Fashion Parsing with Video Context

Si Liu*[1], Xiaodan Liang*[2,1] , Luoqi Liu[1], Ke Lu[3], Liang Lin[2], Shuicheng Yan[1]
[1]National University of Singapore
[2]Sun Yat-Sen University
[3]University of Chinese Academy of Sciences
{dcslius, eleyans}@nus.edu.sg,
{xdliang328, llq667}@gmail.com, {linliang}@ieee.org, {luk}@ucas.ac.cn,

ABSTRACT

In this paper, we explore how to utilize the video context to facilitate fashion parsing. Instead of annotating a large amount of fashion images, we present a general, affordable and scalable solution, which harnesses the rich contexts in easily available fashion videos to boost any existing fashion parser. First, we crawl a large unlabelled fashion video corpus with fashion frames. Then for each fashion video, the cross-frame contexts are utilized for human pose co-estimation, and then video co-parsing to obtain satisfactory fashion parsing results for all frames. More specifically, Sift Flow and super-pixel matching are used to build correspondences across frames, and these correspondences then contextualize the pose estimations and fashion parsing in individual frames. Finally, these parsed video frames are used as the reference corpus for the non-parametric fashion parsing component of the whole solution. Extensive experiments on two benchmark fashion datasets as well as a newly collected challenging Fashion Icon (FI) dataset demonstrate the encouraging performance gain from our general pipeline for fashion parsing.

Categories and Subject Descriptors

H.3.3 [**Information Search and Retrieval**]: Retrieval models; I.2.6 [**Learning**]: Knowledge acquisition

General Terms

Algorithms, Experimentation, Performance

Keywords

Fashion parsing, video segmentation, video context, human pose estimation

*indicates equal contribution

MM'14, November 3–7, 2014, Orlando, Florida, USA.

Figure 1: Illustration of images for fashion parsing and their parsing results obtained by our proposed system. Note that the girls are photographed with quite diverse poses, with some girl in side-view or back-view even, which are very challenging circumstances for fashion parsing. For better viewing of all figures in this paper, please see the original zoomed-in color pdf file.

1. INTRODUCTION

Fashion parsing aims to predict the label (e.g. skin, t-shirt, etc.) for each pixel in a fashion image. The performance of traditional fashion parsers are constrained by the limited training data, and thus cannot well parse the fashion images which contain human bodies in arbitrary, even exaggerated poses and various views. In this paper, we propose a fashion parsing algorithm which can potentially make use of unlimited fashion videos on the web, and thus can parse the challenging fashion images well. Some exemplar parsing results are shown in Fig. 1.

Fashion parsing has the potential to benefit a wide range of applications [2, 12, 11]in the multimedia area. For example, it can be used to analyse the numerous photos shared by users on social networks, so that the information about the users' personalities and preferences can be obtained and exploited for friend recommendation, advertisement, etc. Besides, fashion parsing can also be used in intelligent surveillance, e.g., person re-identification. A good understanding of one's apparel may provide useful cues to identify a person. Despite its various potential applications, current fashion parsing systems [17, 3, 16, 10] still have many limitations,

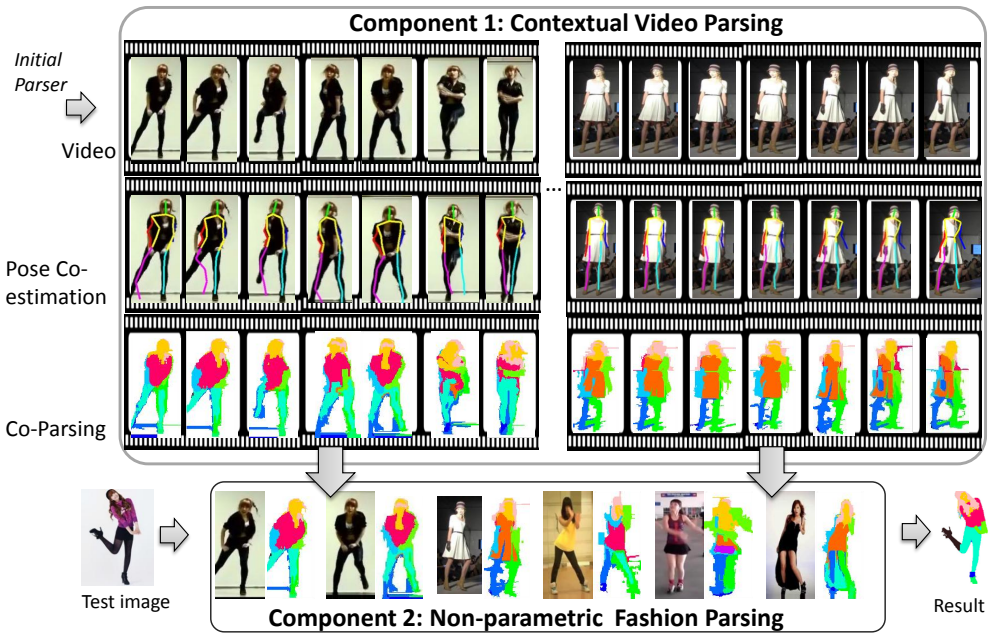

Figure 2: Framework overview of our system. It contains two components, i.e., contextual video parsing and non-parametric fashion parsing. Any off-the-shelf constrained pose estimator and fashion parser can initialize the human pose estimation and parsing results. To leverage the video contexts represented by Sift Flow and super-pixel matching, the videos have much better human pose estimation and fashion parsing results by the proposed video human pose co-estimation and fashion co-parsing algorithms. Since no supervision is required for the video data, a large number of video frames and their video parsing results can be obtained and used as a gallery set to facilitate the non-parametric label transfer to the test image.

since they cannot handle fashion parsing well, which is the common case in real life.

To solve fashion parsing in the wild, there are mainly two challenges. Firstly, there is a great diversity in human poses, which results in the unsatisfactory performance of even the state-of-the-art human pose estimator [18]. As an important embedded component of all fashion parsers, the poor performance of a human pose estimator may greatly degrade the fashion parsing results. Secondly, there has been no dataset specially designed for the fashion parsing task. All existing publicly available fashion parsing datasets [16, 10, 16] contain only constrained images. Lack of training data makes training an unconstrained fashion parser impossible.

In this paper, we propose a novel framework which leverages video contexts to tackle the fashion parsing task without extra annotation. Here, the videos are not required to be labelled, thus can be easily obtained from the web (e.g., youtube.com). The framework is illustrated in Fig 2. It contains two components: 1) contextual video parsing, and 2) non-parametric fashion parsing. As for the contextual video parsing component, the goal is to parse the unlabelled unconstrained videos which are also assumed to be in the wild, yet with valuable cross-frame contexts. Since human pose estimation is the prerequisite for fashion parsing, a human pose co-estimation step (Sec. 4.1.1) is implemented first in the contextual video parsing component. In this step, we first apply the off-the-shelf constrained human pose estimator to the videos, and then refine the estimated results by incorporating the pixel-level correspondences between sequential frames, which are described by Sift Flow [9]. After the human pose co-estimation, we first apply a constrained fashion parser (pre-trained on a small amount of constrained labeled data). It can provide a rough initialization for unconstrained video parsing. Then we use the super-pixel-level correspondences between two sequential frames, which are described by a super-pixel matching technique to refine the parsing results. That is, we co-parse all the frames in one video simultaneously (Sec. 4.1.2). In the fashion parsing in the wild component, these parsed video frames are used as the gallery set which transfers labels to testing images by the proposed non-parametric method (Sec. 4.2). To verify the effectiveness of the proposed framework, we conduct extensive experiments. The results proof that our framework better parses fashion images in the wild than state-of-the-art parsers.

Our whole framework is inspired by physiology. More specifically, our whole system simulates how a baby gradually gains more knowledge. Initially, a little baby has a simple and original understanding of the world (corresponding to the initial fashion parser in our paper). Then the baby gradually learns more knowledge by linking what he/she already knows with the world (corresponding to label propagation from initial fashion parser to the whole video). After gaining much more knowledge about the world (finish the video parsing), the baby can apply his/her enriched knowledge to handle the problem (using enlarged video corpus to transfer labels to testing image). The first component of our system, i.e., contextual video parsing, simulates the process that a baby gradually learns more knowledge. The second component, i.e., the non-parametric label transfer, simulates the process that determines how the baby uses the knowledge he/she learns to better solve a problem.

The contributions of this work can be summarized as:

- Our work differs from other existing fashion parsing works in that we parse fashion images with the help

of the large scale of web videos without extra annotation. Extensive experiments on three datasets prove the effectiveness of the proposed framework.
- In order to robustly parse the web videos, we leverage the rich temporal and semantic video contexts. It contains two components, i.e., the contextual video parsing and the non-parametric fashion parsing.
- We construct two new datasets, one large video dataset and one Fashion Icon (FI) dataset, which can serve as the benchmark datasets for the fashion parsing task.

2. RELATED WORK

In this section we review the recent research development in the fields of fashion parsing and video parsing sequentially.

2.1 Fashion Parsing

The first clothing parsing work was conducted by Yamaguchi *et al.* [17]. Their fashion parsing performance was not quite high due to the large human pose variation and background clutter. Later, Yamaguchi *et al.* [16] dramatically improved the fashion parsing performance by using a retrieval based approach. Their approach combines parsing from pre-trained global clothing models, local clothing models learned on the fly from retrieved examples, and transferred parse masks (paper doll item transfer) from retrieved examples. Dong *et al.* [3] used Parselets as the building blocks of the parsing model. Parselets are a group of parsable segments which can generally be obtained by low-level over-segmentation algorithms. They built a Deformable Mixture Parsing Model (DMPM) for human parsing to simultaneously handle the deformation and multimodalities of Parselets. Liu *et al.* [10] addressed the problem of automatically parsing the fashion images with weak supervision from the user-generated color-category tags. They proposed to combine the human pose estimation module, the MRF-based inference module and the category classifier learning module. However, all the existing algorithms can only parse the constrained fashion images, which are still far from practical use. To our best knowledge, we are the first to explore the fashion parsing task.

2.2 Video Parsing

The main difficulty of video parsing lies in the great burden of labelling training samples. According to the amount of required labelling, video parsing algorithms can be classified into four categories. The first category is supervised video parsing [14], which requires a large amount of labelled video data and thus is extremely tedious for human labelling. To reduce the burden of labelling, semi-supervised video segmentation [1, 15] is proposed, which reduces the labellers' burden to some extent. Later, weakly supervised video segmentation was explored [13] where semantic labels are associated with training videos but not spatially or temporally localized. Our video parsing technique belongs to the last category, i.e., unsupervised video parsing. No pixel-level labels are required when we parse the video data, thus ideally an infinite number of videos can be parsed, which partially alleviates the lack of unconstrained training data.

3. DATASET COLLECTION

We collect two datasets, including a video dataset and a Fashion Icon (FI) image dataset for our ultimate purpose of fashion parsing. The first video dataset contains $1,500$ videos. The second FI image dataset contains $1,082$ images.

Video Dataset: Three requirements are considered during the data collection. First, all the videos are HD videos so that the clothes can be seen clearly to facilitate the later fashion parsing. Second, at least one girl in the video is unoccluded from head to toe during the entire video sequence. Third, to balance the informativeness of the videos and the processing efficiency, the lengths of all video clips retained range from 2 seconds to 10 seconds. We mainly collect five categories of videos, including: 1) MVs of the singers; 2) dance teaching videos; 3) fashion shows of some clothing brands; 4) fashion TV dramas or movies; and 5) big evening parties or talent shows.

To balance the efficiency of video processing and the informativeness of videos, we sample from each video and keep the frame number in each video less than 50. In order to filter the background of each frame, all the crawled videos are automatically preprocessed to be human-centric and with less background clutter before fed to our video parsing system. The automatic preprocessing process contains two steps. We first use Grammar Models [6] to automatically detect the human body. Then, the detected human-centric bounding box is used as the seed for the tracking algorithm [7, 19]. Thus all the video frames are roughly aligned and mostly occupied by the human body, which greatly facilitates the later video parsing. Alternatively, we can detect the human body in each frame, but this solution suffers from the relatively low detection speed. We believe that using detection as the initialization for the later tracking algorithm is a balance between accuracy and efficiency. Due to the fully unsupervised processing of the videos, our video dataset is easily scalable by continuously downloading more data.

Fashion Icon (FI) Image Dataset: We collect $1,082$ images from the web to construct the Fashion Icon dataset (FI). The FI dataset is quite different from existing fashion parsing datasets [16, 3] in two aspects. Firstly, some images in FI each contain multiple humans. Secondly, the humans in the images of the FI dataset are in very diverse poses, which is more consistent with reality. In order to compare the performances of different parsing systems, the FI dataset is thoroughly labelled based on the label set defined by Dong *et al.* [3], which includes 18 categories: face, sunglass, hat, scarf, hair, upper clothes, coat, left-arm, right-arm, belt, pants, left-leg, right-leg, skirt, left-shoe, right-shoe, bag, dress and background.

4. FRAMEWORK

In this section we introduce the proposed framework. We detail the contextual video parsing component and the non-parametric fashion parsing component sequentially in the following.

4.1 Contextual Video Parsing

The goal of our contextual video parsing is to parse all the frames in each video simultaneously. The main challenge comes from the large variations in human poses and views within the video frames. The performance of existing fashion parsers often relies on a perfect human pose estimator to localize the human as well as the body parts. However, most of the previous pose estimators, limited by the small amount of training data, tend to fail in predicting arbitrary poses in images from the web. In this paper, we propose a novel generic graphical model to better infer the poses

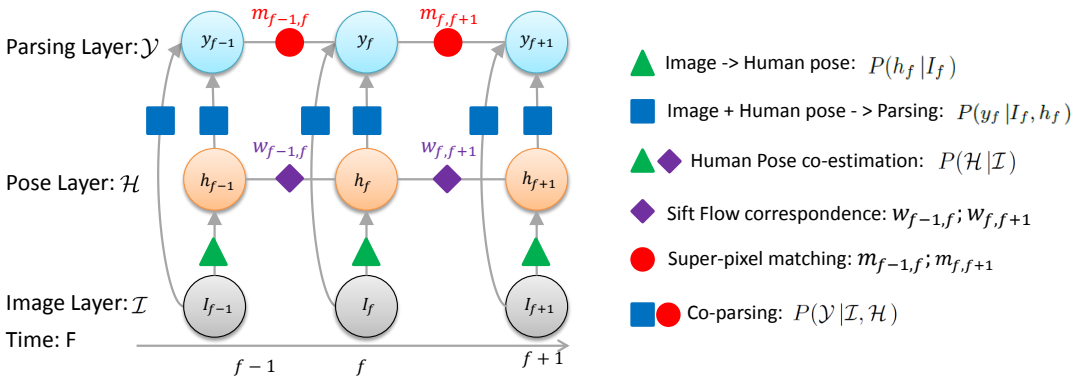

Figure 3: The whole video parsing framework. The graphical model has three layers: image layer, pose layer and parsing layer. Green triangles and blue squares respectively represent the traditional human pose estimation and fashion parsing. By incorporating Sift Flow correspondences indicated by purple diamonds, video pose co-estimation is conducted. The refined human pose results, along with the mined super-pixel matching indicated by red circles, are fed into the video co-parsing step.

and the parsing results of any frames. Intuitively, we utilize the temporal coherence and appearance consistency characteristics within video frames to refine the estimated poses and parsing results obtained from the existing constrained models. By regarding these informative contexts as the regularization constraints, the pose co-estimation and fashion co-parsing can be largely improved.

We denote the parsing results of all frames $\mathcal{I}$ as $\mathcal{Y}$ and human pose estimation results as $\mathcal{H}$. We estimate the pixel-wise semantic labelling, where the whole label set is denoted as $\mathcal{C} = \{1, \cdots, N_C\}$ and N_C is the number of labels. The three factors $(\mathcal{I}, \mathcal{Y}, \mathcal{H})$ are dependent upon each other for the fashion parsing task. Video co-parsing can be viewed as maximizing the conditional probability over parsing results $\mathcal{Y}$, human poses $\mathcal{H}$ and video frames $\mathcal{I}$, formulated as

$$P(\mathcal{H}, \mathcal{Y}|\mathcal{I}) = P(\mathcal{H}|\mathcal{I})P(\mathcal{Y}|\mathcal{I}, \mathcal{H}). \quad (1)$$

As illustrated in Fig. 3, our graphical model can be viewed as the composition of three layers. The bottom layer I contains all the input frames $\mathcal{I} = \{I_f\}_1^F$. The middle layer represents the estimated poses for each frame $\mathcal{H} = \{h_f\}_1^F$. Finally in the top layer, the fashion parsing results for all frames are denoted as $\mathcal{Y} = \{y_f\}_1^F$. For simplicity, only three temporal adjacent frames I_{f-1}, I_f and I_{f+1} are shown. Note that the nodes in the middle layer h_f are conditioned on the input observations I_f and the temporal constraints $w_{f,f+1}$ and $w_{f-1,f}$ from the adjacent human poses. The inference of all nodes is denoted as the video pose co-estimation. Furthermore, the video co-parsing task can be converted to inferring the states of nodes y_f in the top layer. The probability of each node y_f relies on the prediction of the corresponding pose h_f, the appearance constraints $m_{f,f+1}$ and $m_{f-1,f}$ as well as the inputs I_f. Because there are numerous hypotheses of locations of poses and all the frames are required to be parsed simultaneously, the joint inference of $P(\mathcal{H}|\mathcal{I})P(\mathcal{Y}|\mathcal{I}, \mathcal{H})$ can be NP-hard and impossible to solve efficiently. We approximate the inference task in Eq. (1) by separately optimizing the two sequential tasks: pose co-estimation $P(\mathcal{H}|\mathcal{I})$ and video co-parsing $P(\mathcal{Y}|\mathcal{I}, \mathcal{H})$. Note that the results of video pose co-estimation are the inputs of video co-parsing.

4.1.1 *Video Human Pose Co-estimation* $P(\mathcal{H}|\mathcal{I})$

Our pose co-estimation stage has two steps. First, we estimate the initial pose for each frame, illustrated as green triangles in Fig. 3. Second, all poses of all the frames are refined together by considering the confidence ranking of poses and the Sift-Flow correspondences between the successive frames, represented by purple diamonds in Fig. 3.

Image Human Pose Estimation $P(h_f|I_f)$ **:** Human pose estimation in the image has been extensively studied. We adopt the articulated pose estimation technique with flexible mixtures-of-parts method [18]. The human pose model can be represented by a K-node skeleton graph $G_s = (V_s; E_s)$, where the K nodes V_s correspond to different human parts, such as left shoulder, right shoulder, etc., and the edges E_s represent the relationships of human parts.

Given a frame I_f, we estimate the locations $\{l_f^i\}_{i=1}^K$ for all K key-points and the associated part types $\{t_f^i\}_{i=1}^K$ for each point within the human skeleton. The human pose can be calculated as $h_f = \{l_f, t_f\}$, where $l_f = \{l_f^i\}_{i=1}^K$ and $t_f = \{t_f^i\}_{i=1}^K$. We denote the hypotheses set of l_f^i as $\{1, \cdots, L\}$ and that of t_f^i as $\{1, \cdots, T\}$, where L is the image lattice and T is the number of types for each part.

Given a pose configuration h_f (including part types t_f and positions l_f), the confidence $P(h_f|I_f)$ is computed by combining 3 factors: the corresponding confidence for the part type assignments t_f, the unary score for each key point and the pairwise scores for the skeleton relations by [18]. It is worth noting that the probability $P(h_f|I_f)$ can be used to roughly predict the accuracy of the human pose estimator. That is, the high probability means the estimator has strong confidence for the estimated pose. We rank the probabilities of all the poses for all frames in each video, and then we can select the most confident poses, used as the "seeds" for the following pose co-estimation.

Video Human Pose Co-Estimation $P(\mathcal{H}|\mathcal{I})$**:** As aforementioned, even the state-of-the-art pose estimators may fail when parsing the fashion images. To process the numerous frames, we consider the video frames as a chain structure and the contextual relationships between adjacent frames are used to regularize the poses of all the frames. In this chain model, each node is the pose h_f of the frame $\mathcal{I}$, and the edges E_W are the chains. As shown in Fig. 3, the frame

Figure 4: Illustration of human pose co-estimation. The frames ♯4, ♯7, ♯23, ♯24 of the video are shown. The constrained human pose estimator is applied and the confidences are shown in parentheses. Then the frames with high confidences (e.g., ♯7 with confidence 0.9367) are regarded as seeds. Each frame has 3 candidate poses. The first candidate is generated by the frame itself, while the other two are transferred from seeds by Sift-Flow correspondences (in second row). For each possible candidate pair, we calculate the pairwise term according to the Sift Flow field. Based on the context between successive frames, the optimal poses among all pose candidates are selected, which are highlighted by red bounding boxes. The final pose co-estimation results are shown in the last row, which are much better than the original estimated poses, such as the initial pose estimation results for frames ♯23 and ♯24, highlighted by green bounding boxes.

h_f connects only with h_{f-1} and h_{f+1}. By using the temporal constraints, more accurate human pose estimations for all frames can be obtained simultaneously.

We use the Sift Flow [9] method to capture the temporal displacements between successive frames. For the frame pair (i, j), we denote the corresponding flow field as $w_{i,j}$, which is a 2D flow vector indexed by pixel positions. Given the flow field $w_{i,j}$, the position of each pixel p in the frame i can be mapped to $p + w_{i,j}(p)$ in the frame j. Note that w is not symmetric, i.e., $w_{i,j} \neq w_{j,i}$, according to the Sift Flow computational framework.

As for human pose co-estimation, we consider two items for jointly refining the poses of all the frames: single pose confidence for each frame and the pairwise pose coherence. First, the single pose confidence is obtained by $P(h_f|I_f)$, which evaluates the quality of the estimated pose of each frame. Second, the pairwise term assesses the coherence of poses in two adjacent frames. We map the estimated pose of one frame by the flow vector to its adjacent frame, and hope the mapped pose to be close to the estimated pose of the adjacent frame. This means that pose estimation results should be consistent with the temporal flow field. We thus formulate the human pose co-estimation as maximizing the probability $P(\mathcal{H}|\mathcal{I})$,

$$
\begin{aligned}
P(\mathcal{H}|\mathcal{I}) &\propto \prod_{f\in I} \mathcal{P}(l_f, t_f|I_f) \\
&\cdot \exp(-\eta \sum_{\{f_1,f_2\}\in E_W} (\sum_{i\in V_s} \left| l^i_{f_2} - \{l^i_{f_1} + w_{f_1,f_2}(l^i_{f_1})\}\right|^2_2 \\
&+ \sum_{i\in V_s} \left| l^i_{f_1} - \{l^i_{f_2} + w_{f_2,f_1}(l^i_{f_2})\}\right|^2_2))
\end{aligned} \tag{2}
$$

where η is used to balance the single pose confidence and the pairwise pose coherence, which is empirically set in our experiments. Given the pose l_{f_1} of the frame I_{f_1}, we map it into the frame I_{f_2} by using the Sift Flow vector w_{f_1,f_2} for all locations of l_{f_1}, denoted as $l^i_{f_1} + w_{f_1,f_2}(l^i_{f_1})$. The temporal displacement between the estimated pose l_{f_1} and the transfered pose $l^i_{f_1} + w_{f_1,f_2}(l^i_{f_1})$ is calculated using the Euclidean distance. The pairwise term is computed by the summation of the displacements of all key points.

The difficulty of optimizing Eq. (2) lies in two aspects. 1) We estimate the pose locations $\{l_f\}_{i=1}^K$ of all frames simultaneously, which leads to a very huge hypotheses set of the size FK^L. 2) The whole graph for the pose co-estimation can be viewed as a hierarchical model. The bottom is a common skeleton graph, of which the nodes are human key points and the edges are skeleton relations within each frame. Then the top is a chain structure, where the nodes are the single pose confidences obtained from the bottom and the edges E_W are the cross-frame pose coherences. This hierarchical graph makes inferring pose locations intractable for each video, not to mention our large-scale video set.

For efficiency, we consider all within-frame nodes (i.e. key points) for each frame as a super node (i.e. an integrated pose candidate). In this way, our graph can be simplified into a chain structure from a hierarchical model, which can be effectively solved by the well-known belief propagation method [1]. To generate a set of reasonable pose candidates for each frame, we use the pose propagation strategy with the selected pose seeds. Specifically, we rank all confidences $\{P(h_f|I_f)\}_{f=1}^F$ of initialized poses for all frames and select the top 5 candidates with the highest confidences as the pose seeds. We then propagate these seeds to all other frames via Sift Flow [9]. Except the frames with pose seeds, each frame I_f has 6 candidate poses, including 5 propagated pose candidates and the estimated pose from I_f itself. We consider these pose candidates as the possible hypotheses of each frame, which largely reduces the searching space for each node. During the inference procedure, the unary term for each super node I_f is $P(h_f|I_f)$ and the probabilities of propagated pose candidates are directly transferred from the original pose confidence of seeds. In addition, given a specific pair of frames, we obtain different pairwise terms if we select different pose candidate pairs. The pairwise term for each pose candidate pair is calculated by the summation of two temporal displacements using the Sift Flow vector, as described in Eq. (2). The whole procedure of our pose co-estimation is illustrated in Fig. 4. Two pose seeds with highest confidences are selected and then used to generate the candidate poses for the non-seeds frames.

4.1.2 *Video Co-Parsing* $P(\mathcal{Y}|\mathcal{I},\mathcal{H})$

Given the refined human poses for all frames, we can perform the video co-parsing, conditioned on the image and

[1] http://www.di.ens.fr/∼mschmidt/Software/UGM.html

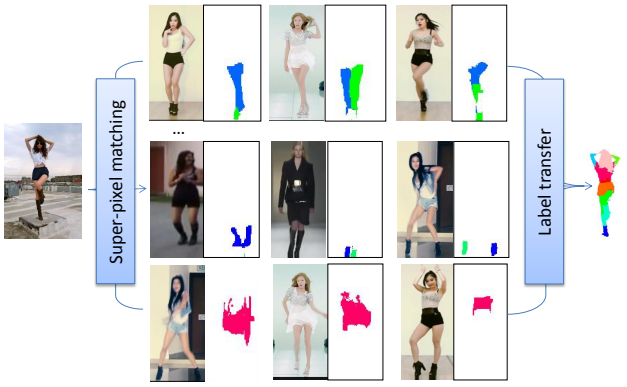

Figure 5: Non-parametric fashion parsing. For each testing image, 25 similar images in the video corpus are retrieved. Parsing results are then transferred from the retrieved video frames to the testing image at the super-pixel level. We show the transfer process of 5 fashion items, i.e., leg/right leg, left/right shoes and upper-clothing.

pose layer as displayed in Fig. 3. Our co-parsing algorithm includes two steps: computing pixel-level confidences w.r.t. the fashion items $P(y_{f,i'}|I_f, h_f)$ for all pixel $i' \in I_f$ (denoted as blue rectangles), and then co-parsing all frames by considering the super-pixel correspondences (denoted as red circles) to obtain $P(\mathcal{Y}|\mathcal{I}, \mathcal{H})$.

Image Parsing $P(y_f|I_f, h_f)$**:** Given one frame I_f and the refined human pose h_f, we compute the confidence score of assigning the possible clothing item label to each pixel. Let us denote $y_{f,i'}$ as the clothing item label at pixel i'. The confidence score $P(y_{f,i'}|I_f, h_f)$ of assigning clothing item labels to $y_{f,i'}$ can be computed by the existing fashion parser, e.g. [16]. And $P(y_f|I_f, h_f)$ can be denoted as the set of $P(y_{f,i'}|I_f, h_f)$.

Note that our algorithm can easily adapt to any other fashion parser, such as [3], by properly redesigning video co-parsing solution.

Video Co-parsing $P(\mathcal{Y}|\mathcal{I}, \mathcal{H})$**:** Based on the pixel-level confidences, we refine the parsing results of all frames together by considering the within-frame and cross-frame super-pixel consistencies. Intuitively, the super-pixels in the spatial neighbours within each frame are encouraged to take the same fashion labels; and similarly, the matched super-pixels across the adjacent frames also favor the same labels. We can thus rectify and smooth the label map of super-pixels of all frames together.

Following the previous parsing works, we build dense appearance correspondences for super-pixels instead of pixels. We first compute over-segmentations of all frames using a fast segmentation method [4]. Then the confidence score of assigning the clothing item label to each super-pixel s is computed as the average of the pixel-wise confidences $P(y_{f,i'}|I_f, h_f)$ of $i' \in s$ which represents all pixels within this super-pixel. We denote the confidence score of each super-pixel by $\phi_s(y_f(s))$. To smooth the label maps of all frames, we utilize two kinds of relationships to consider the appearance consistency. First, the within-image relationship N_{int} is computed for the spatial neighbors of super-pixels. Second, we consider the cross-image relationship N_{ext} for each super-pixel with its most similar counterpart in the previous/subsequent frame.

Mathematically, our co-parsing task aims to maximize the probability $P(\mathcal{Y}|\mathcal{I}, \mathcal{H})$, formulated as

$$P(\mathcal{Y}|\mathcal{I}, \mathcal{H}) \propto \exp(-(\sum_{f\in I}\sum_{s\in\wedge_f}\phi_s(y_f(s))+ \sum_{(s,q)\in N_{\text{int}}}\varphi_{\text{int}}(y_f(s), y_f(q)|I_f) + \sum_{(s,q)\in N_{ext}}\varphi_{ext}(y_f(s), y_{f'}(q)|I_f, I_{f'}))) \tag{3}$$

where $\wedge_f$ denotes all super-pixels within each image I_f. (s, q) represents each pair of neighboured super-pixels within images and across images. The within-image smoothness term φ_{int} and the cross-image smoothness term φ_{ext} are defined as

$$\begin{aligned}&\varphi_{\text{ext}}(y_f(s), y_f(q)|I_f)\\ &= \delta(y_f(s) \neq y_f(q)) \exp(-\lambda_{ext}|\mathcal{F}(s) - \mathcal{F}(q)|)\\ &\varphi_{\text{int}}(y_f(s), y_f(q)|I_f, I_{f'})\\ &= \delta(y_f(s) \neq y_{f'}(q)) \exp(-\lambda_{int}|\mathcal{F}(s) - \mathcal{F}(q)|)\end{aligned} \tag{4}$$

where $\delta(\cdot)$ is the indicator function and $\mathcal{F}(s)$ is the feature of the super-pixel, which is computed by a concatenation of bag-of-words from RGB, Lab and Gradient for each super-pixel. We also pick the closest super-pixel pairs (s, q) across the sequent frames using L2-distance on these bag-of-words features. λ_{ext} and λ_{int} are the weights of two kinds of pairwise terms. Because our pairwise term Eq. (4) is a submodular function, the optimization of maximizing Eq. (3) becomes a tractable graphical model. We solve this optimization problem by the well-studied α-expansion method [5]. Thus the optimal parsing results of all frames can be calculated as $\mathcal{Y}$.

4.2 Non-parametric Fashion Parsing

Based on our contextual video parsing algorithm, we can efficiently process the large scale video data to generate the gallery set of images. In the following, we propose a non-parametric method for transferring the parsing results of our parsed gallery to the test image.

Given a testing image I, we first use the human detection technique [6] to roughly locate the human body. The caffe feature [8] for each human is computed, which can intrinsically capture the style, pose and appearance characteristics of the whole image. We use L2-distance over the DeCAF feature to find 25 nearest neighbors in our gallery. After that, the testing images are over-segmented and each super-pixel of the testing image finds the closest super-pixel from each retrieved image using L2-distance of the caffe features.

More specifically, we denote the retrieved images for the image I as D. For each super-pixel s, the selected corresponding super-pixel from the reference image r in D is denoted as s_r, and the caffe feature of the super-pixel s is denoted as $h(s)$. Then, our transfered label y_s for each super-pixel s is computed by

$$P(y_s|s, D) = \frac{1}{Z}\sum_{r\in D}\frac{M(y_s, s_r)}{1 + \|h(s) - h(s_r)\|} \tag{5}$$

where we define:

$$M(y_s, s_r) = \frac{1}{|s_r|}\sum_{i'\in s_r}\delta(y_{r,i'} = y_s), \tag{6}$$

where $|s_r|$ is the number of pixels within the super-pixel s_r and i' denotes each pixel. Z is a normalization constant. Our parsing results are computed as the weighted average of the parsing results of closest super-pixels for all retrieved images in D. The obtained transferred parsing results $P(y_s|s, D)$ for all super-pixels are further refined by

Figure 6: The top row shows images from FS and CFPD while the bottom row shows images from FI.

Markov Random Filed to respect boundaries of actual clothing items.

5. EXPERIMENTS

We first introduce our experimental setting, including the datasets and the baselines we compare with. Then we report the step-by-step results of the whole framework, including the quantitative and qualitative results of video pose co-estimation, video co-parsing and test image parsing.

5.1 Experimental Setting

We conduct the experiments on three datasets. The first is the Fashionista (FS) dataset [17] containing 685 photos with good visibility of the full body and covering a variety of clothing items. 456 out of the 685 images are used for training and the rest 229 images are used for testing. The second dataset is the Colorful Fashion Parsing Data (CFPD) [10] dataset which consists of 2,682 images. The training set and the testing set are split into half-half. The third dataset is our newly collected Fashion Icon (FI) dataset which contains 1,028 images. The images in this dataset contain one or multiple humans with quite diverse human poses. Some exemplar images of the three datasets are shown in Fig. 6. It can be seen that humans in constrained fashion images are in (near)-frontal view and well-posed. However, in unconstrained fashion images, the girls cross or stretch their arms or legs freely, and may be in arbitrary view. In our experiments, the label sets of FS and CFPD contain 18 and 13 kinds of fashion items, respectively. FI has two sets of label sets, one containing 18 kinds of fashion items as FS, and the other containing 13 kinds as CFPD, where the later is obtained by merging from the former. Our experiments are conducted on a PC with Core I7 3.4GHz GPU and 6GB memory, and the average processing time for testing an image with resolution $600 * 400$ is 2 seconds. The parameter η, λ_{ext} and λ_{int} are set as 0.1, 0.5 and 0.5 empirically in this work.

5.2 Experimental Results

In this subsection, we first evaluate the effectiveness of human pose co-estimation and video co-parsing sequentially. Then, we compare the results of our system and the baselines on the three datasets.

5.2.1 *Video Pose Co-estimation Evaluation*

We evaluate our human pose co-estimation method in predicting the poses of frames. We randomly select 100 videos from our collected video dataset and manually label 14 key points of the human skeleton for each selected video frame. We compare our results with the state-of-the-art image-based pose estimator, mixtures-of-parts model [18], which is trained on FS and predicts the pose of each frame separately. The standard PCK (Probability of Correct Key point) metric [18] is used to evaluate the performance of pose estimation. Table 1 displays the results of the frame based pose estimator (denoted as "Pose") and the video pose co-estimator (denoted as "Co-Pose"). The results demonstrate that our video pose co-estimator can generally improve the key points localization accuracies for 9 out of all 14 key points. In particular, the accuracy for the left knee point has been increased by 9.6%. Moreover, Our average PCK of all 14 key points reaches high accuracy of 78.17% and improves the "Pose" by 1.73%.

In addition, we also visualize the pose estimation results of the two comparison methods in Fig. 7. For the dancing video frames with large variations in pose and view, our method shows superior performance in predicting the key points of human poses, especially for the left and right knees. As shown in Fig. 7(a), only the left knee point of the first frame is predicted correctly for "Pose" while our method can rectify the left knee key points of all frames by benefiting from the Sift-Flow and temporal coherence constraints.

Table 3: Comparison between frame based parsing and video co-parsing

Method	Accuracy	F.g. accuracy	Avg. precision	Avg. recall	Avg. F-1
Parsing	80.48	44.79	31.98	40.64	33.18
Co-Parsing	**82.38**	**48.47**	**33.02**	**42.54**	**34.69**

5.2.2 *Video Co-Parsing Evaluation*

We compare the performances of our video co-parsing method with the existing image-based fashion parser [16], whose code is publicly available [2]. The 100 videos are randomly selected and all frames are manually labeled. Similar to [16], we evaluate the parsing results of all frames with 5 metrics, including accuracy, foreground accuracy, average precision, average recall and average F-1 score. The comparison results are shown in Table 3. Significant improvements of our co-parsing method for all 5 metrics can be observed.

More exemplar results are shown in Fig. 8. Video co-parsing predicts more consistent fashion labels for all video frames than the image-based fashion parser. For example, in the left panel, "Parsing" predicts that the girl wears upper clothing in three frames yet dress in two frames. Through the contextual inference of "Co-Parsing", all five frames are correctly predicted. Another example is shown in the right panel of Fig. 8. The left/right arms and left/right legs are more accurately estimated benefiting from the temporal regularizations during the co-parsing procedure.

5.2.3 *Evaluation of Fashion Parsing in FS and CFPD*

We report the fashion parsing performance of Paper Doll and our method on testing images in the FS and CFPD datasets. In addition, we evaluate the superiority of our pose co-estimator and video co-parsing components. The "Co-Pose+Co-Parsing" utilizes the pose co-estimator and the video co-parser sequentially. The "Co-Parsing" solution does not implement pose co-estimation and directly uses the image-based pose estimator.

The results are listed in the first two rows of Table 2. It is obvious that both of our two solutions achieve higher perfor-

[2]http://www.cs.sunysb.edu/~kyamagu/research/paperdoll/

Table 1: The PCK comparison between frame based pose estimation and video based pose co-estimation.

key point	lank	lkne	lhip	rhip	rkne	rank	lwr	lelb	lsho	rsho	relb	rwr	hbot	htop
Pose	68.88	74.13	83.39	83.92	79.37	**71.50**	43.88	**73.43**	**93.01**	90.04	**71.50**	**47.38**	94.93	94.76
Co-Pose	**74.48**	**83.73**	**87.94**	**91.78**	**80.94**	70.80	**44.58**	72.55	92.65	**91.78**	69.06	38.63	**98.08**	**97.38**

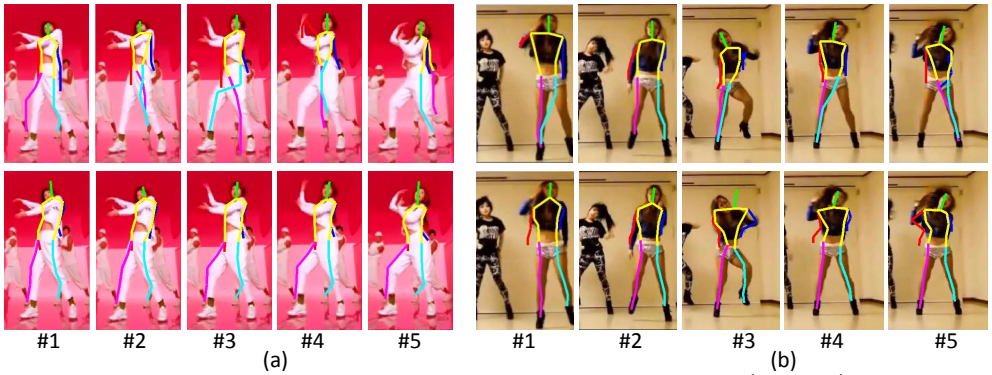

Figure 7: Two comparison examples between image based human pose estimation (top row) and video based human pose co-estimation (bottom row).

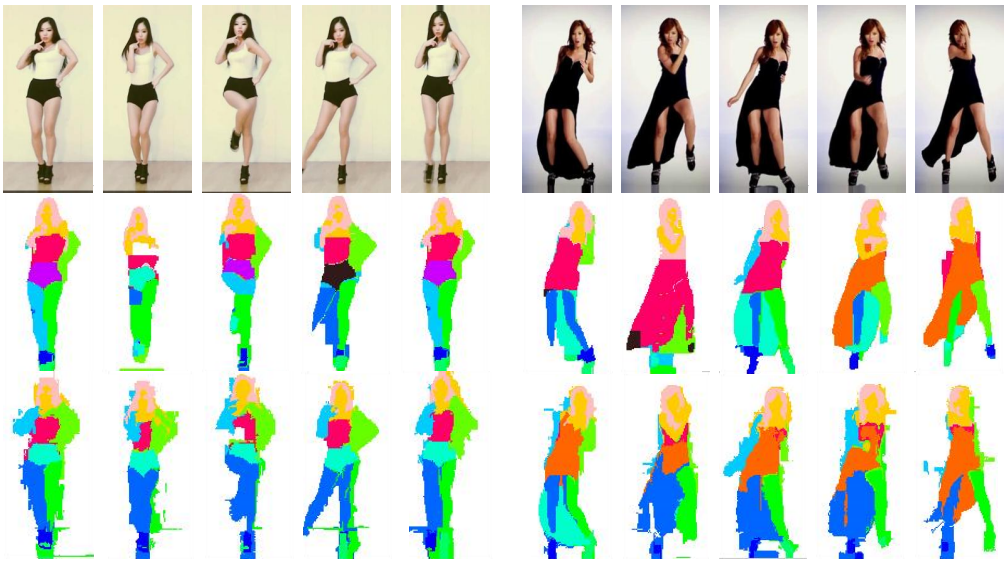

Figure 8: Comparison examples between image based fashion parsing (middle row) and video based co-parsing (bottom row).

Table 2: Comparison among Paper Doll [16] and two solutions of our method in FS, CFPD and FI.

Data set	Method	Accuracy	F.g. accuracy	Avg. precision	Avg. recall	Avg. F-1
FS-FS	Paper Doll	85.69	52.09	41.74	**45.15**	37.43
	Co-Parsing	87.21	54.03	54.73	39.36	39.15
	Co-Pose + Co-Parsing	**88.34**	**57.08**	**56.97**	42.25	**43.69**
CFPD-CFPD	Paper Doll	82.79	44.08	**49.20**	32.00	32.66
	Co-Parsing	83.73	49.03	43.56	40.36	39.96
	Co-Pose + Co-Parsing	**84.70**	**52.49**	42.31	**42.31**	**41.42**
FS-FI	Paper Doll	84.63	47.43	36.12	**39.65**	35.20
	Co-Parsing	86.26	42.09	35.96	29.30	28.31
	Co-Pose + Co-Parsing	**87.33**	**51.09**	**41.63**	39.33	**37.07**
CFPD-FI	Paper Doll	81.81	37.11	34.20	28.04	25.20
	Co-Parsing	83.84	45.77	35.14	36.04	34.00
	Co-Pose + Co-Parsing	**85.65**	**50.29**	**37.13**	**38.05**	**36.05**

mances than the Paper Doll in general, which demonstrates the capability of our contextual video co-parser. In addition, the necessity of human pose co-estimation is proven, where the avg. F1-score of "Co-Pose+Co-Parsing" outperform "Co-Parsing" by 4.54%.

We also present the F1-scores for each fashion item label in Fig. 9(a) and Fig. 9(b). Generally, the "Co-Pose+Co-Parsing" shows the highest performance, especially in predicting the human part labels, such as "LeftLeg", "LeftArm", "RightLeg" and "RightArm". We can also observe this superior performance from the visualization of parsing results in Fig. 11(a) and Fig. 12(a). The first row shows the parsing results of Paper Doll and the second row shows ours. Our parser performs better on predicting the fashion items, such as skirt, pants, and upper-clothes. In addition, our results can be less disturbed by the background clutter and show relatively clear boundary and appearance consistency, e.g. the leg regions of the second and third images in the first

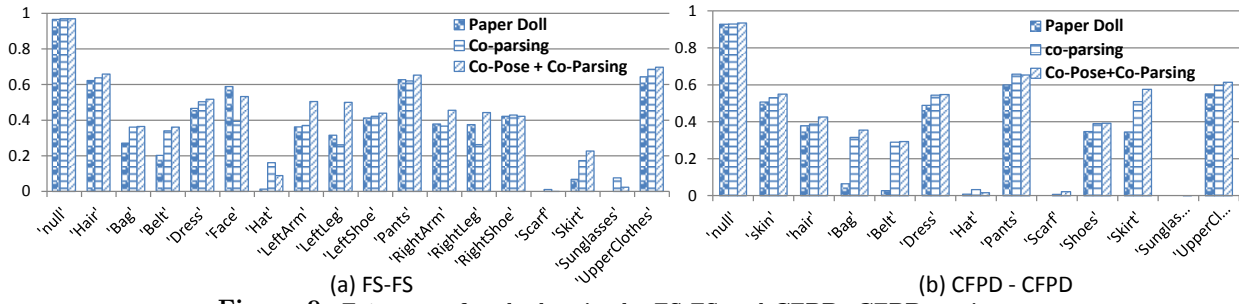

Figure 9: F-1 score of each class in the FS-FS and CFPD -CFPD settings.

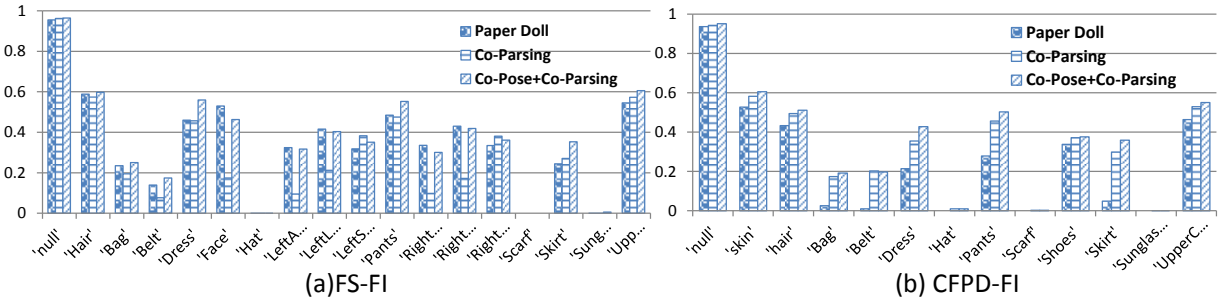

Figure 10: F-1 scores of each class in the FS-FI and CFPD-FI settings.

Figure 11: Comparison of two settings: (a) FS-FS, (b) FS-FI. In each triple, the original image, the parsing result by Paper Doll and our result are shown sequentially.

Figure 12: Comparison of two settings: (a) CFPD-CFPD, (b)CFPD-FI. In each triple, the original image, the parsing result by Paper Doll and our result are shown sequentially.

row of Fig. 11(a). Moreover, our parser can also correctly localize small fashion items, such as the bags in the second and third images in the second row of Fig. 11(a) and the third image in the first row of Fig. 11(b).

5.2.4 *Evaluation of Fashion Parsing in FI*

Our collected FI dataset contains more images with diverse poses and arbitrary views. We parse images in the FI dataset with the trained model from two training image sets, FS and CFPD, separately. The main difference between $FS-FI$ and $CFPD-FI$ is that we train two super-

Figure 13: The results of our system in parsing images with multiple humans in the FS-FI setting.

vised parsing models with different training data and label sets. Similarly, we compare two solutions of our method, i.e., "Co-Pose+Co-Parsing" and "Co-Parsing" with the baseline Paper Doll. The quantitative comparison results show that our method largely improves the performance of Paper Doll in both settings, shown in the last two rows of Table 2. It is worth noting that our method shows larger improvements on the FI dataset than on FS and CFPD datasets. Specifically, with the same training dataset CFPD, the performance of FI is increased by 3.82% compared with 1.91% of CFPD in terms of accuracy. This well proves the advantages of our method on parsing fashion images.

The detailed comparison of each fashion item label among "Paper Doll", "Co-Parsing" and "Co-Pose+ Co-Parsing" in both FS-FI and CFPD-FI settings is illustrated in Fig. 10. In general, "Co-Pose+ Co-Parsing" outperforms "Co-Parsing" and performs much better than Paper Doll.

Moreover, the visual parsing comparisons are shown in Fig. 11(b) and Fig. 12(b) for the FS and CFPD datasets, respectively. Our system can correctly predict the fashion items for images with very diverse human poses, e.g., the second image of the first row of Fig. 11(a).

Additionally, we also conduct experiments of parsing the multi-human images under the FS-FI setting, as shown in Figure. 13. We use the detection method [6] to cut the images into several smaller images with single human only. The single human image is then fed into our system and the parsing results are generated. Then the final parsing result for each multi-human image is merged by combining the parsing of each single image. We show several results of parsing images with multiple humans and prove that our method can predict reliable parsing results when the humans are not heavily occluded.

6. CONCLUSION AND FUTURE WORK

In this paper we propose a novel framework for fashion parsing in the wild which leverages video contexts without extra annotation. It contains two components, i.e., the contextual video parsing and the non-parametric fashion parsing. Extensive experiments on two benchmark fashion datasets as well as a newly collected FI dataset demonstrate the effectiveness of our proposed framework well. We can optionally label more unconstrained images and train an unconstrained fashion parser. However, the great burden of human labelling may considerably limit the scalability of the fashion parser. Since our method only needs the unsupervised videos which can be easily crawled from the web, our solution can easily scale up to new videos with even more challenging poses and views, e.g. lying on the floor or sitting in the chair.

Two possible research directions can be considered in the future. First, we plan to develop a mobile App, which can parse images uploaded by users in an online way. Second, we can select images with uncertain results through active learning, and then annotate more frame poses and fashion parsing ground truths, which may lead to better video parsing results with a small amount of user interaction.

Acknowledgement

This work was supported by the Singapore National Research Foundation under its International Research Centre@Singapore Funding Initiative and administered by the IDM Programme Office. The work is also supported by the Hi-Tech Research and Development (863) Program of China (no. 2013AA013801) and Guangdong Science and Technology Program (no. 2012B031500006).

7. REFERENCES

[1] Vijay Badrinarayanan, Fabio Galasso, and Roberto Cipolla. Label propagation in video sequences. In *CVPR*, 2010.

[2] Huizhong Chen, Andrew Gallagher, and Bernd Girod. Describing clothing by semantic attributes. In *ECCV*. 2012.

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

[4] Pedro F Felzenszwalb and Daniel P Huttenlocher. Efficient graph-based image segmentation. *IJCV*, 2004.

[5] Brian Fulkerson, Andrea Vedaldi, and Stefano Soatto. Class segmentation and object localization with superpixel neighborhoods. In *CVPR*, 2009.

[6] Ross B Girshick, Pedro F Felzenszwalb, and David A McAllester. Object detection with grammar models. In *NIPS*, 2011.

[7] Sam Hare, Amir Saffari, and Philip HS Torr. Struck: Structured output tracking with kernels. In *ICCV*, 2011.

[8] Yangqing Jia, Evan Shelhamer, Jeff Donahue, Sergey Karayev, Jonathan Long, Ross Girshick, Sergio Guadarrama, and Trevor Darrell. Caffe: Convolutional architecture for fast feature embedding. 2014.

[9] Ce Liu, Jenny Yuen, and Antonio Torralba. Sift flow: Dense correspondence across scenes and its applications. *TPAMI*, 2011.

[10] Si Liu, Jiashi Feng, Csaba Domokos, Hui Xu, Junshi Huang, Zhenzhen Hu, and Shuicheng Yan. Fashion parsing with weak color-category labels. *TMM*, 2014.

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

[12] 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*, 2012.

[13] X Liu, D Tao, M Song, J Bu, and C Chen. Discriminative segment annotation in weakly labeled video. In *CVPR*, 2014.

[14] Jamie Shotton, Matthew Johnson, and Roberto Cipolla. Semantic texton forests for image categorization and segmentation. In *CVPR*, 2008.

[15] Meng Wang, Richang Hong, Xiao-Tong Yuan, Shuicheng Yan, and T-S Chua. Movie2comics: Towards a lively video content presentation. *TMM*, 2012.

[16] Kota Yamaguchi, M Hadi Kiapour, and Tamara L Berg. Paper doll parsing: Retrieving similar styles to parse clothing items. In *ICCV*, 2013.

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

[18] Yi Yang and Deva Ramanan. Articulated pose estimation with flexible mixtures-of-parts. In *CVPR*, 2011.

[19] T Zhang, B Ghanem, S Liu, and N Ahuja. Robust visual tracking via multi-task sparse learning. 2012.