on incorporating clothing estimates to improve pose identification (Sec 3.3.3).

Method	Pixel accuracy	Mean AGR
Final result	89.0 ± 0.8	69.6 ± 1.7
with truth	89.3 ± 0.8	71.2 ± 1.5
without pose	86.0 ± 1.0	64.6 ± 1.6
Unary only	88.2 ± 0.8	69.8 ± 1.8
Baseline	77.6 ± 0.6	12.8 ± 0.1

Table 3.2: Clothing-parsing performance with standard deviation. Results are shown for the final model (**top**), the model using unary terms only (**3rd**), and a baseline labeling (**bottom**). The results of the final model are optimized for each performance criteria.

Garment	Final result	with truth	without pose
background	95.3± 0.4	95.6± 0.4	92.5± 0.7
skin	74.6± 2.7	76.3± 2.9	78.4± 2.9
hair	76.5± 4.0	76.7± 3.9	69.8± 5.3
dress	65.8± 7.7	67.7± 9.4	50.4±10.2
bag	44.9± 8.0	47.6± 8.3	33.9± 4.7
blouse	63.6± 9.5	66.2± 9.1	52.1± 8.9
shoes	82.6± 7.2	85.0± 8.8	77.9± 6.6
top	62.0±14.7	64.6±13.1	52.0±13.8
skirt	59.4±10.4	60.6±13.2	42.8±14.5
jacket	51.8±15.2	53.3±13.5	45.8±18.6
coat	30.8±10.4	31.1± 5.1	22.5± 8.8
shirt	60.3±18.7	60.3±17.3	49.7±19.4
cardigan	39.4± 9.5	39.0±12.8	27.9± 8.7
blazer	51.8±11.2	51.7±10.8	38.4±14.2
t-shirt	63.7±14.0	64.1±12.0	55.3±12.5
boots	75.2± 6.2	80.2± 5.9	73.8± 7.3
shorts	84.6± 7.9	79.8± 9.1	63.1±11.2
jeans	80.9±10.2	85.3± 9.2	83.4±10.9
pants	78.7±12.3	81.6±11.6	71.5± 8.9
belt	71.3± 7.0	73.0± 8.9	68.6± 7.8
heels	80.9±12.3	82.6±11.9	79.3±13.1
tights	78.5±12.5	77.8±11.3	65.8±13.6
leggings	82.9±13.1	86.9± 8.5	80.7±12.9
stockings	81.0±10.9	83.3±10.9	77.2± 9.5
socks	67.4±16.1	67.8±19.0	74.2±15.0
necklace	51.3±22.5	46.5±20.1	16.2±10.7
bracelet	49.5±19.8	56.1±17.6	45.2±17.0

Table 3.3: Recall for selected garments with standard deviation.

3.3.1 Clothing parsing accuracy

Performance of clothing-parsing is measured in two ways; using average pixel accuracy, and using mean Average Garment Recall (Mean AGR). Mean AGR is measured by computing the average labeling performance (recall) of the garment items present in an image, and then the mean is computed across all images. Table 3.2 shows a comparison. Parameters of the final models are learned to optimize pixel accuracy and Mean AGR respectively. (Note that the choice of which measure to optimize for is application dependent.) Since the most frequent label present in the Fashionista dataset is *background*, the baseline in the table is a naive prediction of all regions being *background*, which already reaches **77%** accuracy. The proposed model achieves a much improved **89%** pixel accuracy, close to the result from the case when ground-truth estimates of pose are used (**89.3%**). If no pose information is used, clothing-parsing performance drops significantly (**86%**). For Mean AGR, the Unary model achieves slightly better performance (**69.8%**) over the full model because smoothing in the full model tends to suppress infrequent (small) labels.

Figure 3.5 shows a garment confusion matrix for clothing label prediction in the jet color-map. Largest confusions are observed between the *null* (background) label and clothing items, but overall good performance is observed (bright diagonal cells). Clothing items that tend to be more confused are items that occur on similar body regions, e.g. top confused with *blazer*, *cardigan*, or *t-shirt*. This implies layering produces challenges that have not fully been solved. Other typical confusions are observed at "spilled" regions nearby the body, such as a small background region next to a body, or in holes –

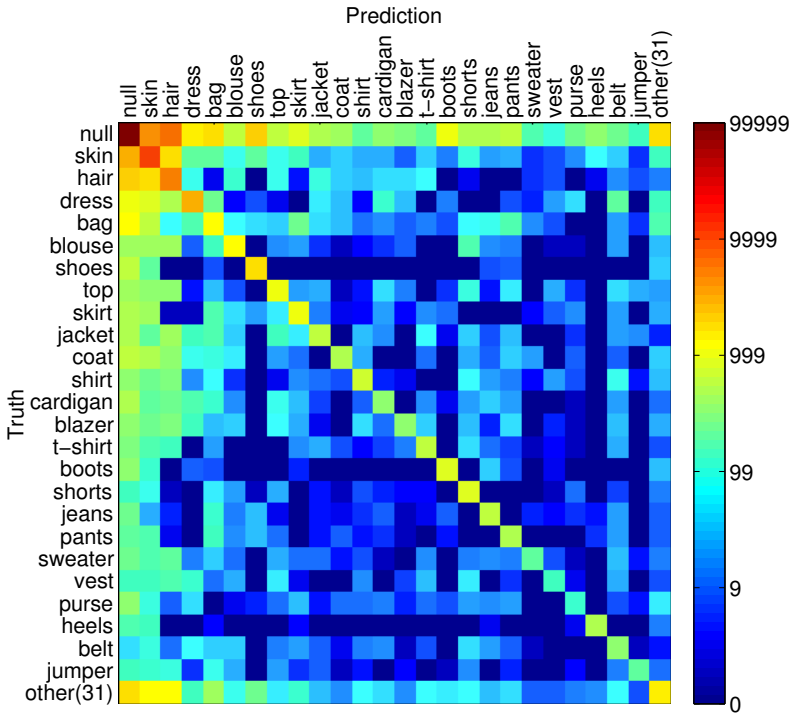


Figure 3.5: A subset of the garment confusion matrix for the 25 most commonly occurring clothing-item types.

e.g. between torso and elbow or left and right legs. Another kind of frequent confusion was skin-color objects in the picture. It was observed that light-brown garment items or wooden objects in the background were sometimes misclassified as *skin*. This color confusion were also seen between *hair* and dark garments such as *jacket* or *blazer*. Yet another issue is with superpixels that they sometimes fail to capture regions with coarse texture, e.g. the striped dress in fig 3.6b is segmented into regions for each stripe. Confusion is likely to occur in these cases since the regions will correspond to very difference appearance models.

(a) Items known (b) Items unknown

Figure 3.6: Clothing-parsing with garment meta-data (left) and without meta-data (right). Confusion between garments increases in the detection scenario, but still improves over the baseline (Table 3.2).

Detection attempt This section also reports performance of the proposed model in a detection scenario. The model was not designed for detection, but this is possible by applying the parser assuming all clothing labels are given. As seen in Fig 3.6, the full parsing problem with all 53 garment possibilities is quite challenging with this formulation, but the proposed method still obtains 80.8% pixel accuracy, a cross-validated gain of 3% over the baseline method.

3.3.2 Qualitative evaluation

This section reports the proposed clothing-parsing on all 158k un-annotated samples in the Fashionista dataset. Since there is no ground-truth labels for these photos, this section just reports qualitative observations. These results confirm that the proposed parser predicts good clothing labels on this large and varied dataset. Figure 3.7 shows some good parsing results, even han-

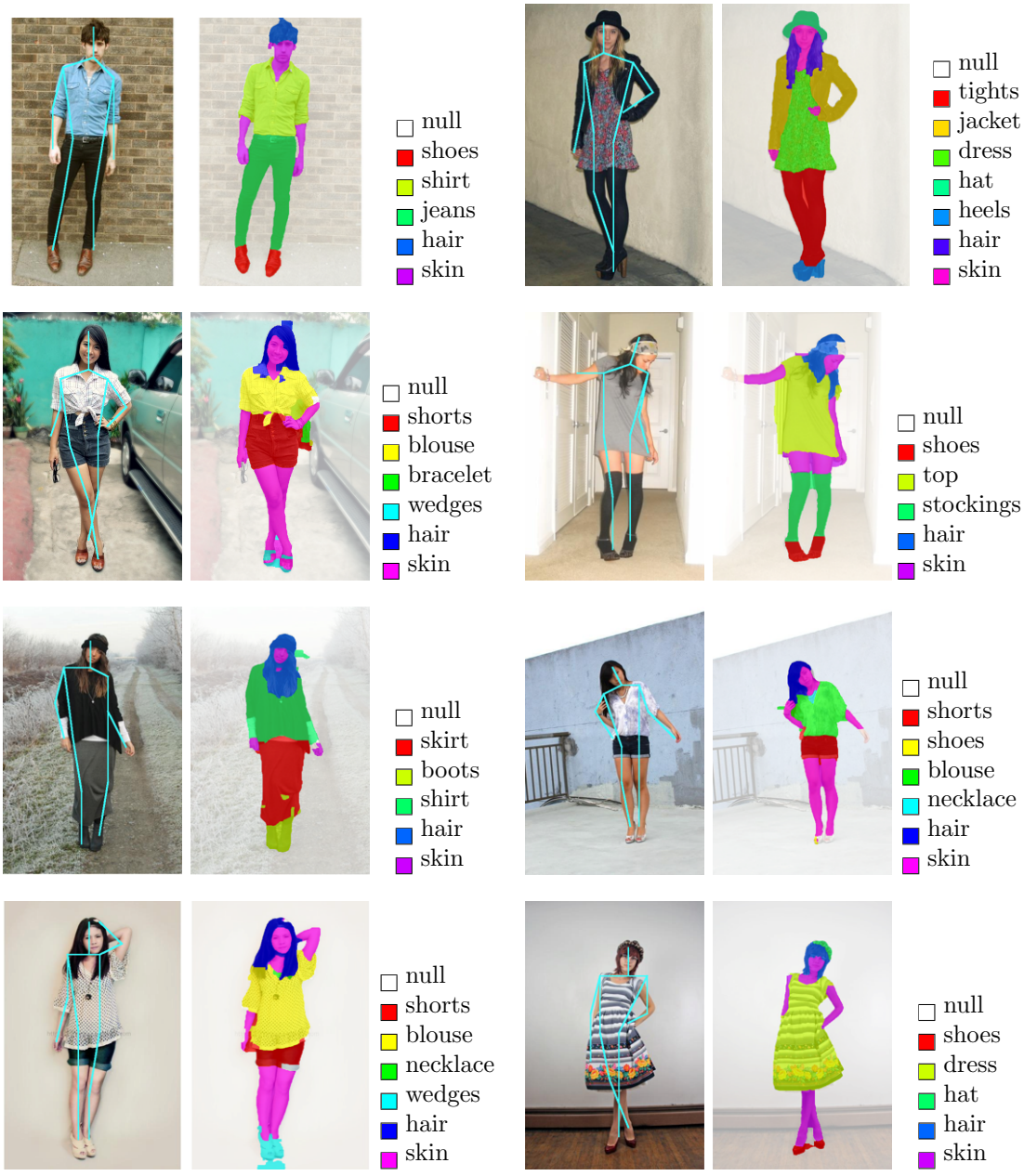
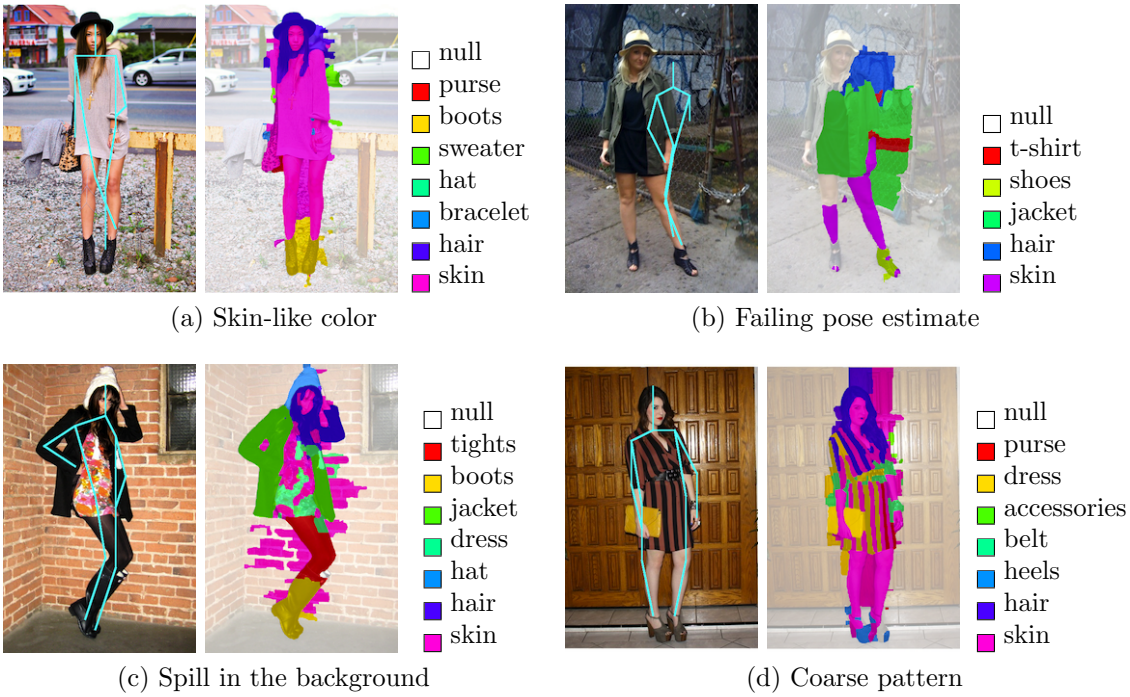


Figure 3.7: Example of successful cases.



(a) Skin-like color (b) Failing pose estimate

(c) Spill in the background (d) Coarse pattern

Figure 3.8: Example of failure cases.

dling relatively challenging clothing (e.g. small hats, and partially occluded shoes). Generally the parsing problem becomes easier in highly distinguishable appearance situations, such as on clean backgrounds, or displaying distinctive clothing regions. Failure cases (Fig 3.8) are observed due to ambiguous boundaries between foreground and background, when initial pose estimates are quite incorrect, or in the presence of very coarse patterns. Other challenges include pictures with out of frame body joints, close ups of individual garment items, or when no relevant entities appear at all.

Discussion of superpixels The proposed approach assumes that each superpixel has the same clothing label and encourages over-segmentation to make this assumption nearly true. However, in some cases the superpixel segmen-

Method	PCP
No clothing (initial)	86.5 ± 1.5
With clothing	**86.9** ± 1.4
True clothing	89.5 ± 1.5

Table 3.4: Pose-estimation performance with standard deviation. Initial state-of-the-art performance (**top** - trained and evaluated on the Fashionista dataset), the re-estimate of pose using a model incorporating predicted clothing estimates (**middle**), and pose re-estimation performance given ground-truth clothing parse (**bottom**).

Part	No clothing	With clothing	True clothing
torso	100.0 ± 0.2	99.9 ± 0.3	100.0 ± 0.1
upper left leg	94.3 ± 2.1	94.3 ± 2.3	94.3 ± 2.9
upper right leg	93.8 ± 2.4	95.3 ± 2.1	96.2 ± 2.0
lower left leg	90.8 ± 3.0	89.4 ± 3.9	90.7 ± 3.3
lower right leg	90.3 ± 3.7	93.3 ± 3.1	94.7 ± 2.7
upper left arm	86.6 ± 3.9	84.7 ± 3.8	87.7 ± 3.6
upper right arm	85.3 ± 3.4	86.6 ± 3.6	89.9 ± 3.1
lower left arm	62.8 ± 6.3	61.8 ± 5.5	70.4 ± 5.0
lower right arm	62.2 ± 6.1	64.9 ± 6.6	71.7 ± 5.9
head	99.5 ± 0.7	99.2 ± 1.1	99.5 ± 0.9

Table 3.5: Limb detection rate with standard deviation.

tation does not correctly separate regions. This is likely to occur in an image with nearly invisible boundaries, such as a black-haired person wearing a black jacket with black pants. This issue is an age old segmentation problem and very difficult to solve. The next chapter will report a pixel-wise formulation to overcome this issue.

3.3.3 Pose re-estimation accuracy

Finally, this section reports initial experiments on pose re-estimation using clothing-parsing. Pose estimation is a well-studied problem with very effective

methods [35, 14, 12, 113, 82, 56]. This experiment measures performance of pose-estimation by the probability of a correct pose (PCP) [113], which computes the percentage of body parts correctly overlapping with the ground-truth parts. Table 3.4 and 3.5 summarizes performance. Current methods [113] obtain a cross-validated PCP of **86.5%** on the Fashionista data set. Using the estimated clothing labels, the parsing-conditioned model achieves **86.9%**. As motivation for future research on clothing-parsing, the pose re-estimation system reaches a PCP of **89.5%** when true clothing labels are given, demonstrating the potential usefulness of incorporating clothing into pose identification. The next chapter will revisit this pose re-estimation evaluation with another clothing-parsing approach.

3.4 Summary

This chapter described a parsing approach to the localization scenario. The Fashionista dataset is introduced for evaluation of the proposed approach, which is collected from the online fashion network and crowdsourced annotation tool. The experimental result showed excellent performance in parsing as well as intriguing implication on using clothing estimates to improve human pose prediction. Parsing in a detection scenario was also attempted, though the performance was not satisfying with this approach. The next chapter will propose a data-driven parsing method for detection.

Chapter 4

Clothing Parsing: Detection Approach

This chapter considers clothing parsing in a detection scenario, where there is no information about clothing categories in a picture. This chapter proposes a data-driven approach to this challenging scenario, since it is known that in general non-parametric methods perform better at classification of large categories when large datasets are available.

The proposed method first collects a large, complex, real-world collection of outfit pictures from a social network focused on fashion. Using a hand-parsed dataset described in the previous chapter, it is possible to locate clothing items in this new, large image database with help from text tags associated with each image in the collection. Now, given a query image without any associated text, the proposed method retrieves similar outfits from the parsed collection. The parsed similar images are then used to detect items by combining three parsing methods: a direct parse of an image with the same parsing model,

a local parsing model that takes the specific distribution more into account in these retrieved similar images, and transferred clothing-masks from the retrieved samples to the query image. Final iterative smoothing produces the end result using a conditional random field. In each of these steps the proposed method takes advantage of the relationship between clothing and body pose to constrain prediction and produce a more accurate parse. This dissertation calls the proposed method *Paper Doll parsing*, because it essentially transfers predictions from retrieved samples to the query, like laying paper cutouts of clothing items onto a paper doll.

This approach uses an over-segmentation algorithm in part of the parsing pipeline and does not rely on superpixels for the final parsing. As discussed in Section 3.3.2, the per-pixel approach does not suffer from the irrecoverable error produced by the superpixel approximation. Although this increases the computational cost, it is empirically observed that the implementation can achieve the same level of computational time.

Given a new image to parse, the proposed approach consists of two major steps:

1. Retrieve similar images from the parsed database.

2. Use retrieved images and tags to parse the query.

Figure 4.1 depicts the overall parsing pipeline. In particular, in the second parsing stage, this dissertation proposes a retrieval-based approach to clothing parsing that combines;

- pre-trained global models of clothing items,

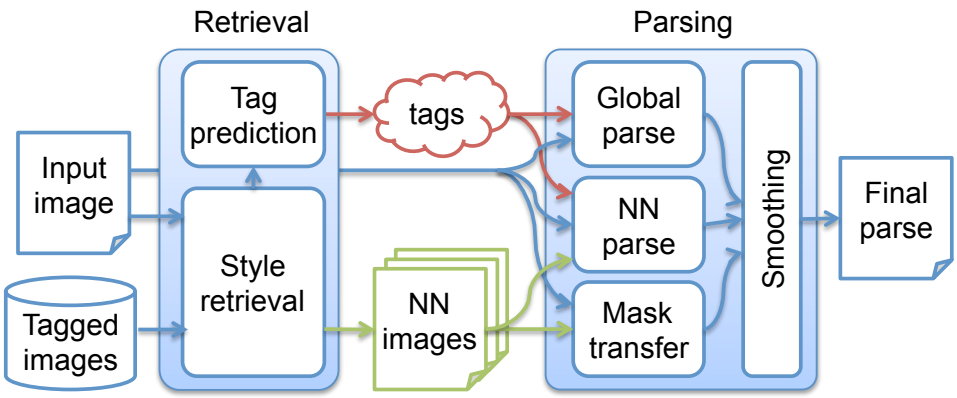


Figure 4.1: Data-driven parsing pipeline.

- local models of clothing items learned on the fly from retrieved examples,
- parse mask predictions transferred from retrieved examples to the query image, and
- iterative label smoothing.

4.1 Paper Doll Dataset

This chapter uses the Fashionista dataset from the previous chapter and a new, weakly-supervised set of images named the Paper Doll dataset. This chapter uses the Fashionista dataset for supervised training and performance evaluation, 456 for training and 229 for testing. The training samples are used for learning feature transforms, building global clothing models, and adjusting parameters. The testing samples are reserved for evaluation.

The Paper Doll dataset is a large collection of tagged fashion pictures from `chictopia.com`. Over 1 million pictures are downloaded with associated meta-data tags denoting characteristics such as color, clothing item, or occasion.

Since the Fashionista dataset also uses Chictopia, the data-collection process automatically excludes any duplicated pictures from the Paper Doll dataset. From the remaining, the data-collection process selects pictures tagged with at least one clothing item and runs a full-body pose detector [113] on them, keeping those that have a person detected. This process results in 339,797 pictures weakly annotated with clothing items and estimated pose in the following experiments. Though the annotations are not always complete – users often do not label all depicted items, especially small items or accessories – it is rare to find images where an annotated tag is not present. The proposed parsing-approach uses the Paper Doll dataset for style retrieval.

4.2 Low-level Features

This section details low-level image features used in the proposed parsing method.

For a new image, the pre-processing step first runs a pose estimator [113] and normalizes the full-body bounding box to a fixed size. The pose estimator is trained using the Fashionista training split and negative samples from the INRIA dataset. During parsing in the later steps, all computations are done in this fixed frame size, and warped back to the original image afterward, assuming regions outside the bounding box are background.

The proposed methods draw from a number of dense feature types (each parsing method uses some subset):

RGB : RGB color of the pixel.

Lab : L*a*b* color of the pixel.

MR8 : Maximum Response Filters [106].

Gradients : Image gradients at the pixel.

HOG : HOG descriptor at the pixel [26].

Boundary Distance : Negative log-distance from the boundary of an image.

Pose Distance : Negative log-distance from 14 body joints and any body limbs.

Skin-hair Detection : Likelihood of skin, hair, other foreground, and background of the pixel.

Whenever a statistical model built upon these features is used in this chapter, the pre-processing step first normalizes features by subtracting their mean and dividing by their 3 standard deviation for each dimension. Also, when logistic regression is used in this chapter, regression models use these normalized features and their squares, along with a constant bias. So, for an N-dimensional feature vector, logistic regression always gets $2N + 1$ parameters.

4.3 Style Retrieval

The proposed parsing algorithm starts by retrieving similar images in the Paper Doll dataset for an input. The purpose of retrieving similar pictures is two-fold:

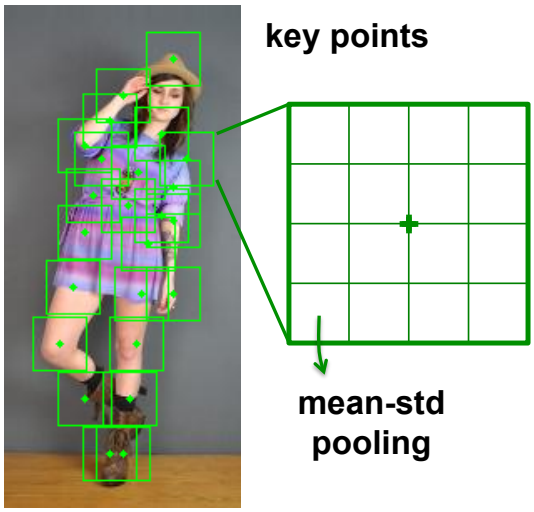


Figure 4.2: Spatial descriptors for style representation.

1. to predict depicted clothing items, and

2. to obtain information helpful for parsing clothing items.

4.3.1 Style descriptor

This section describes the design of a comprehensive fashion image descriptor, which is named Style Descriptor, specifically useful for finding styles with similar appearance. For this purpose, the style descriptor is built upon pose estimation, and reflects information about the items a person is wearing, their arrangement on the body, and their appearance.

The style descriptor is computed as follows.

1. Apply a pose estimator [113] to obtain estimates for 24 body part locations (centered around head, torso, joints, etc.).

2. For local image regions around each detected body part, calculate a vector of the following features at each pixel: RGB, Lab, MR8, HOG,

Boundary Distance, and Skin-hair Detection.

3. Calculate a 4×4 grid of mean-std pooling of the above features. That is, an image patch of 32×32 pixels around the body part location is extracted first, and this patch is split into 4×4 cells. The mean and standard deviation of features are calculated within each of the cell. The result is concatenated to form a single vector.

4. Concatenate the pooled features and apply PCA to reduce dimensionality.

Skin-hair Detection is computed using generalized logistic regression for *skin*, *hair*, *background*, and *clothing* at each pixel. this logistic regression uses RGB, Lab, MR8, HOG, Boundary Distance, and Pose Distance for input. The logistic regression is learned from the Fashionista dataset using one-vs-all approach. Note that the style descriptor does not include Pose Distance in the low-level features, but instead uses Skin-hair Detection to indirectly include pose-dependent information in the representationThis is because the purpose of the style descriptor is to find similar styles independent of pose.

Figure 4.2 illustrates the process of extracting style descriptors. The above process resulted in a 441 dimensional representation for each fashion picture in the experiments in this Chapter. This chapter use the Fashionista training split to build the Skin-hair detector and also to train the PCA model. Including pose estimation, it takes 3-4 seconds to calculate the descriptor in the implementation.

4.3.2 Retrieval

The retrieval of similar styles uses L2-distance over the style descriptors to find the K nearest-neighbors (KNN) in the Paper Doll dataset. For efficiency, the retrieval system builds a KD-tree [107] to index samples. In this chapter, the retrieval size is fixed to $K = 25$ for all the experiments. Figure 4.3 shows two examples of nearest-neighbor (NN) retrievals.

4.3.3 Tag prediction

The retrieved samples are first used to predict clothing items potentially present in a query image. The purpose of tag prediction is to obtain a set of tags that might be relevant to the query, while eliminating definitely irrelevant items for consideration. Later stages can remove spuriously predicted tags, but tags removed at this stage can never be predicted. Therefore, it is preferable to obtain the best possible predictive performance in the high recall regime.

Tag-prediction is based on a simple voting approach from KNN. Each tag in the retrieved samples provides a vote weighted by the inverse of its distance from the query, which forms a confidence for presence of that item. The prediction is made by thresholding this confidence.

This simple KNN prediction is experimentally selected among other models because it turns out KNN works well for the high-recall prediction task. Figure 4.4 shows performance of linear vs KNN with retrieval size at 10 and 25. While linear classification (clothing item classifiers trained on subsets of body parts, e.g. *pants* on lower body keypoints), works well in the low-recall

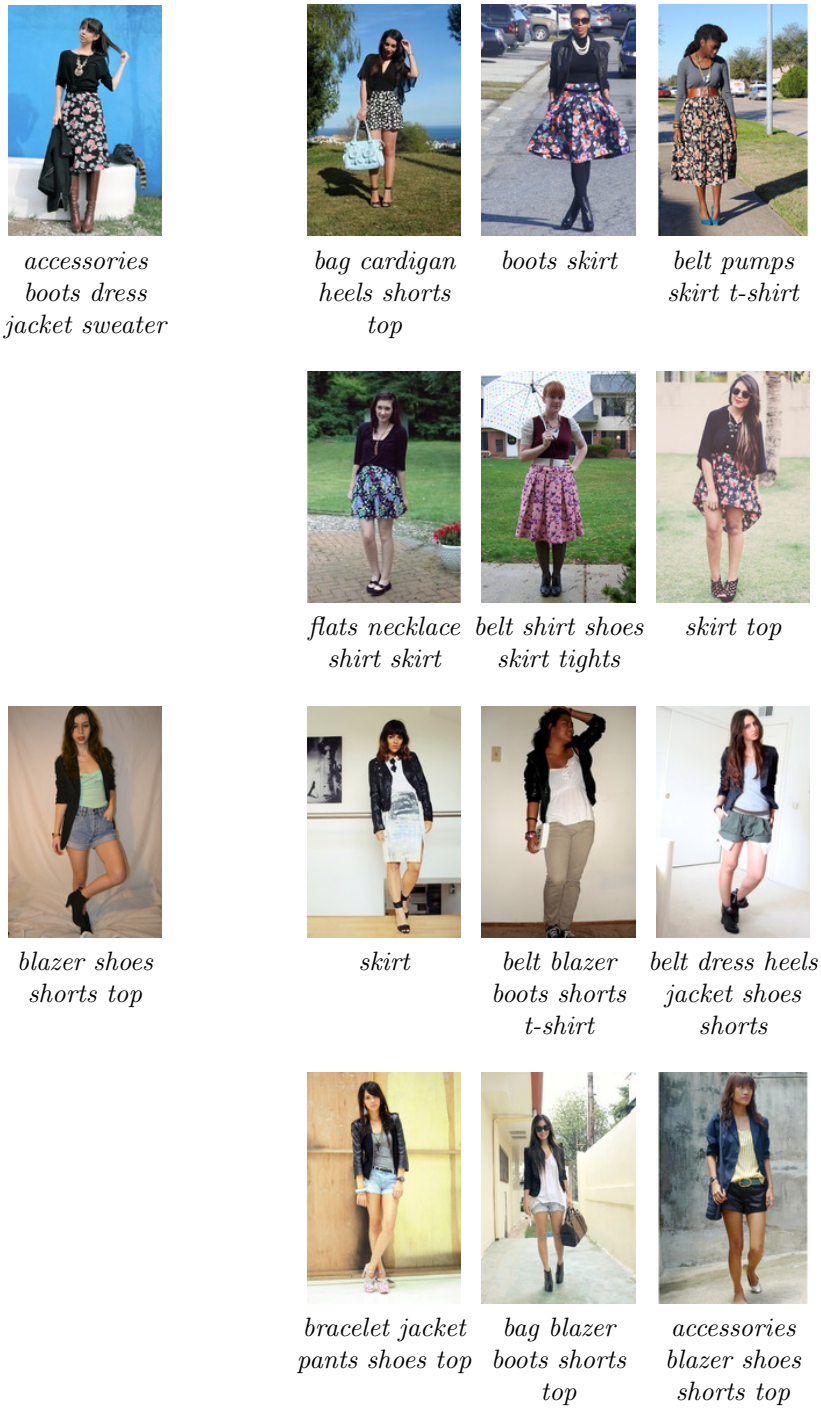


Figure 4.3: Retrieval examples. The leftmost column shows query images with ground-truth item annotation. The rest are retrieved images with associated tags in the top 25. Notice retrieved samples sometimes have missing item tags.

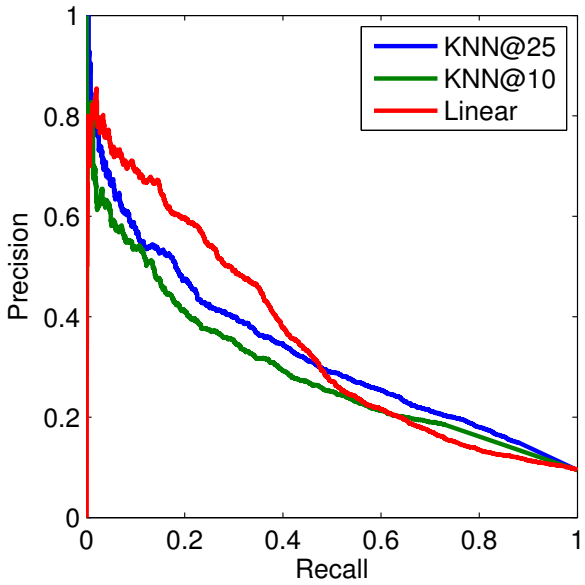


Figure 4.4: PR-plot of tag-prediction.

high-precision regime, KNN outperforms the linear model in the high-recall range. KNN at 25 retrievals also outperforms 10. The effect of retrieval size in parsing is discussed more in Section 4.5.1.

Since the goal here is only to eliminate obviously irrelevant items while keeping most potentially relevant items, the threshold is tuned to give 0.5 recall in the Fashionista training split. Due to the skewed item distribution in the Fashionista dataset, the same threshold is applied to all items to avoid overfitting the predictive model. After tag-prediction, the parsing stage always includes *background*, *skin*, and *hair*, in addition to the predicted clothing tags.

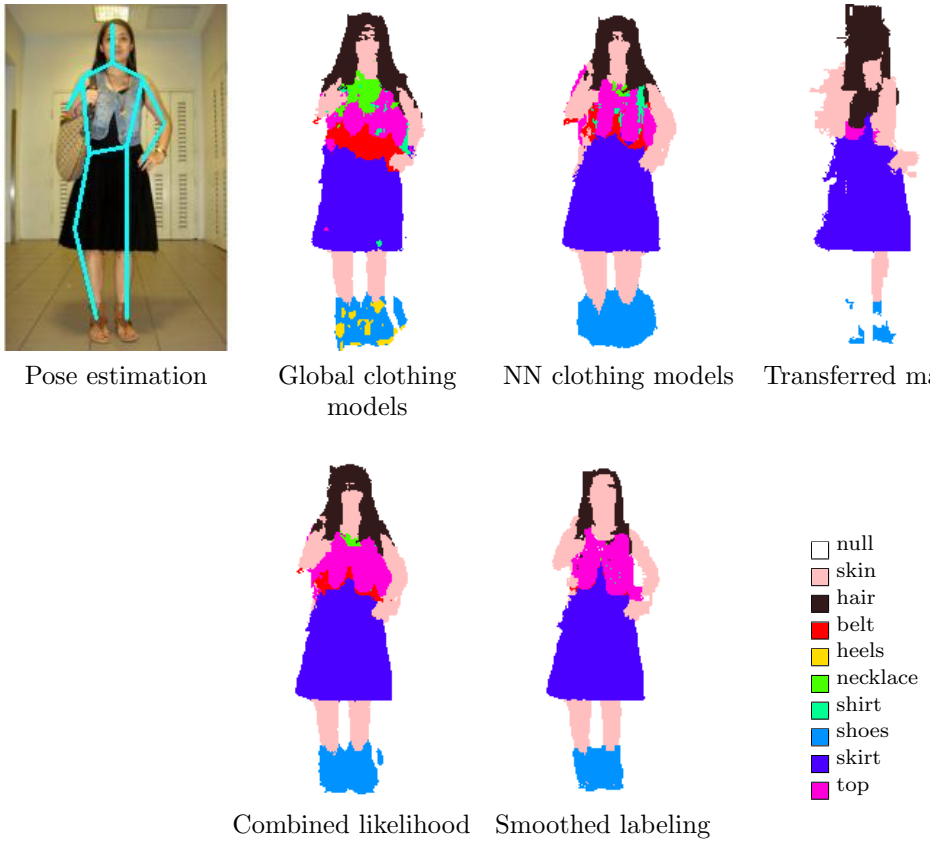


Figure 4.5: Parsing outputs at each step. The labels are the MAP assignments of the scoring functions.

4.4 Clothing Parsing

Following tag-prediction, the Paper Doll pipeline proceeds to parse a query image in a per-pixel fashion. Parsing has two major phases:

1. Compute pixel-level likelihood from three methods: global clothing models, nearest-neighbor clothing models, and soft-mask transfer.

2. Apply iterative label smoothing to get a final parse.

Figure 4.5 illustrates outputs from each parsing stage.

4.4.1 Pixel likelihood

Let us denote y_i as the clothing item label at pixel i. The first step in parsing is to compute likelihood of assigning clothing item l to y_i. This likelihood function S is modeled as a mixture of three functions.

$$\begin{aligned} S(y_i|\mathbf{x}_i, D) \equiv\ & S_{\text{global}}(y_i|\mathbf{x}_i, D)^{\lambda_1} \cdot \\ & S_{\text{nearest}}(y_i|\mathbf{x}_i, D)^{\lambda_2} \cdot \\ & S_{\text{transfer}}(y_i|\mathbf{x}_i, D)^{\lambda_3}, \end{aligned} \tag{4.1}$$

where $\mathbf{x}_i$ denotes pixel features, $\Lambda \equiv [\lambda_1, \lambda_2, \lambda_3]$ are mixing parameters, and D is a set of nearest-neighbor samples.

Global parse

The first term in the model is a global clothing likelihood, trained for each clothing item on the hand-parsed Fashionista training split. This is modeled as logistic regression that computes a likelihood of a label assignment to each pixel for a given set of possible clothing items:

$$S_{\text{global}}(y_i|\mathbf{x}_i, D) \equiv P(y_i = l|\mathbf{x}_i, \theta_l^g) \cdot \mathbf{1}[l \in \tau(D)], \tag{4.2}$$

where P is logistic regression given feature $\mathbf{x}_i$ and model parameter θ_l^g, $\mathbf{1}[\cdot]$ is an indicator function, and $\tau(D)$ is a set of predicted tags from nearest-neighbor retrieval. This logistic regression uses RGB, Lab, MR8, HOG, and Pose Distances as features. Any unpredicted items receive zero probability.

The model parameter θ_l^g is trained on the Fashionista training split. To

train each θ_l^g, negative pixel-samples are selected only from those images having at least one positive pixel. That is, the model gives localization probability given that a label l is present in the picture. This could potentially increase confusion between similar item types, such as *blazer* and *jacket* since they usually do not appear together, in favor of better localization accuracy. This approach relies on the tag-prediction τ to resolve such confusion.

Because of the tremendous number of pixels in the dataset, pixels are subsampled in the training of each logistic regression. The subsampling method tries to draw pixel-samples so that the resulting label distribution is close to uniform in each image, preventing learned models from only predicting large items.

Nearest-neighbor parse

The second term in the model is also logistic regression, but trained only on the retrieved nearest-neighbor images. Unlike the global model, the NN model is trained on examples that are similar to the query, e.g. *blazers* that look similar to the query blazer because they were retrieved via style similarity. These local models are considerably better models for the query image than those trained globally (because *blazers* in general can take on a huge range of appearances). The model is defined:

$$S_{\text{nearest}}(y_i|\mathbf{x}_i, D) \equiv P(y_i = l|\mathbf{x}_i, \theta_l^n) \cdot \mathbf{1}[l \in \tau(D)]. \tag{4.3}$$

The model parameter θ_l^n is locally learned from the retrieved samples D, using RGB, Lab, Gradient, MR8, Boundary Distance, and Pose Distance.

In this step, predicted pixel-level annotations from the retrieved samples are used (computed during pre-processing detailed in Section 4.4.3) to learn local appearance models. NN models are trained using any pixel (with sub-sampling) in the retrieved samples in a one-vs-all fashion.

Transferred parse

The third term in the parsing-likelihood is obtained by transferring the likelihoods estimated by the global parse S_{global} from the retrieved images to the query image (Figure 4.6 visualizes an example). This approach is similar in spirit to approaches for general segmentation that transfer likelihoods using over-segmentation and matching [8, 59, 62, 74], but here a parsing-algorithm can take advantage of pose estimation during transfer because segmentation is performed on human body.

This approach finds dense correspondence based on superpixels instead of pixels (e.g., [101]) to overcome the difficulty in naively transferring deformable, often occluded clothing items pixel-wise. The approach first computes an over-segmentation of both query and retrieved images using a fast and simple segmentation algorithm [33], then finds corresponding pairs of super-pixels between the query and each retrieved image based on pose and appearance:

1. For each super-pixel in the query, find the 5 nearest super-pixels in each retrieved image using L2 Pose Distance.

2. Compute a concatenation of bag-of-words from RGB, Lab, MR8, and Gradient for each of those super-pixels.

3. Pick the closest super-pixel from each retrieved image using L2 distance

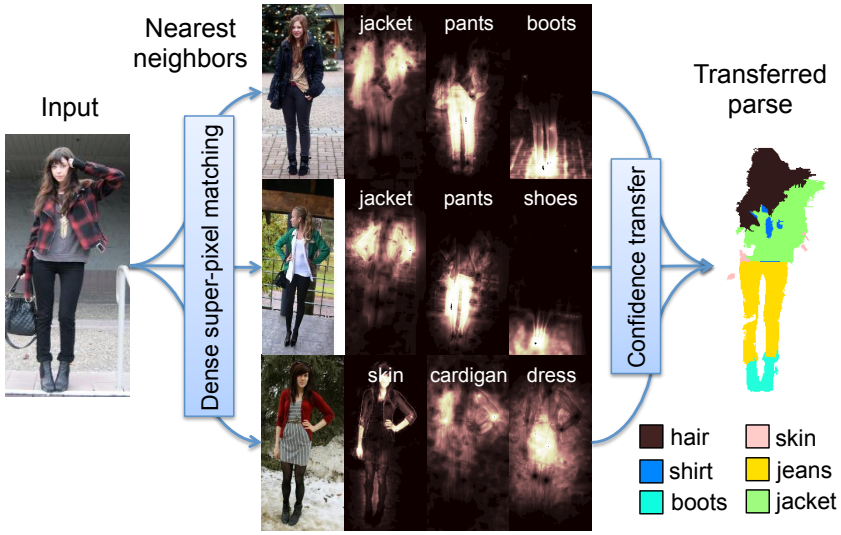


Figure 4.6: Transferred parse. Likelihoods in nearest-neighbors are transferred to the input via dense matching.

on the bag-of-words feature.

Denoting the super-pixel of pixel i by s_i, the selected corresponding super-pixel from image r by $s_{i,r}$, and the bag-of-words features of super-pixel s by $h(s)$, the transferred parse is expressed by:

$$S_{\text{transfer}}(y_i|\mathbf{x}_i, D) \equiv \frac{1}{Z}\sum_{r\in D}\frac{M(y_i, s_{i,r})}{1+\|h(s_i)-h(s_{i,r})\|}, \tag{4.4}$$

where $M(y_i, s_{i,r})$ is defined:

$$M(y_i, s_{i,r}) \equiv \frac{1}{|s_{i,r}|}\sum_{j\in s_{i,r}} P(y_i = l|\mathbf{x}_i, \theta_l^g)\cdot \mathbf{1}[l\in\tau(r)], \tag{4.5}$$

which is a mean of the global parse over the super-pixel in a retrieved image. Here a set of tags of image r is denoted by $\tau(r)$, and Z is the normalization constant.

Combined likelihood

After computing the three likelihoods, they are combined according to parameter Λ to get the final pixel likelihood S as described in Equation 4.1. The best mixing parameter is chosen such that the MAP assignment of pixel labels gives the best foreground accuracy in the Fashionista training split, by solving the following optimization (on foreground pixels F):

$$\max_{\Lambda} \sum_{i \in F} \mathbf{1} \left[\tilde{y}_i = \arg\max_{y_i} S_{\Lambda}(y_i | \mathbf{x}_i) \right], \tag{4.6}$$

where $\tilde{y}_i$ is the ground-truth annotation of pixel i. The nearest-neighbors D in S are dropped in the notation for simplicity. A simplex-search algorithm is employed to solve for the optimum parameter starting from uniform values. In the experiment, the obtained result was $(0.41, 0.18, 0.39)$.

This optimization excludes background pixels because of the skew in the label distribution – background pixels in Fashionista dataset represent 77% of total pixels, which tends to direct the optimizer to find meaningless local optima; i.e., predicting everything as *background*.

4.4.2 Iterative label smoothing

The combined likelihood gives a rough estimate of item localization. However, it does not respect boundaries of actual clothing items since it is computed per-pixel and ignores any relationship between spatially local neighbors. Therefore, the next step introduces an iterative smoothing approach that considers all pixels together to provide a smooth parse of an image. Following the approach

of [93], this smoothing approach is formulated by considering the joint labeling of pixels $Y \equiv \{y_i\}$ and item appearance models $\Theta \equiv \{\theta_l^s\}$ in a conditional random field, where θ_l^s is a model for a label l. The goal is to find the optimal joint assignment Y^* and item models Θ^* for a given image.

The smoothing approach starts by initializing the current predicted parsing $\hat{Y}_0$ with the MAP assignment under the combined likelihood S. Then, the approach treats $\hat{Y}_0$ as training data to build initial image-specific item models $\hat{\Theta}_0$ (from logistic regressions). These models only use RGB, Lab, and Boundary Distance since otherwise the models easily over-fit. Also, the models use a higher regularization parameter for training instead of finding the best cross-validation parameter, assuming the initial training labels $\hat{Y}_0$ are noisy.

After obtaining $\hat{Y}_0$ and $\hat{\Theta}_0$, the smoothing method solves for the optimal assignment $\hat{Y}_t$ at the current step t with the following optimization problem:

$$\hat{Y}_t \in \arg\max_Y \prod_i \Phi(y_i|\mathbf{x}_i, S, \hat{\Theta}_t) \prod_{i,j \in V} \Psi(y_i, y_j|\mathbf{x}_i, \mathbf{x}_j), \tag{4.7}$$

where the potential functions are defined:

$$\Phi(y_i|\mathbf{x}_i, S, \hat{\Theta}_t) \equiv \quad S(y_i|\mathbf{x}_i)^\lambda \cdot P(y_i|\mathbf{x}_i, \theta_l^s)^{1-\lambda}, \tag{4.8}$$

$$\Psi(y_i, y_j|\mathbf{x}_i, \mathbf{x}_j) \equiv \quad \exp\left\{\gamma e^{-\beta\|\mathbf{x}_i - \mathbf{x}_j\|^2} \cdot \mathbf{1}\left[y_i \neq y_j\right]\right\}. \tag{4.9}$$

Here, V is a set of neighboring pixel pairs, λ, β, γ are the parameters of the model, which is experimentally determined. This method uses the graph-cut algorithm [16, 15, 53] to find the optimal solution.

With the updated estimate of the labels $\hat{Y}_t$, the smoothing approach trains

logistic regressions $\hat{\Theta}_t$ and repeat each step. Note that this iterative approach is not guaranteed to converge. Therefore, in the implementation, the iteration terminates either when 10 iterations pass, when the number of changes in label assignment is less than 100, or the ratio of the change is smaller than 5%.

4.4.3 Offline processing

The retrieval techniques require the large Paper Doll Dataset to be pre-processed (parsed), for building nearest-neighbor models on the fly from retrieved samples and for transferring parse masks. Therefore, each sample in the Paper Doll dataset is parsed beforehand using pose estimation and the tags associated with the image by the photo owner. This parse makes use of the global clothing models (constrained to the tags associated with the image by the photo owner) and the iterative smoothing.

Although these training images are tagged, there are often clothing items missing in the annotation. This will lead iterative smoothing to mark foreground regions as *background*. To prevent this, this pre-processing step adds an *unknown* item label with uniform probability and initialize $\hat{Y}_0$ together with the global clothing model at all samples. This effectively prevents the final estimated labeling $\hat{Y}$ to mark missing items with incorrect labels.

Offline processing of the Paper Doll Dataset took a few of days with the Matlab implementation in a distributed environment. For an unseen query image, the full parsing pipeline takes 20 to 40 seconds, including pose estimation. The major computational bottlenecks are in nearest-neighbor parse and in iterative smoothing.

4.5 Experimental Results

Parsing performance is evaluated on the 229 testing samples from the Fashionista dataset. The task is to predict a label for every pixel where labels represent a set of 56 different categories – a very large and challenging variety of clothing items.

Performance is measured in terms of standard metrics: accuracy, average precision, average recall, and average F-1 over pixels. In addition, foreground accuracy (See eqn 4.6) is included as a measure of how accurately each method is at parsing foreground regions (those pixels on the body, not on the background). Note that the average measures are over non-empty labels after calculating pixel-based performance for each since some labels are not present in the test set. Since there are some empty predictions, F-1 does not necessarily match the geometric mean of average precision and recall.

Table 4.1 summarizes predictive performance of the parsing method, including a breakdown of how well the intermediate parsing steps perform. For comparison, the performance of the CRF model described in the previous chapter (for the detection scenario) is included in the table. The Paper-Doll approach outperforms the CRF approach in overall accuracy (**84.68**% vs **77.45**%). It also provides a huge boost in foreground accuracy. The previous approach provides **23.11**% foreground accuracy, while the Paper-Doll parsing obtains **40.20**%. The new approach also obtains much higher precision (**10.53**% vs **33.34**%) without much decrease in recall (**17.2**% vs **15.35**%). The reason for lower recall is further discussed in Section 4.5.3.

Figure 4.7 shows examples from the Paper-Doll parsing method, with

Table 4.1: Parsing performance for final and intermediate results (MAP assignments at each step) in percentages.

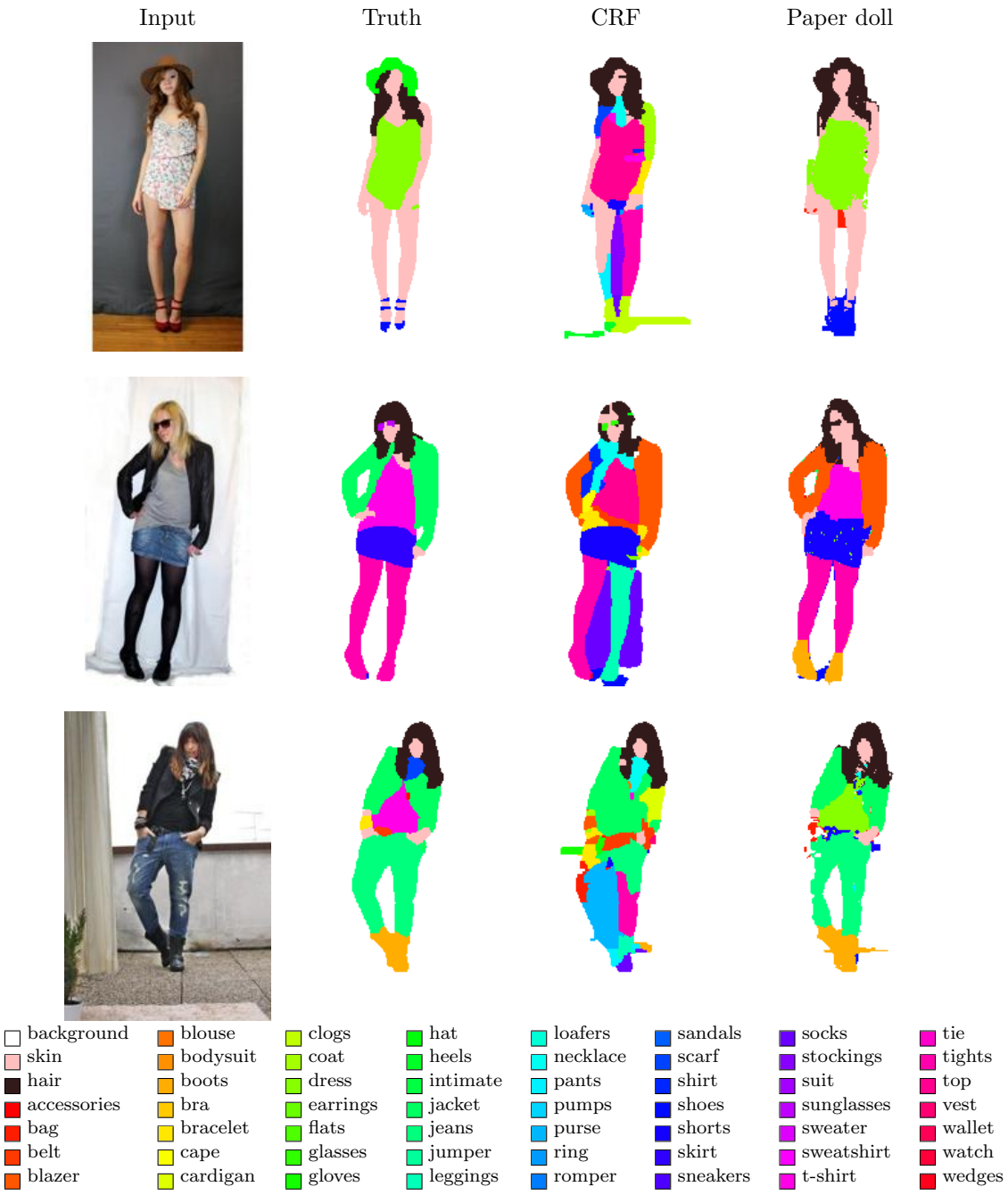
Method	Accuracy	Foreground Accuracy	Avg Precision	Avg Recall	Avg F-1
CRF	77.45	23.11	10.53	**17.20**	10.35
1. Global	79.63	35.88	18.59	15.18	12.98
2. Nearest	80.73	38.18	21.45	14.73	12.84
3. Transferred	83.06	33.20	31.47	12.24	11.85
4. Combined	83.01	39.55	25.84	15.53	14.22
5. Final	**84.68**	**40.20**	**33.34**	15.35	**14.87**

ground-truth annotation and the CRF method. It is observed that the Paper Doll approach produces a parse that respects the actual item boundary, even if some items are incorrectly labeled; e.g., predicting *pants* as *jeans*, or *jacket* as *blazer*. However, often these confusions are due to high similarity in appearance between items and sometimes due to non-exclusivity in item types, i.e., *jeans* are a type of *pants*.

Figure 4.8 plots F-1 scores for non-empty items (items predicted on the test set) comparing the CRF method with the Paper Doll method. The Paper Doll method outperforms the prior work on many items, especially major foreground items such as *dress*, *jeans*, *coat*, *shorts*, or *skirt*. This results in a significant boost in foreground accuracy and perceptually better parsing results.

4.5.1 Big data influence

To see the effect of retrieval in the Paper Doll pipeline in detail, the final parsing performance over retrieval-size and data-size is plotted in Figure 4.9, 4.10, and 4.11.


Input
Truth
CRF
Paper doll
background
skin
hair
accessories
bag
belt
blazer
blouse
bodysuit
boots
bra
bracelet
cape
cardigan
clogs
coat
dress
earrings
flats
glasses
gloves
hat
heels
intimate
jacket
jeans
jumper
leggings
loafers
necklace
pants
pumps
purse
ring
romper
sandals
scarf
shirt
shoes
shorts
skirt
sneakers
socks
stockings
suit
sunglasses
sweater
sweatshirt
t-shirt
tie
tights
top
vest
wallet
watch
wedges

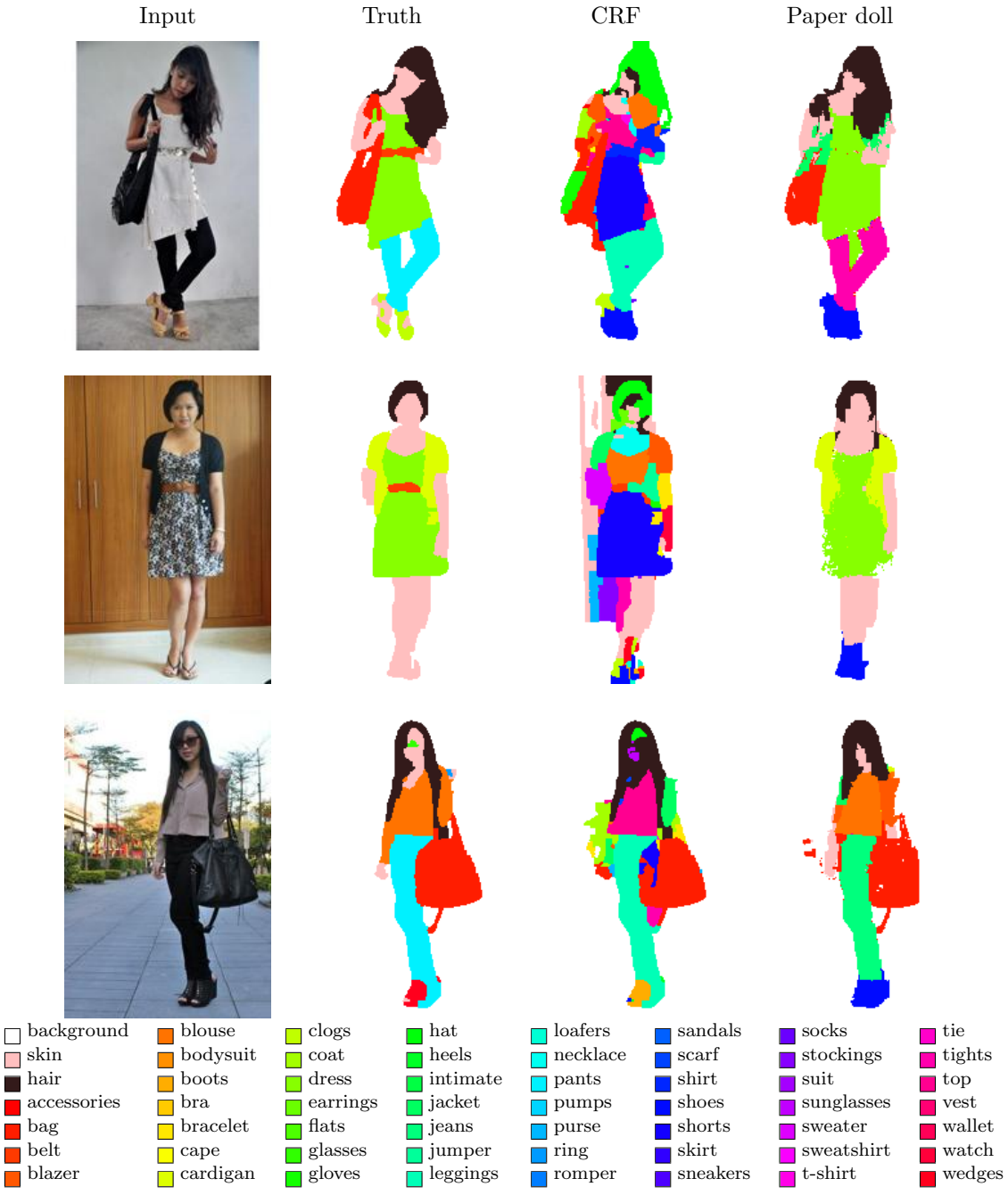


Figure 4.7: Parsing examples. The method sometimes confuses similar items, but gives overall perceptually better results.

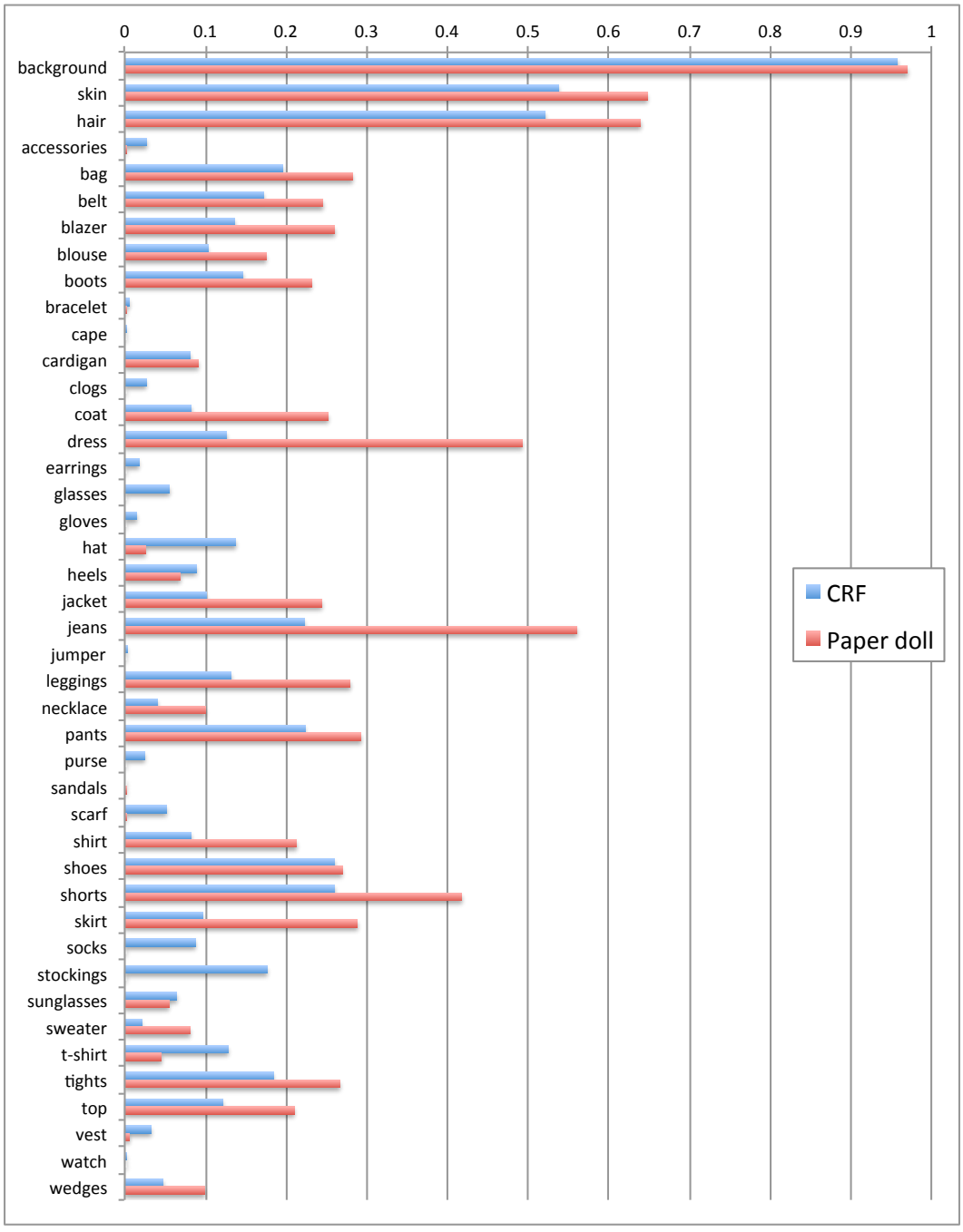


Figure 4.8: F-1 score of non-empty items.

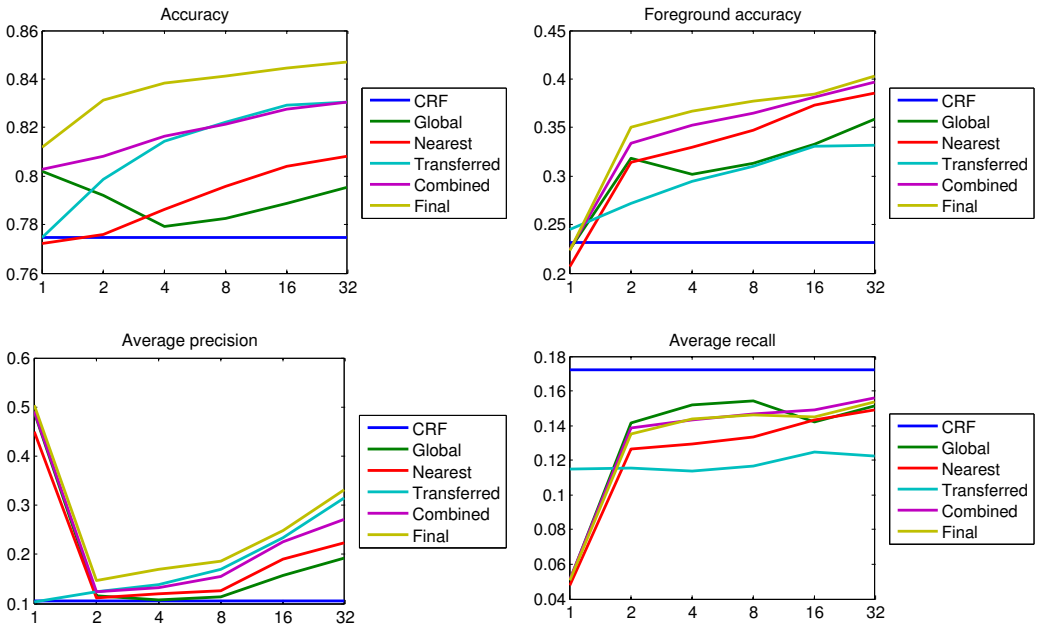


Figure 4.9: Parsing performance over retrieval size when items are unknown. Larger retrieval size results in slightly better parsing, but also takes longer computation time.

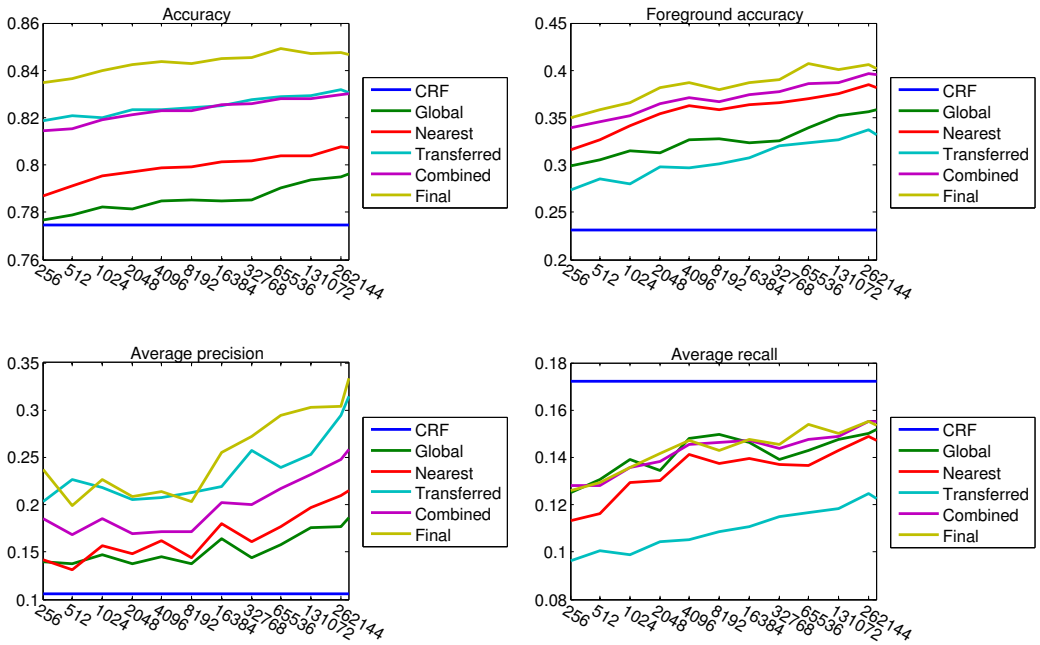


Figure 4.10: Data size and parsing performance when items are unknown (Detection). While average recall tends to converge, average precision grows with data size.

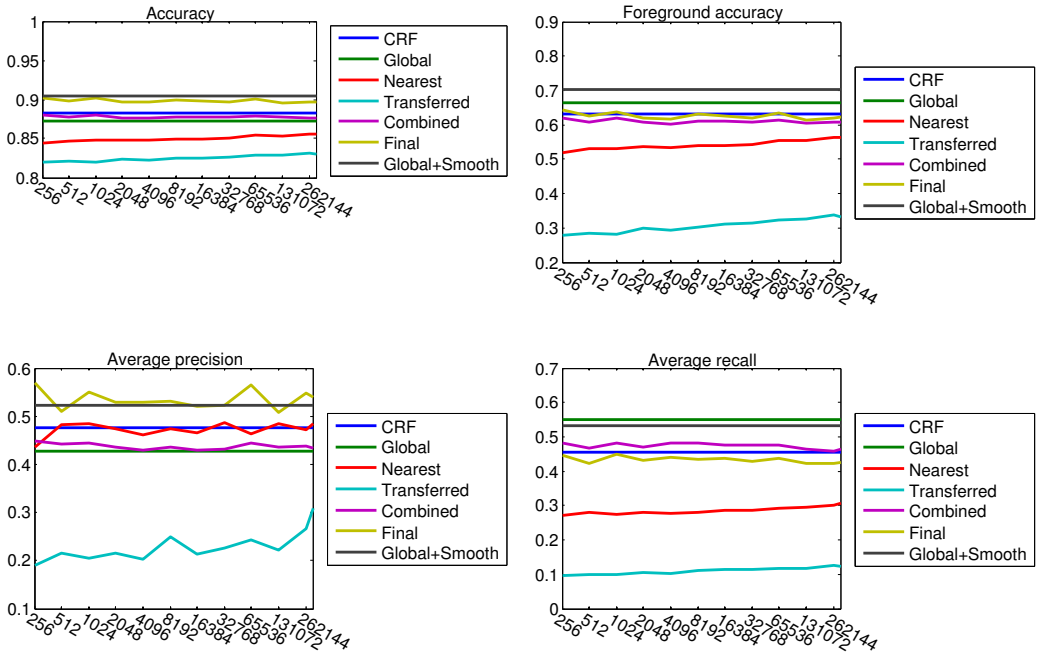


Figure 4.11: Data size and parsing performance when items are known (Localization).

Figure 4.12: Retrieval example for different data sizes. Predicted items are shown at the bottom. Notice at small data size, even a major item like dress or shirt can be missed in prediction.

Figure 4.9 shows the influence of the number of nearest-neighbors to foreground accuracy, average precision, and average recall for each parsing stage as well as CRF as a baseline performance. It is noted that there is a big gap between retrieving 1 image and 2 images, which is mostly due to the missing items appearing in the first nearest-neighbor. When there are more than one nearest-neighbors, the retrieval can prevent a major item such as dress from missing in the tag prediction. Beyond that, the quality of tag prediction gradually increases and that results in performance improvement, with a major effect in average precision. However, this performance increase comes with computational cost – Retrieving 1 image takes 8 seconds to parse one image, while retrieving 32 image takes 25 seconds to parse in the implementation. This is largely due to the increase in computation time at the NN parse and the transfer parse.

Also, to study how performance scales with data size, parsing performance is examined when images are randomly dropped from the Paper Doll dataset for various sizes. The number of retrieval is fixed to 25 in this experiment. Figure 4.10 shows the performance plot against the data size. All measures increase as the data size grow, but their rate differs; Foreground accuracy and average recall shows a moderate increase with respect to the data size, while average precision shows a major improvement at a large data size. This result shows the benefit of big data in the clothing parsing. Figure 4.12 shows examples of retrieval at data sizes = 256 and = 262,144. Clearly, a larger data size improves retrieval quality as well as item-prediction.

The results demonstrate the effectiveness of the big-data approach to clothing parsing. The drawback of this approach, though, is that it requires a lot

of storage space. In the implementation in this study, the Paper Doll dataset required about 71GB of disk space to keep the preprocessed images. Also the performance improvement is proportional to the exponential growth of the data size. However, the emphasis here is that the Paper Doll parsing does not require any manual annotation to the big data – the Paper Doll parsing can take advantage of big data from the online social network only for the cost of disk storage.

4.5.2 Localization and detection

The major motivation of using retrieval is to overcome the difficulty of the *detection* problem. In case of localization, items are known before parsing, and the goal is to *locate* items in a picture. Whereas in detection, the goal is to *identify* what kind of items in a picture in addition to localization. To see how much the retrieval approach helps in detection, this section examines a localization scenario when a list of ground-truth tags are given as input, and compares the performance with the detection case. In this scenario, the global model is given the ground-truth tags, the NN model only learns items included in the ground-truth tags, whereas the transfer parse does not get affected.

Figure 4.11 shows parsing performance vs. data size when items are known before parsing. This plot also adds a parsing result for the case of iterative smoothing applied directly to the global parse (Global+Smooth), in addition to other intermediate results. Note that the CRF model is specifically designed for this localization scenario and constitutes a strong baseline. The final result is performing better at average precision, with comparable result to the base-

line in foreground accuracy and average recall. However, the most effective model in the localization scenario is the global model with iterative smoothing. Note that this result is outperforming the CRF model of the previous chapter in all measures: foreground accuracy (**70.32% vs. 63.14%**), average precision (**52.24% vs. 47.63%**), and average recall (**53.25% vs. 45.54%**).

These results indicate that localization performance is not significantly affected by retrieval. This is an expected result, because the primary role of retrieval in the Paper Doll pipeline is to narrow down the list of potential items and to prevent confusion in parsing. When items are known, the retrieval process no longer serves this role in parsing. Eventually, the global model is sufficient for producing a good result in the localization scenario. In other word, a big-data approach is particularly effective to fill the performance gap between detection and localization scenarios.

4.5.3 Discussion

Though the Paper Doll approach is successful at foreground prediction overall, there are a few drawbacks. By design, the style descriptor is aimed at representing whole outfit style rather than specific details of the outfit. Consequently, small items like accessories tend to be less weighted during retrieval and are therefore poorly predicted during parsing. However, prediction of small items is an inherently and extremely challenging task because small items provide limited appearance information.

Another issue is the prevention of conflicting items from being predicted for the same image, such as *dress* and *skirt*, or *boots* and *shoes* which tend not to

be worn together. Iterative smoothing is effectively reducing such confusion, but the parsing result sometimes contains one item split into two conflicting items.

These two problems are root of the error in tag prediction – either an item is missing or incorrectly predicted – and increase the performance gap between detection and localization. One way to resolve these problems would be to enforce constraints on the overall combination of predicted items, but this leads to a difficult optimization problem.

Lastly, the results suggest that it is still difficult to predict items with skin-like color or coarsely textured items, as discussed in the previous chapter. Because of the variation in lighting condition in pictures, it is very hard to distinguish between actual skin and clothing items that look like skin, e.g. slim khaki pants, even if a similar style is found from the large-scale dataset. Also, it is very challenging to differentiate, for example, between bold stripes and a belt, using low-level image features alone. These cases will require higher-level knowledge about outfits to be correctly parsed.

4.6 Parsing for Pose Estimation

This section revisits the effect of introducing clothing parsing in pose estimation using the Paper Doll approach. Using the pose estimator of [113], this section compare three estimation scenarios.

- **Baseline**: using only HOG feature at each part.
- **Clothing**: using a histogram of clothing in addition to HOG feature.

Table 4.2: Pose-estimation performance with or without conditional parsing input.

Average precision of keypoints (APK)

Method	Head	Shoulder	Elbow	Wrist	Hip	Knee	Ankle	Mean
Baseline	0.9956	0.9879	0.8882	0.5702	0.7908	0.8609	0.8149	0.8440
Clothing	1.0000	0.9927	0.8770	0.5601	0.8937	0.8868	0.8367	0.8639
- Items known	1.0000	0.9966	0.9119	0.6411	0.8658	0.9063	0.8586	0.8829
Foreground	1.0000	0.9926	0.8873	0.5441	0.8704	0.8522	0.7760	0.8461
- Items known	0.9976	0.9949	0.9244	0.5819	0.8527	0.8736	0.8118	0.8624

Percentage of correct keypoints (PCK)

Method	Head	Shoulder	Elbow	Wrist	Hip	Knee	Ankle	Mean
Baseline	0.9956	0.9891	0.9148	0.7031	0.8690	0.9017	0.8646	0.8911
Clothing	1.0000	0.9934	0.9127	0.6965	0.9345	0.9148	0.8843	0.9052
- Items known	1.0000	0.9978	0.9323	0.7467	0.9192	0.9367	0.9017	0.9192
Foreground	1.0000	0.9934	0.9148	0.6878	0.9127	0.8996	0.8450	0.8933
- Items known	0.9978	0.9956	0.9389	0.7183	0.9105	0.9214	0.8734	0.9080

- **Foreground**: using a histogram of figure-ground segmentation in addition to HOG feature.

In all scenarios, this experiment uses 5 mixture components for all body parts.

The foreground model is computed by simply treating non-background regions in clothing parsing as foreground. Comparing the clothing model and the foreground model reveals how *semantic* information helps pose estimation given non-semantic segmentation. For clothing and foreground cases, this experiment also checks the performance of ground-truth pixel annotation used as input, which serves as the performance limit of each model given a perfect segmentation.

Table 4.2 summarizes *average precision of keypoints* (APK) and *percentage of correct keypoints* (PCK) using the Fashionista dataset. Clearly, introducing clothing parsing improves the quality of pose estimation. Furthermore, the improvement of the clothing model over the foreground model indicates that

the contribution is coming from the inclusion of *semantic* parsing, not from a simple figure-ground segmentation.

Note that clothing parsing is particularly effective for body extremities of the body, such as wrist, as the difference between the baseline and the upper-limit suggests. Perhaps this is due to items specific to certain body parts, such as *skin* for wrist and *shoes* for ankle. Note that a figure-ground segmentation cannot provide such semantic context. This result gives an important insight into the pose estimation problem, since improving estimation quality for such body extremities is the key challenge in pose estimation, while state-of-the-art methods can already accurately locate major parts such as head or torso. Semantic parsing perhaps gives a strong context to improve localization of minor parts that often suffers from part articulation.

4.6.1 Iterating parsing and pose estimation

The previous subsection showed that pose estimation can benefit from parsing. Since clothing parsing depends on pose estimation, this subsection also evaluates the effect of iterating between pose estimation and clothing parsing. This iterative process starts by clothing parsing with the baseline pose estimator, followed by the pose estimation conditioned on clothing parsing. Then, the iterative approach replaces the pose estimation input to the parsing pipeline with the output of the conditional pose estimator, and continue the same process for a couple of iterations. Denoting parsing by Y and pose configuration

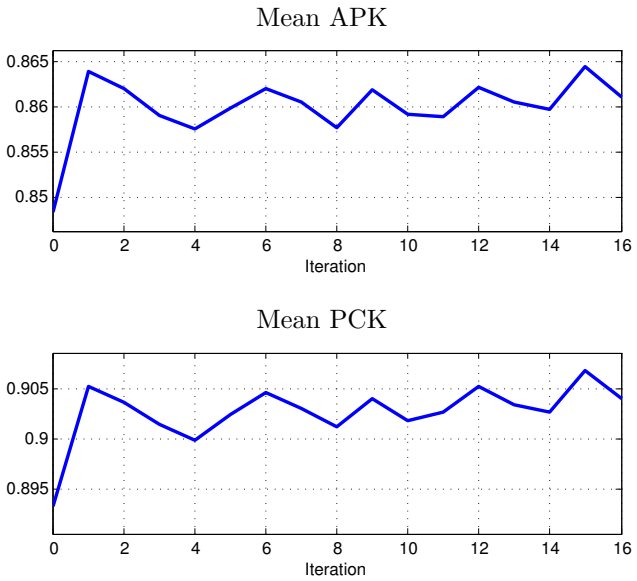


Figure 4.13: Pose-estimation performance over iterations.

by Z, the process can be described in the following for iteration $t = 0, 1, ..., n$:

$$Z_0 \equiv \arg\max_Z P(Z), \tag{4.10}$$

$$Y_t \equiv \arg\max_Y P(Y|Z_t), \tag{4.11}$$

$$Z_{t+1} \equiv \arg\max_Z P(Z|Y_t), \tag{4.12}$$

where $P(Z)$ is the baseline pose estimator, $P(Y|Z)$ is the parsing model, and $P(Z|Y)$ is the conditional pose estimator.

The performance is evaluated for pose estimation and parsing over iterations using the Fashionista dataset. Figure 4.13 and 4.14 plots the performance. The plot shows that the performance starts to oscillate after the first pose re-estimation by the conditional pose model, and there is no benefit of the iterative process in parsing. This result reflects that a slight change in pose estimation does not affect too much the final parsing quality.

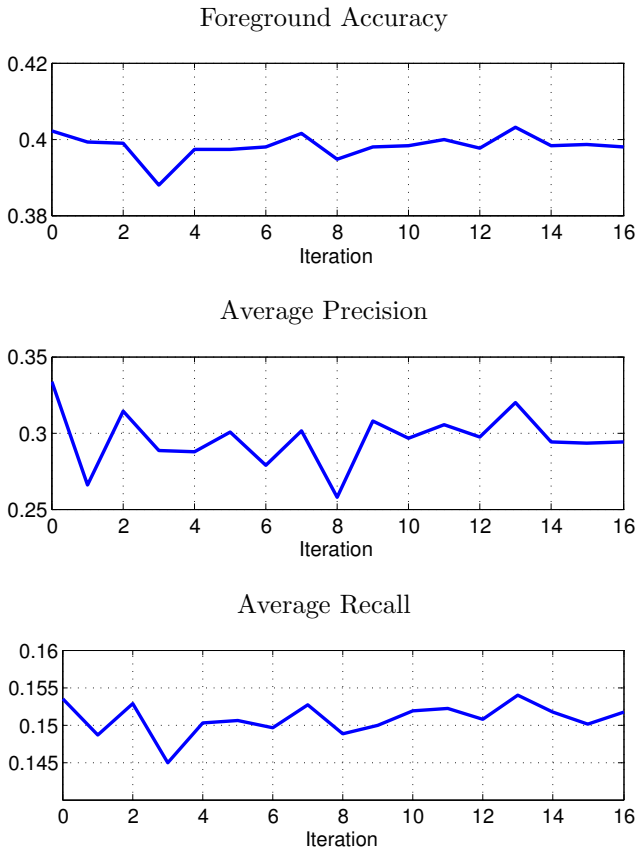


Figure 4.14: Parsing performance over iterations.

Oscillation happens because the iterative model does not guarantee convergence. The approach in this section independently solves pose estimation and clothing parsing, and thus there is a discrepancy in the objective in this iterative process. To make the iterative approach converge, it is necessary to consider a joint model of pose and parsing, and try to optimize for the global objective. Such an approach is an interesting future direction [56].

In the end, this result suggests that 1) the conditional pose estimator can improve the performance in the first pose re-estimation, but 2) further iteration may not improve the performance.

4.7 Summary

This chapter described a clothing-parsing approach in a detection scenario. The proposed parsing framework consists of nearest-neighbor style-retrieval and a pixel-wise parsing approach using global clothing models, local models computed on the fly from retrieved samples, and a transfer of clothing likelihood from the nearest-neighbors. Experimental evaluation showed successful results, demonstrating a significant boost of overall accuracy and especially foreground parsing-accuracy over the CRF model designed for a localization scenario. Also, experimental evaluation indicated that the data-driven approach is key to resolve the difficulty involved in detection of large item categories. While an iterative approach does not help, pose re-estimation with the Paper Doll approach showed the effectiveness of semantic parsing in pose estimation.

Chapter 5

Popularity Analysis

This chapter focuses on the other aspect of the understanding problem in online visual networks, namely, behavior understanding. Specifically, this chapter studies how much visual, textual, and social factors contribute to the popularity of a picture in the real-world network. To analyze the social popularity phenomena, this study makes use of computer vision techniques to characterize visual content related to outfits in addition to social network factors, as well as textual meta data and network information. Then, regression and classification analyses are applied in both in-network and out-of-network scenarios. The empirical results indicate significant statistical evidence that social factors dominate within-network popularity while this dominance does not occur out of network. This result suggests the study of image popularity should carefully consider the strong affect of social factors in the visual network.

5.1 Motivation

Nearly every blog or social network utilizes a combination of images, text, and other modalities (e.g., location) to convey information and promote interaction. In many online communities the amount of visual data is quite vast, sometimes representing the main source of content. For example, Instagram has 40 million daily uploads with a total of 16 billion pictures, Flickr hosts 8 billion photos, newcomer Pinterest already has over 70 million users, and Facebook boasts 350 million photos uploaded daily with over 240 billion pictures total.

Despite the underlying multi-modal nature of the data in many online social communities, many social network analyses have only focused on a single modality, such as examining network structure, or using text processing techniques to access linguistic content. Developing algorithms that make use of image and video information is a clear next step toward exposing the currently unstudied *dark visual matter* for improved social network understanding.

To date very few research has used visual recognition techniques for network analysis, with the only exception of the very recent work by Khosla et al. [49]. This is perhaps because computer vision is a very challenging problem and the results of automatic computer vision techniques are often extremely noisy. However, for specific settings or more constrained visual recognition problems, visual analysis may be feasible and useful; Stone et al [99] demonstrate improved facial recognition when computer vision algorithms are combined with network structure information, Crandall et al [24] use scene based image and text content analysis combined with location-based structure to

organize a large collection of geo-tagged photos. Recently some attempts are made to evaluate the influence of indirect forms of visual information on popularity or behavior [7, 20, 4].

This chapter studies behavior in an online fashion network using visual, textual, and social factors. In particular, the study examines the influence of content and social factors on post popularity. The study purposely chooses a social fashion-network, Chictopia, where fashion pictures are the main form of content, for the following reasons; 1) the network is large and real-world, with over 175k users and 600k pictures; 2) content in this network is mainly visual, consisting of "outfit of the day" pictures uploaded by users; 3) the community is focused on a single topic (fashion), which yields relatively consistent user-based popularity; 4) the relevant data is publicly and readily available online; and 5) the analysis can employ mature computer vision techniques – for person detection and pose estimation – as tools to help extract the visual content most relevant to style and popularity.

The proposed popularity analysis builds on a multimodal content modeling approach. The analysis first quantifies various available information from a fashion picture, including content metadata, computer vision features, natural language features, and social network information. Then statistical analyses are applied to these data to reason about how those factors may affect post-popularity prediction performance. To incorporate useful visual information, a new comprehensive style descriptor is developed in this chapter on top of body pose estimates that captures the visual characteristics of an outfit.

This study revealed concrete statistical evidence of a strong potential for the existence of an inherent "social bias" in the real-world social network.

Finally, the study also looks at how user behavior changes in an out-of-network environment where no social factors are present by design.

The following list summarizes the major focuses of this study.

- A multimodal approach to quantify fashion pictures including visual, network, metadata, and textual factors, using state-of-the-art computer-vision techniques
- A new computer-vision feature for representing outfit style based on clothing parsing
- An empirical study of social vs. content influence on popularity in a large-scale fashion network
- A comparison between out-of-network (i.e., socially-isolated) and in-network popularity modeling through a large crowdsourcing effort

In a network focused around fashion, and style, one might predict visual content to be the most influential factor for popularity. However, this study suggests that social factors actually dominate both visual and textual factors in prediction models, even in a network where outfits are purportedly rated based on their fashion style. This perhaps illustrates how strongly the *rich-get-richer* phenomenon [5] affects content evaluation in a social network. Furthermore, the results of the study of the content effects outside of the social network suggest that the dominant social bias does not appear in the out-of-network scenario, but rather the social and content information provide complementary information to explain popularity. This result may provide potentially useful insights to domain experts or researchers studying human behavior and the

Figure 5.1: An example of a Chictopia post.

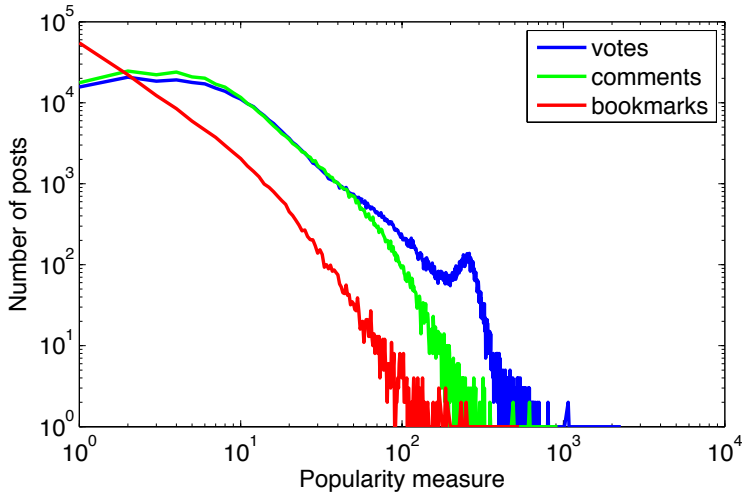


Figure 5.2: Popularity distribution.

popularity of visual content in social networks, and to engineers seeking to exploit user behavior characteristics in social-network applications.

5.2 Dataset

This chapter uses data from `chictopia.com`. In Chictopia, users post not only pictures of their daily outfits but also various textual metadata, including a title, description, and several labels: general style category, occasion, colors, or

free-form tags. Users can also list individual clothing items with color, brand, and one free-form word. The tagging is a *self tagging* model [73] where only the uploader can associate tags. Figure 5.1 shows an example data item.

In this network users can interact with other users through voting, commenting, or bookmarking posts. Also, users can follow other users (fanship = unidirectional relationship) or request others to be a friend (friendship = bidirectional relationship). The friendship is analogous to the Facebook-style relationship, while the fanship is analogous to the Twitter-style relationship. Users can subscribe to feeds from the connections in their user page.

The data collection from Chictopia reached 617,708 posts in total. To compare visual features consistently across images, the pre-processing step runs a state of the art pose detector [113] on all images, which automatically estimates the body pose (i.e., location of arms, legs, torsos, etc.) of people depicted in photos. This leaves us with 328,604 pictures in which a standing person is found, dating from March 2008 to Dec 2012, with 34,327 unique users. Note that the number of users in the website was over 175K at data collection time, but many users follow the website without posting.

The popularity measures considered in this chapter are the number of *votes*, *comments*, and *bookmarks* associated with each post. As is the case with any web content, Chictopia popularity reveals a long-tailed distribution. Figure 5.2 shows the log-log plot of the popularity histogram from all 617K posts. We note that the number of votes shows a slight kink in the distribution perhaps due to front-page highlighting or a special promotion by the website.

Type	Name	Modality	Vector	Size
Social	User identity	Network	Sparse	up to 1,000
	Node degrees	Network	Dense	6
	Previous posts	Metadata	Dense	1
Content	Tag TF-IDF	Textual	Sparse	up to 1,000
	Style descriptor	Visual	Dense	441
	Parse descriptor	Visual	Dense	1060
	Color entropy	Visual	Dense	6
	Image composition	Visual	Dense	6
Other	Date bias	Metadata	Sparse	up to 58

Table 5.1: Summary of the content models.

5.3 Content Representation

The representation of a user-post is a vector of the quantized information-sources available for the post. Social factors (Sec 5.3.1) and content factors (Sec 5.3.2) are the two major components of the model. Additionally, a date term is added to model popularity-bias due to season and growth of the network over time (Sec 5.3.3). Table 5.1 summarizes all of the terms of the model.

5.3.1 Social factors

Social factors capture information related to the user and their social status within the network – factors related to the user's identity, their node degrees, and posting frequency. Note that although social factors only quantify the network information, they are certainly correlated to the content quality. For example, posts from the same user tend to be similar in quality.

User identity The identity of users is represented as a sparse indicator vector. To constrain feature dimensionality, this indicator vector is restricted

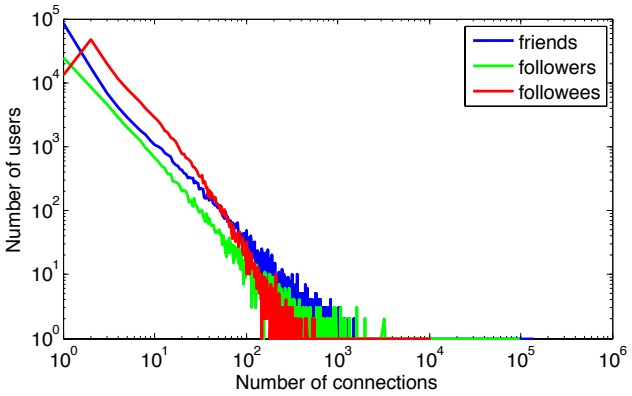


Figure 5.3: Distribution of node degrees in Chictopia.

to the top-1000 most frequent users. Users not in the top-1000 receive a feature vector with all zero elements. Posts from the same user will all have the same feature vector.

Node degrees A six-element feature vector is constructed from the counts and log-counts of friends, followers, and followees of the user. Posts from the same user will all have the same feature vector. Figure 5.3 plots the distribution of node degrees from all 175K users in Chictopia. The distribution displays a long tail, very characteristic of social networks in general.

Note that due to limitations imposed by the network information available, the node degree is calculated at the time of data collection rather than at the time of posting. The same applies to the popularity data. Although this may incur a slight difference in context for older posts, because this number is almost always monotonically increasing over time, the assumption is that they are all under the same condition in the experiments.

This chapter uses only node degrees as social features related to network structure. Other network-related features, such as the identity of friends,

followers, and followees, were not chosen after empirical testing due to their less reliable prediction compared to node degrees.

Previous posts A scalar feature is built from the number of previous posts from the same user. This number is calculated based on the post's timestamp.

5.3.2 Content factors

This section proposes to apply natural-language processing and computer-vision techniques to model textual and image information for post content. Textual features employ a language model based on n-grams and TF-IDF (term frequency - inverse document frequency) [71]. Visual features combine general color measures and two high-level features.

Tag TF-IDF First, a post is represented as a text document using unigrams and bigrams from all the text labels. Then, TF-IDF weights are calculated as a language-based feature representation for posts [71].

In Chictopia, a user can label each individual post with tags indicating general style, occasion, colors, or free-form keywords. In addition, there are structured tags describing individual clothing items: one clothing item has a tuple of associated tags consisting of a color, a brand, or a free-form tag word. Unigrams are computed from all of these user-provided labels. Additionally, bigrams are computed from the listed items by concatenating the item type with any other word in the tuple. For example, for an item described as "white h&m printed t-shirt", bigrams will be `white t-shirt`, `h&m t-shirt`, and `printed t-shirt`.

As in the case of user identity, to constrain the dimensionality of this feature, this chapter only considers the 1000 most frequent n-grams found in the training samples and ignore other words appearing at test time for the analysis.

Note that the model does not use title and description of a post because they are often missing, fairly irrelevant to the content, or simply a duplicate of one of the structured tags.

Style descriptor The content model includes the style descriptor described in Sec 4.3.1 for this study. The style descriptor is not only useful for image retrieval but also considered a comprehensive visual representation of the clothing a person is wearing.

Parse descriptor In addition to the style descriptor, this section proposes a new fashion-focused image descriptor based on clothing parsing (Chapter 4), which is named *parse descriptor*. The parse descriptor represents the appearance of individual garment items found in a picture, and is experimentally verified to give a strong prediction in combination with the style descriptor. The parse descriptor is considered to be the most important representation in content analysis, because the parse descriptor specifically captures the appearance of a person's garment items. Figure 5.4 illustrates both the style and parse descriptors.

The parse descriptor is computed using the following steps;

1. compute clothing parse using the method in Chapter 4, and obtain 10 masks corresponding to specific garment groups, such as *outer top*, *dress*,

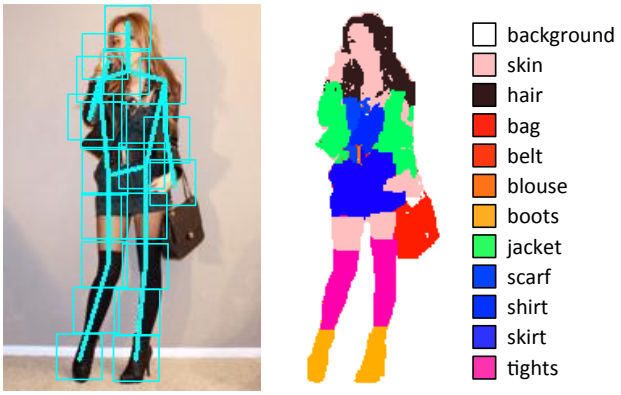


Figure 5.4: Style descriptor (left) and parse descriptor (right). The style descriptor extracts visual information from patches while the parse descriptor extracts information from the predicted clothing parse (semantic assignment of pixels to garment labels).

or *footware*. Note that the original 56 garment categories are mapped to 10 garment sets to improve robustness,

2. extract RGB color, Lab color, Texture response, HOG descriptor, distance from image border, and probability of skin and hair at every pixel,

3. compute mean-std pooling of the extracted features in each region,

4. concatenate all pooled features over 10 regions (1060 dimensions).

Color entropy The entropy of RGB and Lab color is calculated from the image, which yields a 6-element vector. This feature helps distinguish drawings which sometimes occur on Chictopia from natural photos.

Image composition Given a bounding box encompassing a person (estimated by the pose detector), this feature measures the overall composition of how the person is depicted relative to the image frame. The following information is extracted from a bounding-box for each image:

1. normalized width, height, and area;

2. normalized x and y displacement from the center of the image; and

3. normalized distance from the image center.

5.3.3 Other factor

Date bias This is a sparse indicator vector to represent which month the picture is posted. This feature has up to 58-elements in the collected data from Chictopia. The date bias is used in all models in the experiments, because this is designed to take into account the popularity difference due to the site growth and the seasonal influence.

5.3.4 Preprocessing

The following normalization procedures are applied to the data set before any of the statistical analysis;

- every element of the dense factors is scaled so that the mean is zero and $\pm 3\sigma$ range is rescaled to $(-1, 1)$ in the training samples, and

- elements of sparse factors are scaled by their maximum value in the training samples without adjusting by the mean value.

This pre-processing improves numerical stability in the statistical analyses.

5.4 In-Network Popularity

This section provides statistically significant evidence indicating existence of "social bias" in popularity in the network. In this section, the Chictopia posts are analyzed using three approaches: correlation, regression, and classification. Note that the main interest in this section is *not* to identify which individual features within a particular class of factors are most informative, but to reason at the general class level. Hence, this section uses a prediction-based analysis because it allows us to differentiate the predictive power of each class of factors, social vs. content, without having to explicitly consider the specific features within each factor that most contribute to the corresponding prediction performance, or the possible correlations among those features.

5.4.1 Network-popularity correlation

As a preliminary study, this subsection shows the correlation between network structure and the observed popularity. Table 5.2 shows the Pearson and Spearman correlation coefficients between the node degrees and observed popularity measures in Chictopia, using 328K posts.

These correlation measurements reveal that from among the social factors (measures of node degree) the number of fans has an unignorable correlation to all of the popularity measures. However, note that this does not immediately mean that popularity is a function only of social factors as other content based factors could also be important.

Pearson

friends	1.00						
fans	0.58	1.00					
followees	0.12	0.24	1.00				
votes	0.21	0.49	0.06	1.00			
log-votes	0.28	0.45	0.09	0.75	1.00		
comments	0.18	0.37	0.02	0.58	0.44	1.00	
bookmarks	0.32	0.40	0.10	0.56	0.61	0.73	1.00
	fri.	fans	fol.	vot.	log.	com.	boo.

Spearman

friends	1.00					
fans	0.65	1.00				
followees	0.39	0.45	1.00			
votes	0.47	0.64	0.28	1.00		
comments	0.39	0.52	0.14	0.64	1.00	
bookmarks	0.50	0.57	0.28	0.80	0.61	1.00
	fri.	fans	fol.	vot.	com.	boo.

Table 5.2: Pearson and Spearman correlation coefficients between the node degrees in the network and the observed popularity. All values are non-zero ($p < 10^{-6}$). Notice the relative strength of correlation between fans and popularity measures.

5.4.2 Regression analysis

A regression analysis is applied on the votes and log-votes of the posts using social, content, and a combination of social and content factors.

The analysis uses a linear regression model. Let us denote the popularity measure with y (votes or log-votes), model parameters with θ, and a factor representation with $\mathbf{x}$. The regression model is described by:

$$y = \theta^{\mathrm{T}}\mathbf{x} \,. \tag{5.1}$$

In the following experiments, $\mathbf{x}$ is a concatenation of various factors explained

Factors	R^2		Spearman	
	votes	log-votes	votes	log-votes
Social	0.372 ±0.012	0.491 ±0.005	0.591 ±0.005	0.682 ±0.004
Content	0.132 ±0.005	0.248 ±0.005	0.418 ±0.005	0.485 ±0.005
Social+Content	0.341 ±0.010	0.493 ±0.005	0.572 ±0.005	0.685 ±0.004
Content: Textual	0.115 ±0.005	0.166 ±0.004	0.314 ±0.006	0.388 ±0.005
Content: Visual	0.119 ±0.004	0.212 ±0.004	0.395 ±0.006	0.450 ±0.005

Table 5.3: Regression results on the observed popularity with accompanying 95% confidence intervals on error. For cleaner presentation, the tiny asymmetric difference in bootstrapped confidence intervals is rounded.

in the previous sections.

L2-regularized support vector regression is used to learn a model [32], which is the default regression model in this package. The free parameters of the learning algorithm are searched over a grid with 10-fold cross validation on the training samples.

This analysis measures how social factors affect on the votes and log-votes of posts by comparing the fitness of the regression models consisting of social factors only, content factors only, or a combination of both. This study measures two fitness criteria: R^2 and Spearman coefficients. These measures are evaluated on a statistical bootstrapping protocol with the 328K posts; posts are randomly resampled with replacement and subsampled to 10,000 posts (for computational tractability), and the above measures are evaluated with a 9,000 / 1,000 train / test split. This process is repeated 100 times to derive statistical significance.

Table 5.3 shows the results of the regression analysis. The regression models fit significantly better when a regression model contains social factors, suggesting that a user's social connections largely dominate the popularity of their posts over the post's particular content. Additionally, the difference between votes and log-votes indicates that the distribution of popularity is better modeled in log-scale, which is consistent with the long-tailed distribution observed in Figure 5.2. The difference of the social-only model and the combined model indicates that even explicitly incorporating content factors gives a worse fit with a linear model when the distribution follows such a long-tail.

The result implies that the popularity of the content is dominantly determined by the network regardless of the picture quality or clothing / dress aesthetic appeal. However, note that social factors can be highly correlated with content quality – users with many followers may tend to wear highly fashionable outfits. Nevertheless, the results indicate that the effect of content quality is likely to be considerably smaller than network influence.

Table 5.3 also shows the effect of using only textual or visual factors from the content model. The combination of both (Content) gives a better fit.

5.4.3 Classification analysis

Next, this section applies a classification-based analysis for predicting popularity. In this case, instead of predicting the exact popularity values y, the predicted value is a binary indicator $z = 1\,[y > \alpha]$ for some threshold α, popular vs. unpopular, using a linear classifier. This approach can reveal the effect at various definitions of popularity. The model is equivalent to (5.1),

Factors	Popularity thresholds		
	Top 25%	Top 50%	Top 75%
Social	0.847 ±0.002	0.761 ±0.003	0.779 ±0.003
Content	0.778 ±0.003	0.664 ±0.003	0.737 ±0.003
Social+Content	0.845 ±0.002	0.754 ±0.003	0.775 ±0.002

Table 5.4: Accuracies from the social-only model, the content-only model, and the combined model on top K% prediction of the observed popularity with accompanying 95% confidence intervals on error.

but learned using L2-regularized logistic regression.

This experiment varies the threshold value α for 25%, 50%, and 75% quantiles of the votes in the training samples. The predictive performance is measured in terms of accuracy (i.e., 1 minus the 0/1 misclassification error). The experimental protocol is the same as for the regression analysis – performance evaluation consists of a statistical bootstrapping procedure: 100 random resampling with replacement, followed by random subsampling to 10,000 sample size with a 1,000 / 9,000 split in each batch.

Table 5.4 shows prediction accuracies. Classification reveals an asymmetry between top 25% and top 75% prediction, indicating the prediction of the most popular posts is easier than predicting the least popular posts. This is perhaps partly due to the consistently better quality of top-rated pictures and partly due to social bias more strongly affecting popular posts. Also, the slightly smaller difference between the social-only and the content-only model in top 75% prediction suggests that popularity is less affected by social influence in least-popular regime.

5.5 Out-of-Network Popularity

Given these findings of social bias in social network based popularity, this section studies popularity without social influence. Toward this goal, the study in this section utilizes *crowdsourcing* to emulate the voting process in a content network without social relationships. Then, the same statistical analyses are applied to the voting data using the collected "out-of-network" popularity votes.

5.5.1 Crowdsourced popularity

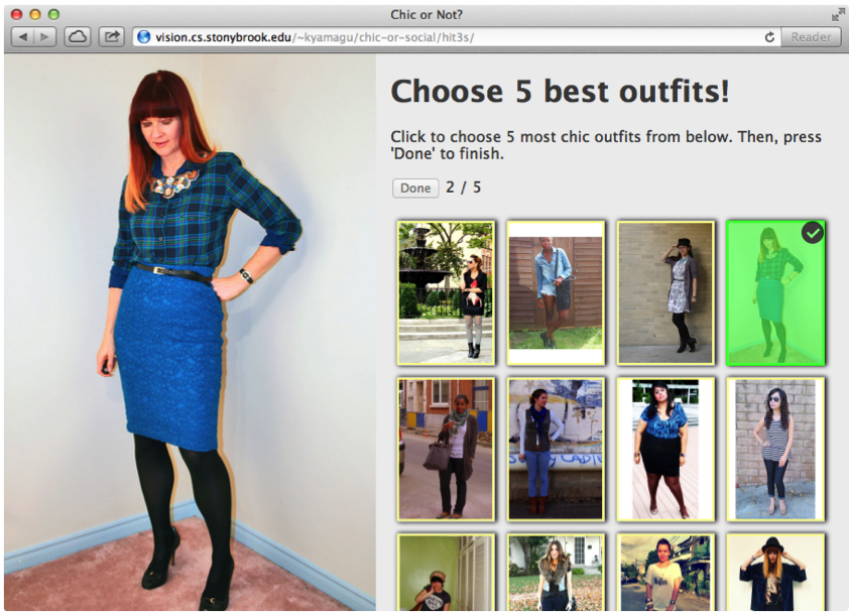
The crowdsourced post-popularity is collected using Amazon Mechanical Turk.

Task design

Here the goal is to design a task that resembles the voting environment of Chictopia but without the social network. The task-design consists of two steps. The first step exposes users to a reasonably large set of pictures individually and at close range (at a similar resolution to what they would observe on Chictopia), to more closely emulate the environment of Chictopia, where users are likely to have browsed through many pictures while interacting with the website. The second step of the task takes advantage of the first step experience. The following describe each step in more detail. Figure 5.5 shows the task interface.

Binary decisions The first step of the task shows the worker 50 random large-resolution images in sequence, and asks them to vote on the picture if

Binary votes



Top-K votes

Figure 5.5: Crowd-voting interface.

they find it "chic"[1],

For quality control, the task measures how long it takes each worker to complete this first step and rejects a worker's votes if they completed the step too quickly or did not display enough variation in their voting procedure.

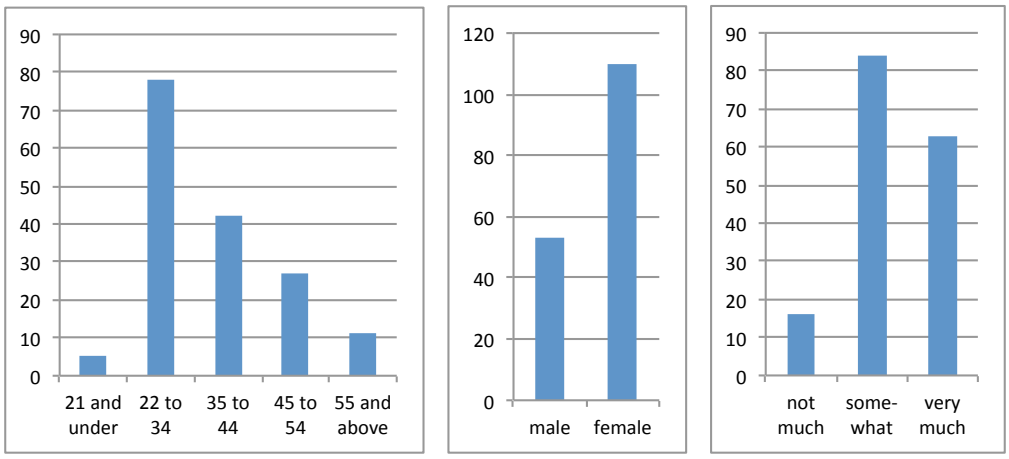
Top K selections After the first step, the task shows the worker an array of thumbnails for the 50 pictures they saw in step one, and ask them to select the 5 most "chic" pictures. To facilitate the worker's choice from this relatively large collection of images, the interface orders the thumbnail pictures so that those pictures where the worker voted "chic" in the first step appear at the top of the array (ranking).

Voting data

The crowd experiment is performed by randomly selecting 3,000 posts (from the dataset of 328k) and instantiating the above tasks 60 times. Each of the 60 resulting tasks assigns 25 workers. Thus, any post on each task obtains up to 25 "chic" votes in both binary decisions and top-K selections.

Prior to starting the MTurk tasks, workers are also asked to answer general demographics information such as gender, age group, and their degree of fashion interest. Even though the crowdsourcing website does not permit precise control over population, according to the answers collected, this experiment attracted young, female workers with interests in fashion, a group likely to overlap / align well with, and thus be generally representative of, the typical Chictopia user population. Figure 5.6 depicts the statistics.

[1]Chictopia uses the term *chic* for popularity voting.



Age Gender Fashion interest

Figure 5.6: Worker demography. The crowdsourced task attracted young, female workers with interests in fashion.

Figure 5.7 shows the distribution of the number of votes obtained from MTurk. For binary decisions, the mode / peak is around 8 votes, while for top-K selections, the mode is at 0. Though this experimental protocol tried to emulate the environment of Chictopia, the distribution does differ from Chictopia's (Figure 5.2). One possible reason for this difference may be the scale (number of samples) of this experiments vs Chictopia. Another possibility is the lack of social-network structure.

5.5.2 Network-crowd correlation

Similarly to Section 5.4.1, this section first studies the correlation between network-based popularity and crowd-based popularity. Table 5.5 shows Pearson and Spearman correlation coefficients between votes and log-votes from Chictopia, and both binary and top-K voting from the crowds. The relative weakness of correlation between Chictopia and the crowd suggests that the

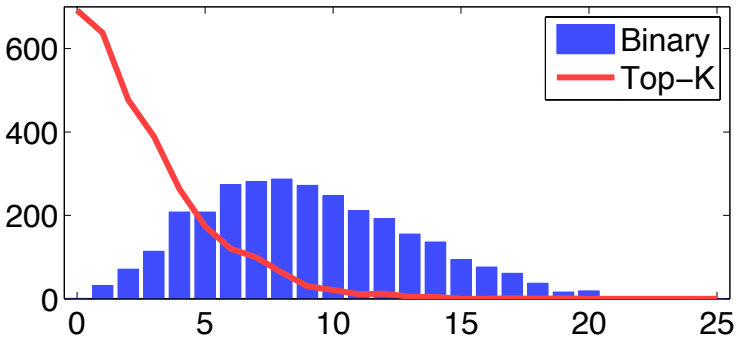


Figure 5.7: Crowd-votes distribution. Binary voting resulted in each post getting varying number of votes while the top-K voting resulted in a long tail.

	votes	log-votes	crowd-bin	crowd-top
votes	1.00			
log-votes	0.75	1.00		
crowd-bin	0.26	0.36	1.00	
crowd-top	0.26	0.32	0.78	1.00

Pearson

	votes	crowd-bin	crowd-top
votes	1.00		
crowd-bin	0.36	1.00	
crowd-top	0.33	0.77	1.00

Spearman

Table 5.5: Pearson and Spearman correlation coefficients between network popularity and crowd popularity. All values are non-zero ($p < 10^{-6}$).

Factors	R^2		Spearman	
	binary	top-K	binary	top-K
Social	0.423	0.387	0.634	0.597
	±0.011	±0.014	±0.007	±0.008
Content	0.428	0.348	0.647	0.560
	±0.012	±0.013	±0.011	±0.008
Social+Content	0.473	0.389	0.686	0.598
	±0.014	±0.014	±0.008	±0.008

Table 5.6: Regression results on crowd popularity with accompanying 95% confidence intervals on error.

crowds disagree with what Chictopia users believe to be *chic*. Granted, such disagreement can come from different interpretations of the word *chic* in the two communities.

5.5.3 Regression analysis

Using the voting data from the crowd, the regression analysis is applied as in Section 5.4.2. The main interest of this experiment is, however, the influence of the social factors observed in Chictopia on the regression based on crowd popularity. This experiment uses the same bootstrap method from the 3,000 posts to compare the social-only, content-only, and combined models in this experiment. Here, social factors are taken from Chictopia dataset which has no direct relationship to crowd popularity. Table 5.6 shows the results of the social-only model, the content-only model, and the combined model.

Given the in-network results, the social factors from Chictopia are initially expected to lead to weaker predictors. However, the results suggest that social factors (from Chictopia) still lead to comparable predictors in binary voting, and stronger predictors in top-K voting, even if the voting data come from the

out-of-network environment. The difference between binary and top-K voting seems to be from the difference in distribution – the distribution of top-K votes look more similar to the long-tail of Chictopia. However, there is no social network in this experiment. It is likely that the solid regression result obtained from the social factors is due to the function of social factors serving as a content evaluation; i.e., user and content quality correlation. But clearly there needs to be more research to parse this out.

Also, the result of the combined model (social + content) is significantly better than the social-only model for binary voting. One possible explanation may be that the content factors are providing complementary information to the social factors in predicting the *unbiased* popularity from the crowd, as opposed to the biased popularity in the network where social influence is by far the stronger predicator of popularity.

5.5.4 Classification analysis

Table 5.7 summarizes the results of classification analysis. Interestingly, here the content-only model gives a much stronger predictive performance than the social-only model, in contrast to the mixed results obtained from the regression analysis in Table 5.6.

Also, the asymmetry of popularity prediction (25% vs. 75%) still holds in this results, which was observed in the in-network analysis in the previous section. Clearly this result indicates the importance of looking at various definitions of popularity.

In summary, the out-of-network popularity analysis yields insights that

Vote	Factors	Popularity thresholds		
		Top 25%	Top 50%	Top 75%
Binary	Social	0.845 ±0.004	0.740 ±0.005	0.787 ±0.005
	Content	0.888 ±0.004	0.835 ±0.004	0.862 ±0.004
	Soc.+Con.	0.884 ±0.004	0.825 ±0.005	0.858 ±0.004
Top-K	Social	0.861 ±0.004	0.743 ±0.005	0.728 ±0.004
	Content	0.896 ±0.003	0.834 ±0.004	0.826 ±0.004
	Soc.+Con.	0.890 ±0.004	0.813 ±0.005	0.806 ±0.004

Table 5.7: Accuracies of the social-only model, the content-only model, and the combined model on top $K\%$ prediction of the crowd popularity with accompanying 95% confidence intervals on error.

suggest that

- social factors contain not only network information but also some aspect of content evaluation,
- content factors probably capture different aspects of popularity than social factors, and
- their combination yields better performance depending on the distribution of popularity.

5.6 Discussion

Apart from factorizing content and social influence, the learned models can be used to predict the popularity of photos. Figure 5.8 shows an example of

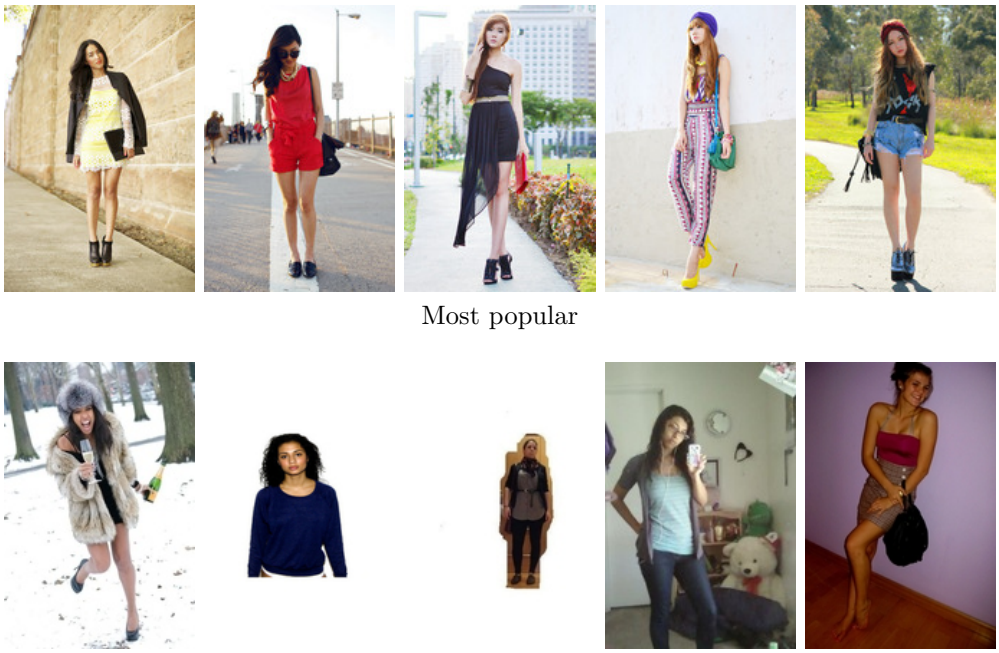


Figure 5.8: Prediction examples.

the most and least popular pictures predicted by one of the in-network models with both social and content factors. There is clearly a distinction in visual quality between the most and the least popular pictures. Perhaps it is also possible to build a system that can predict *unbiased* content popularity. Such prediction could be useful for many e-commerce applications, such as automatic outfit quality feedback [64] and socially-aware fashion recommendation [66]. Stable popularity prediction can benefit in online ad optimization and traffic balancing. It is an interesting future work to use this insight to build a socially-aware multimedia system.

The findings of strong social-influence on popularity implies that any attempt to learning subjective measures such as aesthetics or interestingness from an online community should explicitly consider social bias in the data.

For example, learning a regression function to evaluate content quality that incorporates social information leads to a much better model for popularity prediction. It also indicates that researchers should be careful to learn general content-quality measures useful outside of a particular social network, to learn from data free of social influence.

5.7 Summary

This chapter presented a multimodal approach to quantitatively model a picture shared on an online fashion network and analyzed its usefulness for predicting network popularity. The content model takes advantage of various sources of information, including computer vision, natural language processing, social network information, and other content metadata. Through correlation, regression, and classification analyses in both in-network and out-of-network conditions, it was shown that there is statistical evidence that content popularity under network is mostly the outcome of the social network itself regardless of content quality. With the in-depth analysis under out-of-network condition, the study also finds that social factors can actually serve as a predicate of unbiased content popularity that is rather complementary to the direct content representation provided by computer-vision and natural-language processing.

The study suggests that social factors should be carefully considered for research involving social network photos. It would be interesting to apply the insights from this study in various applications, including analyzing temporal trends of user fashion contents, combining sentiment analysis in reasoning about user behavior, and building a retrieval system that explicitly takes into

account both socially biased and unbiased popularity. Another possible direction is to infer individual preferences from user behavior and content popularity, e.g., using regression and classification techniques to learn and model preference functions implicit in clothing choices and popularity votes, to predict group behavior. Once popularity is predicted, predictions can be used in optimizing Web traffic or advertisement. Popular posts naturally attract more traffic. Thus, popularity prediction benefits in load balancing or online advertisement.

Chapter 6

Conclusion

This dissertation studied the computational approach to understand online social visual networks focused on fashion. The study explored two distinct aspects of the computational understanding of online visual networks.

The first was a computational approach to visual content understanding in fashion images. The goal in this problem was recognition of clothing items. To this goal, two approaches were explored; one that tries to locate items knowing the kind of clothing appearing in the picture (localization), and the other that also detects items without any information about the picture (detection). The first approach formulates the recognition problem using a conditional random field, where human pose estimation is explicitly modeled as an conditional input to the image segmentation problem. The detection approach takes advantage of a large scale dataset of fashion images from the online fashion network, where users provide noisy text annotations of their garment items. This data-driven method showed a successful result in clothing parsing under the challenging detection scenario. The extended experimental results also showed

that the proposed clothing recognition framework helps human pose estimation, by considering clothing recognition as an contextual input. This clothing recognition framework serves as a fundamental component in building wide range of applications, including clothing retrieval, fashion recommendation, social identity recognition, person identification, or fashion trend analysis.

This dissertation also studied the other understanding problem, behavior understanding in the network, in the form of photo-popularity analysis. One might think visual content is the trigger for user action in online visual networks. However, the study in this dissertation statistically revealed that the network itself is the dominant factor in user decision, and visual content has quantitatively a much smaller impact on photo popularity. To get more insights from this observation, additional popularity data are collected by emulating *out-of-network* condition using crowdsourcing. The results indicated that under no network influence, user behavior depends more on visual content. However, the results also suggested that social factors contain information that is not captured by content-only modeling. The insights obtained in this study are useful in building many applications, such as automatic fashion feedback, estimation of unbiased popularity, or load balancing or advertisement optimization based on popularity prediction.

The computational framework proposed in this dissertation establishes the fundamental methodology in analyzing online visual networks.

Bibliography

[1] Karteek Alahari, Guillaume Seguin, Josef Sivic, Ivan Laptev, et al. Pose estimation and segmentation of people in 3d movies. In *ICCV 2013-IEEE International Conference on Computer Vision*, 2013.

[2] Dragomir Anguelov, Kuang-chih Lee, SB Gokturk, and Baris Sumengen. Contextual identity recognition in personal photo albums. In *Computer Vision and Pattern Recognition, 2007. CVPR'07. IEEE Conference on*, pages 1–7. IEEE, 2007.

[3] P. Arbelaez, M. Maire, C. Fowlkes, and J. Malik. Contour detection and hierarchical image segmentation. *PAMI*, 33(5):898 –916, 2011.

[4] S. Bakhshi, D. Shamma, and Eric Gilbert. Faces engage us: Photos with faces attract more likes and comments on instagram. *CHI*, 2014.

[5] Albert-László Barabási and Réka Albert. Emergence of scaling in random networks. *science*, 286(5439):509–512, 1999.

[6] Tamara L Berg, Alexander C Berg, and Jonathan Shih. Automatic attribute discovery and characterization from noisy web data. In *Computer Vision–ECCV 2010*, pages 663–676. Springer, 2010.

[7] J Biel and Daniel Gatica-Perez. The youtube lens: Crowdsourced personality impressions and audiovisual analysis of vlogs. *Multimedia*, 15(1):41–55, 2013.

[8] Eran Borenstein and Jitendra Malik. Shape guided object segmentation. In *CVPR*, volume 1, pages 969–976, 2006.

[9] Youmna Borghol, Sebastien Ardono, and Niklas Carlsson. The untold story of the clones: Content-agnostic factors that impact youtube video popularity. *KDD*, 2012.

[10] Agnés Borràs, Francesc Tous, Josep Lladós, and Maria Vanrell. High-level clothes description based on colour-texture and structural features. In *Pattern Recognition and Image Analysis*, pages 108–116. Springer Berlin / Heidelberg, 2003.

[11] Lukas Bossard, Matthias Dantone, Christian Leistner, Christian Wengert, Till Quack, and Luc Van Gool. Apparel classification with style. *ACCV*, pages 1–14, 2012.

[12] Lubomir Bourdev, Subhransu Maji, Thomas Brox, and Jitendra Malik. Detecting people using mutually consistent poselet activations. In *ECCV*, 2010.

[13] Lubomir Bourdev, Subhransu Maji, and Jitendra Malik. Describing people: A poselet-based approach to attribute classification. In *ICCV*, pages 1543–1550, 2011.

[14] Lubomir Bourdev and Jitendra Malik. Poselets: Body part detectors trained using 3D human pose annotations. In *ICCV*, 2009.

[15] Yuri Boykov and Vladimir Kolmogorov. An experimental comparison of min-cut/max-flow algorithms for energy minimization in vision. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 26(9):1124–1137, 2004.

[16] Yuri Boykov, Olga Veksler, and Ramin Zabih. Fast approximate energy minimization via graph cuts. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 23(11):1222–1239, 2001.

[17] Anders Brodersen, Salvatore Scellato, and Mirjam Wattenhofer. Youtube around the world: geographic popularity of videos. In *WWW*, pages 241–250. ACM, 2012.

[18] Alexandru O. Bălan and Michael J. Black. The naked truth: Estimating body shape under clothing. *ECCV*, pages 15–29, 2008.

[19] Meeyoung Cha, Haewoon Kwak, Pablo Rodriguez, Yong-Yeol Ahn, and Sue Moon. I tube, you tube, everybody tubes: analyzing the world's largest user generated content video system. In *IMC*. ACM, 2007.

[20] Shuo Chang, Vikas Kumar, Eric Gilbert, and Loren G. Terveen. Specialization, homophily, and gender in a social curation site: Findings from pinterest. In *CSCW*, pages 674–686. ACM, 2014.

[21] Hong Chen, Zi Jian Xu, Zi Qiang Liu, and Song Chun Zhu. Composite templates for cloth modeling and sketching. In *CVPR*, 2006.

[22] Huizhong Chen, Andrew Gallagher, and Bernd Girod. Describing clothing by semantic attributes. In *ECCV*, pages 609–623. 2012.

[23] Qi Chen, Gang Wang, and Chew Lim Tan. Modeling fashion. In *Multimedia and Expo (ICME), 2013 IEEE International Conference on*, pages 1–6. IEEE, 2013.

[24] David J Crandall, Lars Backstrom, Daniel Huttenlocher, and Jon Kleinberg. Mapping the world's photos. In *WWW*, pages 761–770. ACM, 2009.

[25] George A Cushen and Mark S Nixon. Mobile visual clothing search. In *Multimedia and Expo Workshops (ICMEW), 2013 IEEE International Conference on*, pages 1–6. IEEE, 2013.

[26] Navneet Dalal and Bill Triggs. Histograms of oriented gradients for human detection. In *CVPR*, volume 1, pages 886–893. IEEE, 2005.

[27] Matthias Dantone, Juergen Gall, Christian Leistner, and Luc Van Gool. Human pose estimation using body parts dependent joint regressors. In *Computer Vision and Pattern Recognition (CVPR), 2013 IEEE Conference on*, pages 3041–3048. IEEE, 2013.

[28] Sagnik Dhar, Vicente Ordonez, and Tamara L Berg. High level describable attributes for predicting aesthetics and interestingness. In *CVPR*, pages 1657–1664. IEEE, 2011.

[29] Wei Di, Catherine Wah, Anurag Bhardwaj, Robinson Piramuthu, and Neel Sundaresan. Style finder: Fine-grained clothing style detection

and retrieval. In *Computer Vision and Pattern Recognition Workshops (CVPRW), 2013 IEEE Conference on*, pages 8–13. IEEE, 2013.

[30] Jian Dong, Qiang Chen, Wei Xia, Zhongyang Huang, and Shuicheng Yan. A deformable mixture parsing model with parselets. *ICCV*, 2013.

[31] David Eigen and Rob Fergus. Nonparametric image parsing using adaptive neighbor sets. In *Computer Vision and Pattern Recognition (CVPR), 2012 IEEE Conference on*, pages 2799–2806. IEEE, 2012.

[32] Rong-En Fan, Kai-Wei Chang, Cho-Jui Hsieh, Xiang-Rui Wang, and Chih-Jen Lin. LIBLINEAR: A library for large linear classification. *J Machine Learning Research*, 9:1871–1874, 2008.

[33] Pedro F Felzenszwalb and Daniel P Huttenlocher. Efficient graph-based image segmentation. *IJCV*, 59(2):167–181, 2004.

[34] Pedro F. Felzenszwalb and Daniel P. Huttenlocher. Pictorial structures for object recognition. *Int. J. Computer Vision*, 61:55–79, 2005.

[35] P.F. Felzenszwalb, R.B. Girshick, D. McAllester, and D. Ramanan. Object detection with discriminatively trained part-based models. *PAMI*, 32(9):1627 –1645, 2010.

[36] P.F. Felzenszwalb and D.P. Huttenlocher. Efficient matching of pictorial structures. In *CVPR*, pages 66 –73, 2000.

[37] V. Ferrari, M. Marin-Jimenez, and A. Zisserman. Progressive search space reduction for human pose estimation. In *CVPR*, 2008.

[38] V. Ferrari, M. Marin-Jimenez, and A. Zisserman. Pose search: Retrieving people using their pose. In *CVPR*, 2009.

[39] Andrew C Gallagher and Tsuhan Chen. Clothing cosegmentation for recognizing people. In *CVPR*, pages 1–8, 2008.

[40] Eric Gilbert, Saeideh Bakhshi, Shuo Chang, and Loren Terveen. I need to try this?: a statistical overview of pinterest. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, pages 2427–2436. ACM, 2013.

[41] Georgia Gkioxari, Pablo Arbeláez, Lubomir Bourdev, and Jitendra Malik. Articulated pose estimation using discriminative armlet classifiers. In *Computer Vision and Pattern Recognition (CVPR), 2013 IEEE Conference on*, pages 3342–3349. IEEE, 2013.

[42] Stephen Gould, Tianshi Gao, and Daphne Koller. Region-based segmentation and object detection. In *NIPS*, 2009.

[43] Peng Guan, Oren Freifeld, and Michael J. Black. A 2D human body model dressed in eigen clothing. *ECCV*, pages 285–298, 2010.

[44] Basela Hasan and David Hogg. Segmentation using deformable spatial priors with application to clothing. In *BMVC*, 2010.

[45] Nils Hasler, Carsten Stoll, Bodo Rosenhahn, Thorsten Thormählen, and Hans-Peter Seidel. Estimating body shape of dressed humans. *Computers and Graphics*, 33(3):211 – 216, 2009.

[46] K. Hevner. Experimental studies of the affective value of colors and lines. *Journal of Applied Psychology*, 19(4):385–398, August 1935.

[47] Xin Jin, Andrew Gallagher, Liangliang Cao, Jiebo Luo, and Jiawei Han. The wisdom of social multimedia: using flickr for prediction and forecast. In *Multimedia*, pages 1235–1244. ACM, 2010.

[48] Yannis Kalantidis, Lyndon Kennedy, and Li-Jia Li. Getting the look: clothing recognition and segmentation for automatic product suggestions in everyday photos. In *Proceedings of the 3rd ACM conference on International conference on multimedia retrieval*, pages 105–112. ACM, 2013.

[49] Aditya Khosla, Atish Das Sarma, and Raffay Hamid. What makes an image popular? *WWW*, 2014.

[50] Akisato Kimura, Katsuhiko Ishiguro, Makoto Yamada, Alejandro Marcos Alvarez, Kaori Kataoka, and Kazuhiko Murasaki. Image context discovery from socially curated contents. In *ACM Multimedia*, pages 565–568, 2013.

[51] Pushmeet Kohli, Jonathan Rihan, Matthieu Bray, and Philip HS Torr. Simultaneous segmentation and pose estimation of humans using dynamic graph cuts. *International Journal of Computer Vision*, 79(3):285–298, 2008.

[52] Pushmeet Kohli, Philip HS Torr, et al. Robust higher order potentials for enforcing label consistency. *International Journal of Computer Vision*, 82(3):302–324, 2009.

[53] Vladimir Kolmogorov and Ramin Zabin. What energy functions can be minimized via graph cuts? *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 26(2):147–159, 2004.

[54] Adriana Kovashka, Devi Parikh, and Kristen Grauman. Whittlesearch: Image search with relative attribute feedback. In *Computer Vision and Pattern Recognition (CVPR), 2012 IEEE Conference on*, pages 2973–2980. IEEE, 2012.

[55] Iljung S Kwak, Ana C Murillo, Peter N Belhumeur, David Kriegman, and Serge Belongie. From bikers to surfers: Visual recognition of urban tribes. *BMVC*, 2013.

[56] Lubor Ladicky, Philip HS Torr, and Andrew Zisserman. Human pose estimation using a joint pixel-wise and part-wise formulation. In *Computer Vision and Pattern Recognition (CVPR), 2013 IEEE Conference on*, pages 3578–3585. IEEE, 2013.

[57] Himabindu Lakkaraju, Julian McAuley, and Jure Leskovec. What's in a name? Understanding the interplay between titles, content, and communities in social media. *ICWSM*, 2013.

[58] Jong Gun Lee, Sue Moon, and Kavé Salamatian. An approach to model and predict the popularity of online contents with explanatory factors. In *WI-IAT*, volume 1, pages 623–630. IEEE, 2010.

[59] Bastian Leibe, Aleš Leonardis, and Bernt Schiele. Robust object detection with interleaved categorization and segmentation. *IJCV*, 77(1-3):259–289, 2008.

[60] Kristina Lerman and Tad Hogg. Using a model of social dynamics to predict popularity of news. In *WWW*, pages 621–630. ACM, 2010.

[61] Kristina Lerman and Laurie A Jones. Social browsing on flickr. *ICWSM*, 2007.

[62] Ce Liu, Jenny Yuen, and Antonio Torralba. Nonparametric scene parsing: Label transfer via dense scene alignment. In *Computer Vision and Pattern Recognition, 2009. CVPR 2009. IEEE Conference on*, pages 1972–1979. IEEE, 2009.

[63] Ce Liu, Jenny Yuen, and Antonio Torralba. Sift flow: Dense correspondence across scenes and its applications. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 33(5):978–994, 2011.

[64] Luoqi Liu, Hui Xu, Junliang Xing, Si Liu, Xi Zhou, and Shuicheng Yan. Wow! you are so beautiful today! In *ACM Multimedia*, pages 3–12, 2013.

[65] Si Liu, Jiashi Feng, Csaba Domokos, Hui Xu, Junshi Huang, Zhenzhen Hu, and Shuicheng Yan. Fashion parsing with weak color-category labels. *IEEE Transactions on Multimedia*, 16(1), January 2014.

[66] Si Liu, Jiashi Feng, Zheng Song, Tianzhu Zhang, Hanqing Lu, Changsheng Xu, and Shuicheng Yan. Hi, magic closet, tell me what to wear! In *ACM international conference on Multimedia*, pages 619–628. ACM, 2012.

[67] Si Liu, Zheng Song, Guangcan Liu, Changsheng Xu, Hanqing Lu, and Shuicheng Yan. Street-to-shop: Cross-scenario clothing retrieval via parts alignment and auxiliary set. In *CVPR*, pages 3330–3337, 2012.

[68] Babak Loni, Maria Menendez, Mihai Georgescu, Luca Galli, Claudio Massari, Ismail Sengor Altingovde, Davide Martinenghi, Mark Melenhorst, Raynor Vliegendhart, and Martha Larson. Fashion-focused creative commons social dataset. In *Proceedings of the 4th ACM Multimedia Systems Conference*, pages 72–77. ACM, 2013.

[69] Rastislav Lukac. *Perceptual Digital Imaging: Methods and Applications*, volume 6. CRC Press, 2012.

[70] Marco Manfredi, Costantino Grana, Simone Calderara, and Rita Cucchiara. A complete system for garment segmentation and color classification. *Machine Vision and Applications*, pages 1–15, 2013.

[71] Christopher D Manning and Hinrich Schütze. *Foundations of statistical natural language processing*. MIT press, 1999.

[72] Eichner Marcin and Ferrari Vittorio. Better appearance models for pictorial structures. In *BMVC*, September 2009.

[73] Cameron Marlow, Mor Naaman, Danah Boyd, and Marc Davis. Ht06, tagging paper, taxonomy, flickr, academic article, to read. *HYPERTEXT*, pages 31–40, 2006.

[74] Marcin Marszałek and Cordelia Schmid. Accurate object recognition with shape masks. *IJCV*, 97(2):191–209, 2012.

[75] Julian McAuley and Jure Leskovec. Image labeling on a network: using social-network metadata for image classification. In *ECCV*, pages 828–841. Springer, 2012.

[76] S. Miller, M. Fritz, T. Darrell, and P. Abbeel. Parametrized shape models for clothing. In *ICRA*, 2011.

[77] Joris M. Mooij. libDAI: A free and open source C++ library for discrete approximate inference in graphical models. *JMLR*, 11:2169–2173, August 2010.

[78] Louis-Philippe Morency, Rada Mihalcea, and Payal Doshi. Towards multimodal sentiment analysis: Harvesting opinions from the web. In *ICMI*, pages 169–176. ACM, 2011.

[79] G. Mori, Xiaofeng Ren, A.A. Efros, and J. Malik. Recovering human body configurations: combining segmentation and recognition. In *CVPR*, 2004.

[80] Ana C Murillo, Iljung S Kwak, Lubomir Bourdev, David Kriegman, and Serge Belongie. Urban tribes: Analyzing group photos from a social perspective. In *CVPR Workshops*, pages 28–35, 2012.

[81] Devi Parikh and Kristen Grauman. Relative attributes. In *Computer Vision (ICCV), 2011 IEEE International Conference on*, pages 503–510. IEEE, 2011.

[82] Leonid Pishchulin, Mykhaylo Andriluka, Peter Gehler, and Bernt Schiele. Poselet conditioned pictorial structures. *CVPR*, 2013.

[83] Srikumar Ramalingam, Pushmeet Kohli, Karteek Alahari, and Philip Torr. Exact inference in multi-label crfs with higher order cliques. In *Computer Vision and Pattern Recognition, 2008. CVPR 2008. IEEE Conference on*, pages 1–8. IEEE, 2008.

[84] Deva Ramanan. Learning to parse images of articulated bodies. In *NIPS*, pages 1129–1136, 2006.

[85] Xiaofeng Ren, A.C. Berg, and J. Malik. Recovering human body configurations using pairwise constraints between parts. In *ICCV*, 2005.

[86] Carsten Rother, Vladimir Kolmogorov, and Andrew Blake. Grabcut: Interactive foreground extraction using iterated graph cuts. In *ACM Transactions on Graphics (TOG)*, volume 23, pages 309–314. ACM, 2004.

[87] Bryan C Russell, Alexei Efros, Josef Sivic, William T Freeman, and Andrew Zisserman. Segmenting scenes by matching image composites. 2009.

[88] Matthew J Salganik, Peter Sheridan Dodds, and Duncan J Watts. Experimental study of inequality and unpredictability in an artificial cultural market. *science*, 311(5762):854–856, 2006.

[89] Jose San Pedro and Stefan Siersdorfer. Ranking and classifying attractiveness of photos in folksonomies. In *WWW*, pages 771–780. ACM, 2009.

[90] Jose San pedro and Poonam Suryanarayan. Your opinion counts!: leveraging social comments for analyzing aesthetic perception of photographs. In *CHI*, pages 2519–2522. ACM, 2012.

[91] Jose San Pedro, Tom Yeh, and Nuria Oliver. Leveraging user comments for aesthetic aware image search reranking. In *WWW*, pages 439–448. ACM, 2012.

[92] Ming Shao, Liangyue Li, and Yun Fu. What do you do? occupation recognition in a photo via social context. *ICCV*, 2013.

[93] Jamie Shotton, John Winn, Carsten Rother, and Antonio Criminisi. Textonboost: Joint appearance, shape and context modeling for multi-class object recognition and segmentation. *ECCV*, pages 1–15, 2006.

[94] Stefan Siersdorfer, Sergiu Chelaru, Wolfgang Nejdl, and Jose San Pedro. How useful are your comments?: analyzing and predicting youtube comments and comment ratings. In *Proceedings of the 19th international conference on World wide web*, pages 891–900. ACM, 2010.

[95] Gautam Singh and Jana Košecká. Semantic context for nonparametric scene parsing and scene classification. *CVPR*, 2013.

[96] J. Sivic, C. L. Zitnick, and R. Szeliski. Finding people in repeated shots of the same scene. In *BMVC*, 2006.

[97] Malcolm Slaney. Web-scale multimedia analysis: does content matter? *MultiMedia, IEEE*, 18(2):12–15, 2011.

[98] Zheng Song, Meng Wang, Xian-sheng Hua, and Shuicheng Yan. Predicting occupation via human clothing and contexts. In *ICCV*, pages 1084–1091, 2011.

[99] Zak Stone, Todd Zickler, and Trevor Darrell. Autotagging facebook: Social network context improves photo annotation. In *Computer Vision and Pattern Recognition Workshops, 2008. CVPRW'08. IEEE Computer Society Conference on*, pages 1–8. IEEE, 2008.

[100] Gabor Szabo and Bernardo A Huberman. Predicting the popularity of online content. *Communications of the ACM*, 53(8):80–88, 2010.

[101] Joseph Tighe and Svetlana Lazebnik. Superparsing: scalable nonparametric image parsing with superpixels. *ECCV*, pages 352–365, 2010.

[102] Joseph Tighe and Svetlana Lazebnik. Finding things: Image parsing with regions and per-exemplar detectors. *CVPR*, 2013.

[103] Michele Trevisiol, Luca Chiarandini, Luca Maria Aiello, and Alejandro Jaimes. Image ranking based on user browsing behavior. In *SIGIR*, pages 445–454. ACM, 2012.

[104] Hang M Ung. Social influence, popularity and interestingness of online contents. In *ICWSM*, 2011.

[105] Roelof van Zwol, Adam Rae, and Lluis Garcia Pueyo. Prediction of favourite photos using social, visual, and textual signals. In *Multimedia*, pages 1015–1018. ACM, 2010.

[106] Manik Varma and Andrew Zisserman. A statistical approach to texture classification from single images. *Int. J. Computer Vision*, 62(1-2):61–81, 2005.

[107] A. Vedaldi and B. Fulkerson. VLFeat: An open and portable library of computer vision algorithms, 2008.

[108] Nan Wang and Haizhou Ai. Who blocks who: Simultaneous clothing segmentation for grouping images. In *ICCV*, pages 1535–1542, 2011.

[109] Xianwang Wang, Tong Zhang, D.R. Tretter, and Qian Lin. Personal clothing retrieval on photo collections by color and attributes. *Multimedia, IEEE Transactions on*, 15(8):2035–2045, Dec 2013.

[110] Michael Weber, Martin Bäuml, and Rainer Stiefelhagen. Part-based clothing segmentation for person retrieval. In *AVSS*, 2011.

[111] Kota Yamaguchi, M Hadi Kiapour, Luis E Ortiz, and Tamara L Berg. Parsing clothing in fashion photographs. In *CVPR*, pages 3570–3577, 2012.

[112] Ming Yang and Kai Yu. Real-time clothing recognition in surveillance videos. In *ICIP*, 2011.

[113] Yi Yang and Deva Ramanan. Articulated pose estimation with flexible mixtures-of-parts. In *CVPR*, pages 1385–1392, 2011.

[114] Haipeng Zhang, Mohammed Korayem, David J Crandall, and Gretchen LeBuhn. Mining photo-sharing websites to study ecological phenomena. In *WWW*, pages 749–758. ACM, 2012.

[115] Xianjun Sam Zheng, Ishani Chakraborty, James Jeng-Weei Lin, and Robert Rauschenberger. Correlating low-level image statistics with users-rapid aesthetic and affective judgments of web pages. In *CHI*, pages 1–10. ACM, 2009.

[116] Silvia Zuffi, Oren Freifeld, and Michael J Black. From pictorial structures to deformable structures. In *Computer Vision and Pattern Recognition (CVPR), 2012 IEEE Conference on*, pages 3546–3553. IEEE, 2012.