Collaborative Clothing Segmentation and Identification Through Image Analysis

Abstract

This research introduces a comprehensive clothing co-parsing system designed to analyze a collection of clothing images, which are unsegmented but include descriptive tags. The system aims to segment these images into meaningful configurations. The proposed method uses a two-stage, data-driven approach. The first stage, termed "image co-segmentation," iteratively refines image regions, using the exemplar-SVM (E-SVM) method to enhance region consistency across images. The second stage, "region co-labeling," utilizes a multi-image graphical model where segmented regions serve as nodes. This incorporates contextual information about clothing, such as item placement and interactions, which can be solved using the efficient Graph Cuts algorithm. The system's performance is tested on the Fashionista dataset and a newly developed dataset called CCP, which contains 2098 high-resolution street fashion images. The results show a segmentation accuracy of 90.29% and 88.23% and a recognition rate of 65.52% and 63.89% on the Fashionista and CCP datasets, respectively, demonstrating an improvement over current leading methods.

1 Introduction

The growth of online clothing sales has increased the demand for accurate clothing recognition and retrieval technologies. This has led to the development of several vision-based solutions. A key challenge in these systems is the detailed, pixel-level labeling of clothing, which is often resource-intensive. However, image-level tags from user data offer a viable alternative. This paper focuses on the development of a system to segment clothing images and assign semantic labels to these segments.

The main contribution of this work is an effective system for parsing groups of clothing images and providing precise pixel-level annotations. The system addresses the following significant challenges:

- Clothes exhibit a wide variety of styles and textures, making them difficult to segment and identify using only basic visual features.
- Variations in human poses and the way clothes can obscure themselves complicate the recognition process.
- The existence of numerous, highly specific clothing categories, such as over 50 in the Fashionista dataset, far more than in existing co-segmentation systems which typically handle fewer categories.

To overcome these challenges, the system employs two sequential stages: image co-segmentation to isolate distinct clothing regions and region co-labeling to identify different clothing items, as illustrated below. It also utilizes contextual cues related to how clothing items are typically arranged and related to each other.

The co-segmentation phase refines regions across images using the E-SVM method. Initially, images are divided into superpixels, which are then grouped into regions. Many of these regions may not be

.

meaningful due to the diversity of clothing and human poses. However, certain stable regions are identified based on criteria like size and position. E-SVM classifiers are trained for these selected regions using HOG features, creating region-based detectors that help identify similar regions across images. This approach is based on the observation that similar clothing items often share visual patterns.

The co-labeling phase uses a data-driven approach, constructing a multi-image graph where regions are treated as nodes. Connections are made between adjacent regions within an image, as well as between regions in different images that share visual or tag similarities. This strategy allows for collective label assignment, leveraging similarities across images. The optimization is performed using the Graph Cuts algorithm, considering various clothing context constraints.

2 Related Work

Previous research on clothing and human segmentation has often focused on creating detailed models to handle the diversity in clothing styles and appearances. Some of the classic work used And-Or graph templates to model and parse clothing configurations. Subsequent studies explored blocking models for segmenting clothes in images where items were heavily obscured, and deformable spatial models to enhance segmentation accuracy. Recent approaches have used shape-based human models or combined pose estimation with supervised region labeling, achieving notable results. However, these methods have not been applied to clothing co-parsing and typically demand significant labeling effort.

Research on image/object co-labeling, which jointly processes a set of images containing similar objects, has been explored. Methods include unsupervised shape-guided approaches for single-category co-labeling and incorporating automatic image segmentation with spatially coherent latent topic models for unsupervised multi-class labeling. These unsupervised methods can struggle with a large number of categories and diverse appearances. Recent efforts have focused on supervised label propagation, using pixel-level label maps to assign labels to new images. However, these methods are often limited by the need for detailed annotations and rely on pixel-level correspondences, which may not be effective for clothing parsing.

3 Methodology

This research introduces a probabilistic model for the co-parsing of clothing images. The input consists of a set of clothing images, denoted as $I = \{I_i\}_{i=1}^{N}$, each associated with tags T_i. Each image I_i is represented by a set of superpixels, $I_i = \{s_j\}_{j=1}^{M}$, which are subsequently grouped into coherent regions. Each image is associated with four additional variables:

(a) Regions $\{r_k\}_{k=1}^{K}$, each comprising a set of superpixels.

(b) Garment labels for each region, denoted as $\ell_k \in T$, where $k = 1, \ldots, K$.

(c) E-SVM weights w_k trained for each selected region.

(d) Segmentation propagations $C = (x, y, m)$, where (x, y) is the location and m is the segmentation mask of an E-SVM, indicating that mask m can be propagated to position (x, y) of I_i.

The objective is to optimize parameters by maximizing the posterior probability:

$$\{L^*, R^*, W^*, C^*\} = \arg\max P(L, R, W, C|I)$$

This probability can be factorized into co-labeling and co-segmentation components:

$$P(L, R, W, C|I) \propto P(L|R, C) \times \prod_{i=1}^{N} P(R_i|C_i, I_i)P(W_i|R_i)P(C_i|W_i, I_i)$$

The optimization process involves two phases: clothing image co-segmentation and co-labeling.

In the co-segmentation phase, optimal regions are obtained by maximizing $P(R|C,I)$. A superpixel grouping indicator $o_j \in \{1,\ldots,K\}$ is introduced, indicating the region to which superpixel s_j belongs. Each region r_k is defined as $r_k = \{s_j|o_j = k\}$. The probability $P(R|C,I)$ is defined as:

$$P(R|C,I) = \prod_i \left[P(r_i|C,I) \prod_{s_j \in I_i} P(o_j|C,I_i) \prod_{(m,n)} P(o_m, o_n, s_m, s_n|C) \right]$$

The unary potential $P(o_j, s_j)$ indicates the probability of superpixel s_j belonging to a region, and the pairwise potential $P(o_m, o_n, s_m, s_n|C)$ encourages smoothness between neighboring superpixels.

Coherent regions are selected to train E-SVMs by maximizing $P(W|R)$:

$$P(W|R) = \prod_k P(w_k|r_k) \propto \prod_k \exp\{-E(w_k, r_k) - \phi(r_k)\}$$

where $\phi(r_j)$ indicates whether r_j has been chosen for training E-SVM, and $E(w_k, r_k)$ is the convex energy function of E-SVM.

Finally, $P(C_i|W_i, I_i)$ is defined based on the responses of E-SVM classifiers, maximized by selecting the top k detections of each E-SVM as segmentation propagations.

In the co-labeling phase, a multi-image graphical model is used to assign a garment tag to each region:

$$P(L|R,C) \propto \prod_i^N \prod_k^K \left[P(\ell_{ik}, r_i) \prod_{(m,n)} P(\ell_{im}, \ell_{in}, r_i, r_j) \prod_{(u,v)} Q(\ell_{iu}, \ell_{iv}, r_u, r_v|C) \right]$$

where $P(\ell_{ik}, r_i)$ is the singleton potential, $P(\ell_{im}, \ell_{in}, r_i, r_j)$ is the interior affinity model, and $Q(\ell_{iu}, \ell_{iv}, r_u, r_v|C)$ is the exterior affinity model.

3.1 Unsupervised Image Co-Segmentation

The co-segmentation process involves iteratively refining regions, E-SVM weights, and segmentation propagations.

Superpixel Grouping: A linear programming problem is formulated to determine the number of regions automatically:

$$\arg\min \sum_e d(s_{e1}, s_{e2}) o_e + \sum_{c \in C} h(c) o_c$$

where $d(s_{e1}, s_{e2})$ is the dissimilarity between superpixels and $h(c)$ measures the consistency of grouping superpixels covered by an E-SVM mask.

Training E-SVMs: The energy function for training E-SVMs is:

$$E(w_k, r_k) = \frac{\lambda_1}{2}||w_k||^2 + \sum_{s_j \in r_k} \max(0, 1 - w_k^T f(s_j)) + \lambda_2 \sum_{s_n \in NE} \max(0, 1 + w_k^T f(s_n))$$

Segmentation Propagation: The E-SVM response is calibrated using a logistic distribution:

$$S_E(f; w) = \frac{1}{1 + \exp(-\alpha_E(w^T f - \beta_E))}$$

3.2 Contextualized Co-Labeling

In this phase, a multi-image graphical model connects all images, incorporating two types of clothing contexts. The singleton potential is defined as:

$$P(\ell_k, r_k) = sig(S(f(r_k), \ell_k)) \cdot G_{\ell_k}(X_k)$$

where $S(f(r_k), \ell_k)$ is the appearance model score and $G_{\ell_k}(X_k)$ is the location context.

The interior affinity model is:

$$P(\ell_{im}, \ell_{in}, r_m, r_n) = \phi(\ell_{im}, \ell_{in}, r_m, r_n) \cdot U(\ell_{im}, \ell_{in})$$

and the exterior affinity model is:

$$Q(\ell_{iu}, \ell_{iv}, r_u, r_v|C) = G_{\ell_{iu}}(X_u) \cdot G_{\ell_{iv}}(X_v) \cdot \phi(\ell_{iu}, \ell_{iv}, r_u, r_v)$$

4 Experiments

The framework is evaluated on two datasets: Clothing Co-Parsing (CCP) and Fashionista. CCP includes 2,098 high-resolution fashion photos with extensive variations in human appearance and clothing styles. The Fashionista dataset contains 158,235 fashion photos, with a subset of 685 images annotated at the superpixel level.

4.1 Quantitative Evaluation

The method is compared with three state-of-the-art methods: PECS, Bi-layer Sparse Coding (BSC), and Semantic Texton Forest (STF). Performance is measured using average Pixel Accuracy (aPA) and mean Average Garment Recall (mAGR).

Table 1: Clothing parsing results (%) on the Fashionista and CCP datasets.

2*Methods	Fashionista		CCP	
	aPA	mAGR	aPA	mAGR
Ours-full	90.29	65.52	88.23	63.89
PECS	89.00	64.37	85.97	51.25
BSC	82.34	33.63	81.61	38.75
STF	68.02	43.62	66.85	40.70
Ours-1	89.69	61.26	87.12	61.22
Ours-2	88.55	61.13	86.75	59.80
Ours-3	84.44	47.16	85.43	42.50
Baseline	77.63	9.03	77.60	15.07

The proposed method outperforms BSC, STF, and PECS on both datasets, demonstrating the effectiveness of the iterative co-segmentation and co-labeling phases.

5 Conclusion

This paper presents a framework for jointly parsing a collection of clothing images using image-level tags. The framework includes a new dataset of high-resolution street fashion photos with detailed annotations. The experiments show that the proposed method is effective and performs favorably compared to existing methods. Future work will focus on improving inference by iterating between the two phases and exploring parallel implementations for large-scale applications.