Fashion Parsing With Video Context

Si Liu, *Member, IEEE*, Xiaodan Liang, Luoqi Liu, Ke Lu, Liang Lin, Xiaochun Cao, *Member, IEEE*, and Shuicheng Yan, *Senior Member, IEEE*

***Abstract*—In this paper, we propose a novel semi-supervised learning strategy to address human parsing. Existing human parsing datasets are relatively small due to the required tedious human labeling. We present a general, affordable and scalable solution, which harnesses the rich contexts in those easily available web videos to boost any existing human parser. First, we crawl a large number of unlabeled videos from the web. Then for each video, the cross-frame contexts are utilized for human pose co-estimation, and then video co-parsing to obtain satisfactory human parsing results for all frames. More specifically, SIFT flow and super-pixel matching are used to build correspondences across different frames, and these correspondences then contextualize the pose estimation and human parsing in individual frames. Finally these parsed video frames are used as the reference corpus for the non-parametric human parsing component of the whole solution. To further improve the accuracy of video co-parsing, we propose an active learning method to incorporate human guidance, where the labelers are required to assess the accuracies of the pose estimation results of certain selected video frames. Then we take reliable frames as the seed frames to guide the video pose co-estimation. Our human parsing framework can then easily incorporate the human feedback to train a better fashion parser. Extensive experiments on two benchmark fashion datasets as well as a newly collected challenging Fashion Icon dataset well demonstrate the encouraging performance gain from our general pipeline for human parsing.**

***Index Terms*—Information retrieval, professional communication.**

I. Introduction

HUMAN parsing aims to predict the label (e.g. face, bag, left-arm, etc.) for each pixel in a human photo. Human parsing can benefit a wide range of real applications. For example, human parsing can be used in intelligent surveillance, such as person re-identification. A good understanding of one's apparel may provide useful cues to identify a person. Moreover, reliable human parsing results can facilitate the clothing retrieval by predicting which pixels belong to the each label [40]. The clothing retrieval can thus utilize the specific feature for each label instead of using the feature extracted from the whole image.

Fig. 1. Examples of human parsing results. Note that the girls are with quite diverse poses, and some girls are in side-view or back-view, which are very challenging. For better viewing of all figures in this paper, please see original zoomed-in color pdf file.

Despite the great progresses achieved [40], [7], [39], [22], human parsing has not been fully solved. It is very challenging due to the following reasons. Firstly, the same label has very diverse appearances. For example, the hair styles of the girls in Fig. 1 are different. Secondly, the relationships among all labels are complicated yet critical for human parsing. Thirdly, the clothing items are always occlued by the human or other clothing items. Finally, all current human parsing datasets are very small.

The explosive development of social networks and image/video sharing websites provides easy access to inexhaustible fashion images and videos. This inspires us to explore whether we can train a robust human parser by making use of limited labeled data and inexhaustible unsupervised web images or videos. To improve the quality of the enriched parsing samples and avoid semantic drifting, we argue that the obtained web videos (from, e.g., YouTube.com) are better choices than web images, since the videos contain richer contextual information. The video context can improve both the intermediate result, i.e., human pose estimation, and the final goal, i.e., human parsing. Video contexts include the temporal correspondence and semantic consistency. Based on the temporal and semantic contexts, reliable pixel-wise cross-frame correspondences can be constructed. Then the correct human pose estimation and parsing results of single frames can be transferred to difficult frames. Thus the wrongly estimated poses or incorrectly parsing results can be refined based on the neighboring parsing results and the cross-frame correspondences. To sum up, parsing the whole video collaboratively can filter the possible noises of parsing each frame individually. The acceptable quality of the enriched data makes re-training a better human parser possible.

Manuscript received August 12, 2014; revised February 17, 2015; accepted May 22, 2015. Date of publication June 10, 2015; date of current version July 15, 2015. This work was supported in part by the National Natural Science Foundation of China under Grant 61422213, Grant 61332012, and Grant 61328205, the 100 Talents Programme of The Chinese Academy of Sciences, the Hi-Tech Research and Development Program of China under Grant 2013AA013801, the Guangdong Natural Science Foundation under Grant S2013050014548, and the Program of Guangzhou Zhujiang Star of Science and Technology under Grant 2013J2200067. The associate editor coordinating the review of this manuscript and approving it for publication was Prof. K. Selcuk Candan.

Si Liu is with the State Key Laboratory Of Information Security, Institute of Information Engineering, Chinese Academy of Sciences, Beijing 100093, China, and also with the National University of Singapore, Singapore 100093.

X. Liang and L. Lin are with Sun Yat-Sen University, Guangzhou 510006, China.

L. Liu and S. Yan are with the Department of Electrical and Computer Engineering, National University of Singapore, Singapore 119077.

K. Lu is with the Graduate University, Chinese Academy of Sciences, Beijing 100049, China.

X. Cao is with the State Key Laboratory Of Information Security, Institute of Information Engineering, Chinese Academy of Sciences, Beijing 100093, China.

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TMM.2015.2443559

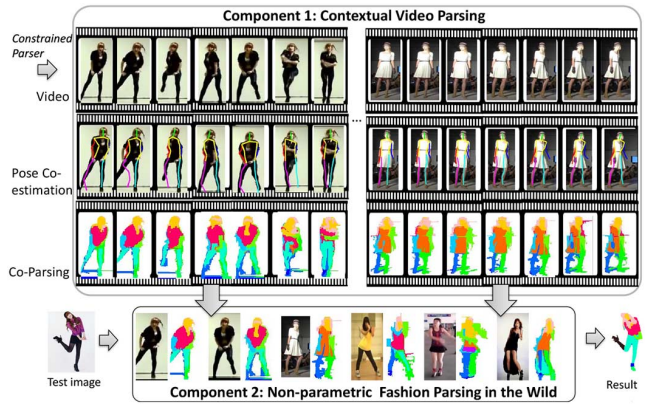


Fig. 2. Overview of our framework. The proposed framework contains two components: contextual video parsing and non-parametric human parsing. Any off-the-shelf pose estimator and human parser can initialize human pose estimaiton and parsing results. By leveraging the video contexts represented by SIFT Flow and super-pixel matching, video human pose co-estimation and human co-parsing algorithms are proposed to achieve much better pose estimation and human parsing results in the video. To further improve the accuracy of video parsing, an active learning method is proposed. The users judge the correctness of the pose estimation results of several recommended video key frames, which then serve as seeds to guide video pose co-estimation. Then, a large number of video frames and their video parsing results can be obtained and used as a gallery set to facilitate the nonparametric label transferring to the testing image.

The framework is illustrated in Fig. 2. It contains two components: 1) contextual video parsing, and 2) non-parametric image human parsing. For the contextual video parsing component, the goal is to parse the unlabeled videos crawled from the web. Since human pose estimation is as a prerequisite for human parsing, a human pose co-estimation step (Section V) is implemented first. In this step, we first apply the off-the-shelf human pose estimator to the videos, and then refine the estimated results by incorporating the pixel-level correspondences between sequential frames, which is described by Sift Flow [20]. Furthermore, for better video pose co-estimation, we propose an active learning method (Section VI) to recommend a human pose set. The pose set should satisfy two conditions: 1) the poses are correctly estimated; 2) the poses are not redundant. Users are required to assess the predicted poses of the selected frames to generate pose seeds which are then used to estimate the remaining poses of the video. Feedback from the users can thus be easily integrated into the video pose co-estimation framework. After human pose co-estimation, we first apply an existing human parser (pre-trained on a small amount of labeled data). It can provide a rough initialization for the video parsing. Then we use the correspondences between super-pixels of two sequential frames, which are described by super-pixel matching technique to refine the parsing results. That is, we co-parse all the frames in one video simultaneously (Section VII). In the non-parametric human parsing component, these parsed video frames are used as the gallery set which transfers labels to the testing images (Section VIII).

The contributions of this work can be summarized as follows.

- We propose a novel semi-supervised framework which can train a reliable parsing model with limited labeled training data by exploring the easily available fashion videos. Our framework is general, affordable and scalable.
- To avoid the semantic drifting, we propose a human pose co-estimation and co-parsing technique to achieve more reliable video parsing.
- We propose an active leaning method to automatically recommend several human poses to the users to assess pose estimation quality. The feedback from the users can be easily integrated into our human parsing framework.

Some components of the paper have been published in an early conference version [23]. This journal version add a few components. The key differences betweeen the conference and journal version are listed as following. The conference version only proposes how to train a human parser in a semi-supervised manner. However, in this paper, we further consider the human interaction and propose an active learning method to automatically and adaptively recommend video frames to labelers to check. Then the users feedback can be easily integrated into our semi-supervised framework to train a better human parser. Moreover, We further study how human parsing can assist clothing retrieval. Finally, more experimental results are added.

II. Related Work

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

A. Human Parsing

The human parsing problem is a special kind of semantic segmentation, which has been studied for a long time [32], [4], [40], [35], [11]. Yamaguchi *et al.* [40] proposed to perform human pose estimation and attribute labeling sequentially. Their human parsing performance was not quite high due to the large human pose variation and background clutters. Later, Yamaguchi *et al.* [39] dramatically improved the human 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 from retrieved examples. Liu *et al.* [22] addressed the problem of automatically parsing the human photos 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. Kohli *et al.* [17] proposed an approach for joint pose estimation and human segmentation. The hierarchical compositions based on the segment shapes were also utilized to assemble the candidate parts [3]. Tran *et al.* [35] proved the advantages of representing a full set of relations between segments than the standard tree model for human parsing. However, this method has high computational cost. Although these works have made great progress in human parsing, the involved representative models usually require a lot of prior knowledges about the specific tasks and heavily rely on the over-segmentation and pose estimation.

Recently, with the development of deep learning structures, many researchers explore how to apply the deep model, especially deep Convolutional Neural Network (CNN), to the semantic segmentation. Farabet *et al.* [8] trained a multi-scale convolutional network from raw pixels to extract dense features for assigning the label to each pixel. The recurrent convolutional neural network [28] was used to speed up the scene parsing and

the state-of-art performances were achieved for scene parsing. Girshick *et al.* [12] also proposed to classify the candidate regions by CNN for semantic segmentation. These deep learning based methods are limited due to the lack of enough training data to capture the diverse appearances and poses of the fashion items and human body.

B. Video Parsing

Similarly, comparing with parsing each video frame separately, more reliable parsing results can be achieved by co-parsing all the frames collaboratively. The cross-frame contexts are used as regularization to smooth and refine the results produced by parsing each frame. Previous efforts on image co-segmentation [16], [2] also utilize a similar spirit.

The main difficulty of video parsing lies in the great burden of labeling training samples. Vijayanarasimhan *et al.* [36] proposed an active learning based solution which can select k frames for manual labeling, such that automatic pixel-level label propagation can proceed with minimal expected error. In this paper, we mainly produce a scalable semi-supervised video parsing framework, where only a small-scale labeled training samples and a larges-scale unlabeled web videos are used. We also propose an active learning based framework where users can label very few video frames, and better results can be achieved.

III. Dataset Collection

We collect two datasets, including a video dataset and a Fashion Icon (FI) image dataset. The video dataset contains 1,500 unlabeled videos downloaded from youtube.com. It is crawled to assist our ultimate goal of parsing human photos. The second FI dataset contains 1,082 images. Compared with existing human parsing datasets, e.g., the Fashionista (FS) dataset [40] and the Colorful human parsing Data (CFPD) [22] dataset, the FI dataset is more challenging, since each image may contain multiple people and each person may take more diverse poses.

Video Dataset: Firstly, we apply Grammar Models [13] to the first frame of the video and automatically detect all the human bodies. For the frames containing multiple people, we only keep the detection bounding box of the people with the largest size and ignore other detected human. Secondly, the detected human-centric bounding box is used as the seed for the tracking algorithm, i.e., Struck [15]. Thus all the video frames are roughly aligned and mostly occupied by the human body, which greatly facilitates the later video parsing. During the data collection process, other detction algorithms [12], [31] or tracking algorihtms [44], [43] can also be used. 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 tracking algorithm is a balance between accuracy and efficiency. Thanks 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 human parsing datasets [39], [7] in two aspects. Firstly, some images in FI may 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 labeled based on the label set defined by Dong *et al.* [7], which includes 18 categories: face, sunglass, hat, scarf, hair, upper clothes, left-arm, right-arm, belt, pants, left-leg, right-leg, skirt, left-shoe, right-shoe, bag, dress and background.

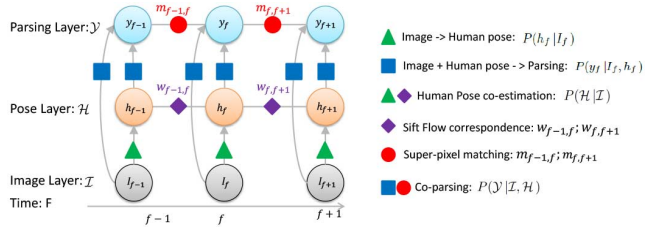


Fig. 3. Whole video parsing framework. The graphical model has three layers: image layer, pose layer, and parsing layer. Green triangles and blue squares respectively represent the traditional human pose estimation and human parsing. By incorporating SIFT Flow correspondences indicated by purple diamonds, 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.

IV. 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 human 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 and obtain better parsing results. 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 models. By taking these informative contexts as the regularization constraints, the pose co-estimation and human co-parsing can be largely improved.

We denote the parsing results of all frames $\mathcal{I}$ as $\mathcal{Y}$ and the human pose estimation results as $\mathcal{H}$. We estimate the pixel-wise semantic labeling, 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 interdependent for the human parsing task. Video co-parsing can be formulated as maximization of the conditional probability over parsing results $\mathcal{Y}$, human poses $\mathcal{H}$ and video frames $\mathcal{I}$, expressed by

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

As illustrated in Fig. 3, our graphical model composes 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 human 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. 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 for the locations of the 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 (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.

V. Video Pose Co-Estimation

Our pose co-estimation stage has two steps. First, we estimate the initial pose for each frame, illustrated by green triangles in Fig. 3. Second, the 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.

1) Image 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 the flexible mixtures-of-parts method [41]. 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 [41]. 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 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.

2) Video Pose Co-Estimation $P(\mathcal{H}|\mathcal{I})$: As aforementioned, even the state-of-the-art pose estimators may fail when parsing the human photos. To process the numerous video 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 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 [20] 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 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 $P(\mathcal{H}|\mathcal{I})$,

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

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 $l_{f_1}^i + w_{f_1,f_2}(l_{f_1}^i)$. The temporal displacement between the estimated pose l_{f_1} and the transfered pose $l_{f_1}^i + w_{f_1,f_2}(l_{f_1}^i)$ 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 (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, in 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

[1][Online]. Available: http://www.di.ens.fr/~mschmidt/Software/UGM.html

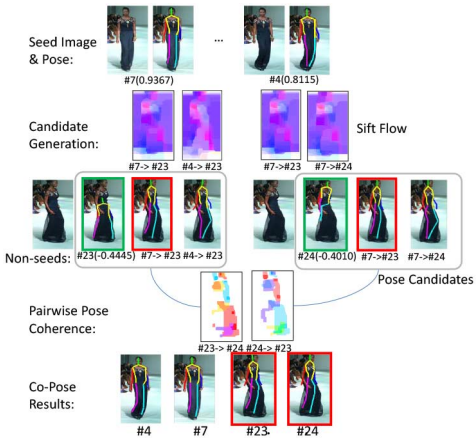


Fig. 4. Illustration of human pose co-estimation. The frames # 4, # 7, # 23, # 24 of the video are shown as examples. The constrained human pose estimator is applied and the confidences are shown in parentheses. Then the frames with high confidences (e.g., # 7 and # 4) are regarded as the seeds. In the active learning scenario, users manually check the recommended human poses, and only the poses considered as correct are used as pose 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 the second row). For each possible candidate pair, we calculate the pairwise correspondence according to the SIFT Flow field in the fourth row. Based on the context between successive frames, the optimal poses among all pose candidates are selected.

these seeds to all other frames via SIFT Flow [20]. 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 confidences of the 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 (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-seed frames.

VI. Human-Assisted Video Pose Co-Estimation

Note that video pose co-estimation is still far from perfectly solved even with the aforementioned co-estimation techniques in Section V-2. Actually, its performance is greatly determined by the accuracy of the pose seeds, because the seeds determine the hypothesis set for each frame in the propagation process.

As shown in Fig. 4, inaccurately estimated poses of the seed frames may spread to other non-seed frames and damage the video pose co-estimation results. Previously in Section V, we select the key poses based on only the pose estimation confidences produced by [41]. Generally, the videos contain extremely diverse human poses, which brings a great challenge for pose estimation algorithms. The poses with high confidences computed by these algorithms are not always correct, thus need to be checked manually. Next, we propose an active learning strategy to select human pose seeds. In order to reduce human labor, we only require labelers to judge whether the estimated poses are correct or not. For each video, an adaptive (e.g. 10 frames) with their human pose estimation results overlaid are shown to the labelers who need to check every human pose estimation results.

Our active video pose co-estimation procedure contains two steps. Step 1 aims to determine which frames to be shown to labelers and Step 2 aims to incorporate human interaction into the existing video pose co-estimation framework. Next, we will elaborate the two steps sequentially.

The first step automatically recommends video frames to the labelers to assess. Two criteria are considered during the pose recommendation. The first criterion is that the pose estimation confidences should be high, which is the same requirement as in Section V. Generally, similar poses will produce similar scores by [41]. However, if all the seed poses are visually similar, they are not representative enough to cover all the variations in the video and this will damage the co-estimation results. Thus we have the second criterion: the pose seeds should sufficiently diverse. Also, reducing the redundancy among the selected video frames saves the labelers much time on checking similar frames.

Mathematically, our target is to select a subset from $\mathcal{I} = \{I_f\}_1^F$. To satisfy the first criterion, we build a matrix A, where the diagonal elements represent the confidence of each frame I_f, i.e., $P(h_f|I_f)$, by the initialized pose estimator, and the non-diagonal elements measure the dissimilarities between samples. For example, $A_{i,j}$ is calculated by first calculating the Euclidean distances between the ℓ_1 normalized deep convolutional activation features by Caffe implementation[2] of frames f_a and f_b. The distance is then fed into the sigmoid function to be scaled into [0,1] and produce $A_{i,j}$. We choose the activations of the first fully connected layer as our feature (4096-d). According to the two criteria, we want to select a dense subgraph, all frames of which have high pose confidences and low redundancies. We aims to maximize the energy function Q which is defined by

$$\begin{aligned} \max_{t,n} \quad & Q_{t,n} = \frac{1}{n^2} t^T A t, \\ s.t. \quad & t_i \in \{0,1\}, |t| = n \end{aligned} \tag{3}$$

where t is an F-dimensional binary vector. If the fth element satisfies $t_f = 1$, the corresponding fth frame is selected. Otherwise, it is not selected. n is the number of selected frames.

Since it is difficult to solve (3) due to the binary constraint on the indicator vector t, we relax this constraint by replacing the vector t by $s = \frac{t}{n}$. Then the formulation (3) is equivalent to

$$\max_{s,n} \quad Q_{s,n} = s^T A s, \quad s.t. \quad s_i \in \left\{0, \frac{1}{n}\right\}, \quad |s| = 1. \tag{4}$$

Since each coordinate s_i of s is nonnegative, $|s| = 1$ is equivalent to $\sum_{i=1}^N s_i = 1$. By relaxing s_i to be within the range of [0, 1], we obtain the final formulation of the frame selection problem

$$\max_{s} \quad Q_s = s^T A s, \quad s.t. \quad s \in \Delta^N \tag{5}$$

[2]"Caffe: An open source convolutional architecture for fast feature embedding," [Online]. Available: http://caffe.berkeleyvision.org/

where $\Delta^F = \{s|s_i \geq 0, \forall i \text{ and } \sum_{i=1}^{F} s_i = 1\}$ is the standard simplex in the F-dimensional Euclidean space. By relaxing (3) to (5), the maximum over the original two variables, t and n, is replaced with the maximum over a single variable s. Once the solution s^* of (5) is obtained, we can easily recover the number of the selected frames n and the index of the selected frames: a frame I_f is selected if and only if $s_f^* > 0$. Consequently, the number of selected frames n is determined by the number of positive coordinates of s^*. Note that the number of selected frames is determined by the non-zero elements of s, which is essentially dynamically determined by the complexity of the video. Intuitively, complex and long videos will need more human labor to check the pose estimation results. The formulation is solved by the pairwise optimization method [21], [24].

The second step aims to integrate human feedback into video pose co-estimation. Little modification is required in the video pose co-estimation process compared with Section V. Instead of fixing the hypothesis number as 6 in Section V, the number of hypotheses is dynamically determined. Suppose that n frames generated by (3) are presented to the labelers. The labelers annotate n_1 among the n frames as correct, and the remaining $n_2 = n - n_1$ frames as wrong. Thus, for the frame I_f, the original human pose result as well as the selected n_1 frames (transformed by Sift Flow) constitute its hypothesis. The n_2 frames annotated as "Bad" poses are removed from the hypothesis set to avoid any error propagation. Except for the seed poses, the remaining video pose co-estimation procedure is the same.

VII. Video Co-Parsing

Given the refined human poses for all frames, we can perform the video co-parsing, conditioned on the image and 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 pixels $i' \in I_f$ (denoted as blue rectangles), and then co-parsing all frames by considering the super-pixel correspondences (denoted as red circles) to obtain $P(\mathcal{Y}|\mathcal{I}, \mathcal{H})$.

3) 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 the pixel i'. The confidence score $P(y_{f,i'}|I_f, h_f)$ of assigning the clothing item label to $y_{f,i'}$ can be computed by the existing fashion parser, e.g., [39]. 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 [7], by properly redesigning video co-parsing solution.

4) 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 work [39], we build dense appearance correspondences for super-pixels instead of pixels. We first compute over-segmentations of all frames using a fast segmentation method [9]. Then the confidence score of assigning the clothing item label to each super-pixel s is computed by 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. $\phi_s(y_f(s))$ is defined as the cost converted by its corresponding confidence. 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, which aims to maximize the probability $P(\mathcal{Y}|\mathcal{I}, \mathcal{H})$, which we define by

$$\begin{aligned} P(\mathcal{Y}|\mathcal{I}, \mathcal{H}) \propto \exp(-(&\sum_{f \in I} \sum_{s \in \wedge_f} \phi_s(y_f(s)) \\ &+ \sum_{(s,q) \in N_{int}} \varphi_{int}(y_f(s), y_f(q)|I_f) \\ &+ \sum_{(s,q) \in N_{ext}} \varphi_{ext}(y_f(s), y'_f(q)|I_f, I'_f))) \end{aligned} \tag{6}$$

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

$$\begin{aligned} &\varphi_{\text{ext}}(y_f(s), y_f(q)|I_f) \\ &\qquad = \delta(y_f(s) \neq y_f(q)) \exp(-\lambda_{ext} |\mathcal{F}(s) - \mathcal{F}(q)|) \\ &\varphi_{\text{int}}(y_f(s), y_f(q)|I_f, I_f{}') \\ &\qquad = \delta(y_f(s) \neq y'_f(q)) \exp(-\lambda_{int} |\mathcal{F}(s) - \mathcal{F}(q)|) \end{aligned} \tag{7}$$

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 the ℓ_2-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 (7) is a submodular function, the optimization of maximizing (6) becomes a tractable graphical model. We solve this optimization problem by the well-studied α-expansion method [11]. Thus the optimal parsing results of all frames can be calculated as $\mathcal{Y}$.

VIII. Human Parsing With Video Context

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 [13] to roughly locate the human body. The caffe feature for each human is computed, which can intrinsically capture the style, pose and appearance characteristics of the whole image. We use the ℓ_2-distance over the Caffe feature to find 25

TABLE I
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	mean
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	76.43
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**	**78.17**

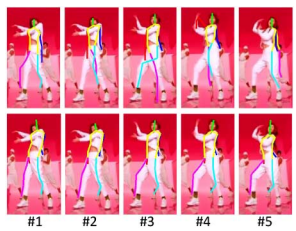


Fig. 5. Two comparison examples between image based human pose estimation (top row) and video based human pose co-estimation (bottom row).

nearest neighbors in our gallery. After that, we follow [39] and use the image segmentation algorithm [9] to obtain superpixels with the parameters of $\sigma = 0.8$, $k = 30$. Each super-pixel of the testing image finds the closest super-pixel from each retrieved image using the ℓ_2-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

$$P(y_s|s, D) = \frac{1}{Z} \sum_{r \in D} \frac{M(y_s, s_r)}{1 + \|h(s) - h(s_r)\|} \tag{8}$$

where we define

$$M(y_s, s_r) = \frac{1}{|s_r|} \sum_{i' \in s_r} \delta(y_{r,i'} = y_s) \tag{9}$$

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 by the weighted average of the parsing results of the 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 Markov Random Field to respect boundaries of actual clothing items.

IX. EXPERIMENTS

A. Experimental Setting

We conduct the experiments on three datasets. The first is the Fashionista (FS) dataset [40] containing 685 photos with good visibility of the full body and covering a variety of labels. 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 human parsing Data (CFPD) [22] dataset which consists of 2,682 images. The training set and the testing set are 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.

TABLE II
COMPARISON BETWEEN IMAGE-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**

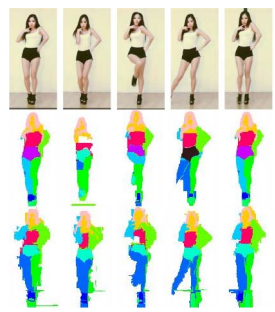


Fig. 6. Comparison examples between image based human parsing (middle row) and video based co-parsing (bottom row). For the color-label map, please refer to Fig. 1.

The FIdataset is more challenging than FS and CFPD since some girls cross or stretch their arms or legs freely, and may be in arbitrary views. In our experiments, the label sets of FS and CFPD contain 18 and 13 kinds of labels, respectively. FI has two sets of label sets, one containing 18 kinds of labels as FS, and the other containing 13 kinds as CFPD, where the later is obtained by merging from the former.

We can parse an image of resolution 600×400 in 2 seconds on a PC with Core I7 3.4 GHz GPU and 6 GB memory. The parameter η, λ_{ext} and λ_{int} are set as 0.1, 0.5 and 0.5 empirically in this work.

B. Experimental Results

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

1) Video Parsing—Video Pose Co-Estimation: We evaluate the performance of the human pose co-estimation method on predicting the poses of video 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 [41], which is trained on FS and predicts the pose of each frame separately. The standard PCK (Probability of Correct Key point) metric [41] is used to evaluate the performance of pose estimation. Table I 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

TABLE III
COMPARISON AMONG PAPER DOLL [39] AND THREE VERSIONS 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
	Active-Co-Pose + Co-Parsing	**88.92**	**58.95**	**57.68**	43.57	**44.90**
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**
	Active-Co-Pose + Co-Parsing	**85.97**	**54.66**	**43.90**	**42.42**	41.35
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
	Active-Co-Pose + Co-Parsing	**88.40**	**53.67**	**43.74**	**39.68**	**38.79**
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
	Active-Co-Pose + Co-Parsing	**86.11**	**55.05**	**43.89**	**42.54**	**41.41**

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 visualize the pose estimation results of the two comparison methods in Fig. 5. For the dancing video frames with large variations in poses and views, our method shows superior performance in predicting the key points of human poses, especially for the left and right knees. As shown in Fig. 5, 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.

2) Video Parsing—Video Co-Parsing: We compare the performances of our video co-parsing method with the existing image-based human parser [39], whose codes are publicly available.[3] The 100 videos are randomly selected and all frames are manually labeled. Similar to [39], 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 II. Significant improvements of our co-parsing method for all five metrics can be observed.

More exemplar results are shown in Fig. 6. The video co-parser predicts more consistent labels for all video frames than the image-based human 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.

3) Human Parsing—FS and CFPD: We report the human parsing performance of the baseline, i.e., Paper Doll and our method on testing images in 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. "Active-Co-Pose+Co-Parsing" indicates the results after active learning is used during the video parsing process. We invite 10 participants (3 females and 7 males who are university students and staffs) to label 500 videos for us. The labelers need judge the correctness of the estimated human poses for the videos and the labeling interface.

[3][Online]. Available: http://www.cs.sunysb.edu/~kyamagu/research/paperdoll/

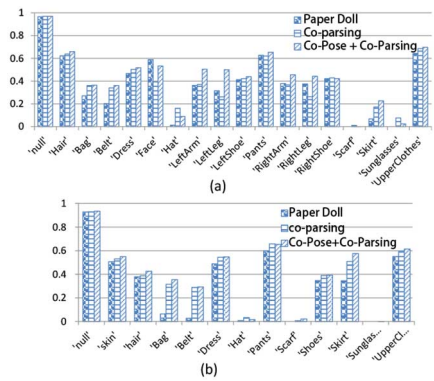


Fig. 7. F-1 scores of each class of Paper Doll [39] and three versions of our methods (Co-Parsing, Co − Pose + Co − Parsing and Active-Co-Pose+Co-Parsing) in the FS-FS and CFPD-CFPD settings. (a) FS-FS. (b) CFPD-CFPD.

The results are listed in the first two rows of Table III. It is obvious that both of our two solutions achieve higher performances 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 proved, where the avg. F1-score of "Co-Pose+Co-Parsing" outperforms "Co-Parsing" by 4.54% in the FS-FS experiment setting. Finally, "Active-Co-Pose+Co-Parsing" also outperforms "Co-Pose+Co-Parsing", e.g., by 2.17% in the FS-CFPD experiment setting.

We present the F1-score for each label in Fig. 7(a) and Fig. 7(b). Generally, the "Active-Co-Pose+Co-Parsing" shows the highest performance. Besides, our semi-supervised human parser "Co-Pose+Co-Parsing" achieves the second best results, especially when predicting the human body labels, such as "LeftLeg", "LeftArm", "RightLeg" and "RightArm". The inferior performance of "Co-Pose+Co-Parsing" than "Active-Co-Pose+Co-Parsing" is due to the lack of human labeling. We can observe this superior performance from the visualization of parsing results in Fig. 8(a). The first row shows the parsing results of Paper Doll and the second row shows the results of "Co-Pose+Co-Parsing". It can be observed that our parser performs better on predicting the fashion labels, such as "skirt", "pants", and "upper-clothes". In addition, our results

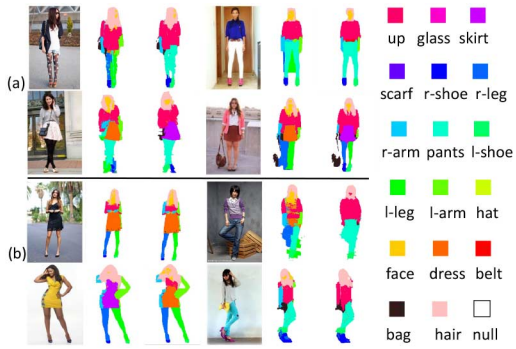


Fig. 8. Comparison of two settings: (a) FS-FS, (b) FS-FI. In each triple, the original image, the parsing result by Paper Doll and our result are shown sequentially.

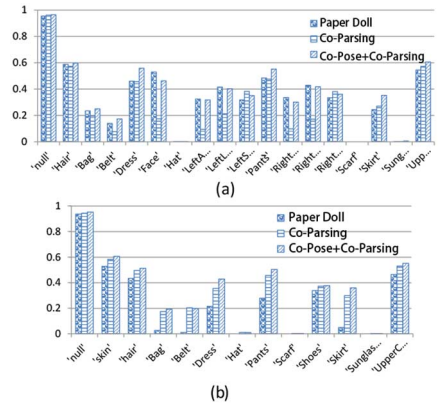


Fig. 9. F-1 scores of each class of Paper Doll [39] and three versions of our methods (Co-Parsing, Co − Pose + Co − Parsing and Active-Co-Pose+Co-Parsing) in the FS-FI and CFPD-FI settings. (a) FS-FI. (b) CFPD-FI.

can be less disturbed by the background clutter and show relatively clearer boundary and appearance consistency, e.g. the leg regions of the second and third images in the first row of Fig. 8(a). Moreover, our parser can also correctly localize small fashion labels, such as the bags in the second and third images in the second row of Fig. 8(a) and the third image in the first row of Fig. 8(b).

4) Human Parsing—FI: Our collected FI dataset contains more images with diverse poses and arbitrary views. We parse images in the FI dataset with the trained models from two training image sets, FS and CFPD, separately. The main difference between FS-FI and CFPD-FI is that we train two supervised parsing models with different training data and label sets. Similarly, we compare three solutions of our method, i.e., "Active-Co-Pose+Co-Parsing", "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 III. It is worth noting that our method shows much larger improvements on our collected FI dataset than on the existing datasets (i.e., FS and CFPD). Specifically, with the same training dataset CFPD, the performance of our "Active-Co-Pose+Co-Parsing" outperforms Paper Doll by 16.21% in the CFPD-FI setting, which is much higher than by 8.69% in the CFPD-CFPD setting. This well proves the advantages of our method on parsing challenging human photos.

The detailed comparison of each label among "Paper Doll", "Co-Parsing", "Co-Pose+Co-Parsing" and "Active-Co-Pose+Co-Parsing" in both FS-FI and CFPD-FI settings is illustrated in Fig. 9. In general, "Active-Co-Pose+Co-Parsing" performs the best. "Co − Pose + Co − Parsing" outperforms "Co-Parsing" and performs much better than Paper Doll [39].

Moreover, the visual parsing comparisons are shown in Fig. 8(b) and Fig. 10(b) for the FS and CFPD datasets, respectively. Our system can correctly predict the labels for images with very diverse human poses, e.g., the second image of the first row of Fig. 9(a).

Additionally, we conduct experiments of parsing the multi-human images under the FS-FI setting, as shown in Fig. 11. We use the detection method [13] to cut the images into several smaller images with a 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.

We show several human parsing results of "Active-Co-Pose+Co-Parsing" in Fig. 12. From the results, we can see that our human parser can handle the half body human photos and humans with very flexible poses.

Fig. 10. 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.

Fig. 11. The results of our system in parsing images with multiple humans in the FS-FI setting.

Fig. 12. Six comparisions of our method and Paper Doll. In each triplet, the original images, results of Paper Doll, and our results are shown sequentially. From the resutls, we can see that our method produced much more accurate parsing results than Paper Doll.

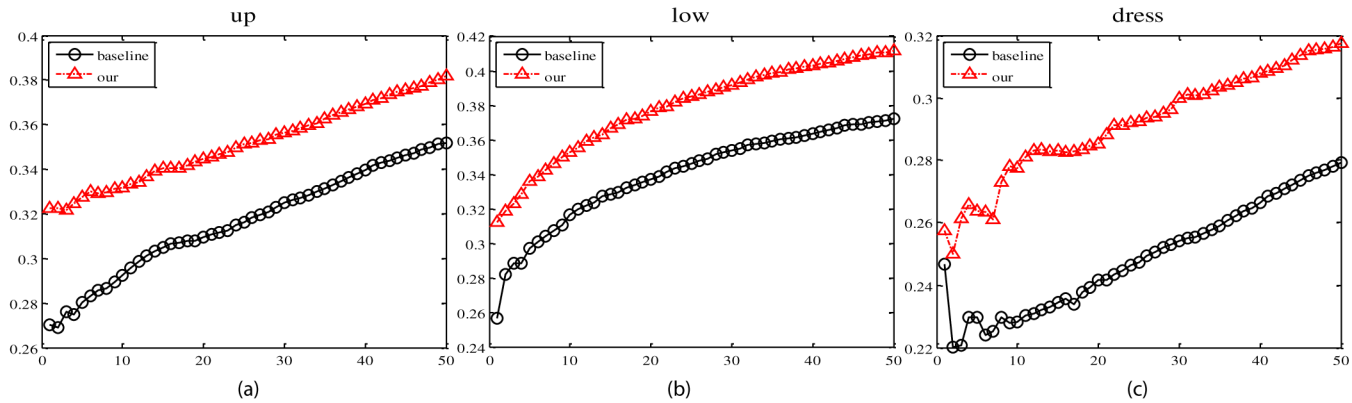


Fig. 13. Quantitative comparisons between our method and baseline method for the upper (a), lower (b) and dress (c). X-axis represents the number of rank, Y-axis is the corresponding NDCG value.

From the quantitative results in Fig. 7 and Fig. 9, the qualitative parsing results in Fig. 8, Fig. 10 and Fig. 11, we can see that our algorithm tends to work better in parsing larger labels, such as "upper clothes", "dress", but sometimes fails in parsing small and less common labels, such as "Scarf" and "Sunglasses". Moreover, although our algorithm still suffer from incorrectly estimated human pose estimation, which will results in wrong foreground-background separation and wrongly estimated foreground labels.

5) Human Parsing Assisted Clothing Retrieval: In this subsection, we will show how the human parsing can help clothing retrieval. Given a query image, we first parse it using the proposed human parsing method. Besides the query image, the users also decide which label to be retrieved. For example, the users want to find skirts which are similar with the one in the query image. We use the CNN feature because it has shown to achieve good performance in object recognition [18], [29], [18], [42], we use the deep convolutional activation feature to represent the region. More specifically, the Caffe feature is extracted from the corresponding region. Finally, the Euclidean distance is calculated and used as the ranking criterion.

To test the clothing retrieval performance, we use all the FI human photos as query and use the on-line shop images in street-to-shop work [25] as repository. Then we evaluate the retrieval performance for upper and lower-body clothing respectively. Similar with the street-to-shop work [25], we evaluate the retrieval performance based on whether the attributes of the query and the retrieval results are the same. All the online shop images in the street-to-shop [25] are extensively labeled by 10 kinds of attributes. So we manually label all the attributes of all images in the FI dataset. We compare with the baseline, i.e., street-to-shop work [25]. Normalized Discounted Cumulative Gain (NDCG) is used as the evaluation metric. The comparisons between our method and street-to-shop are shown in Fig. 13(a), (b) and (c) for dress, upper and lower-body respectively. We can see that for all ranks, our method is better than the baseline. The gain is mainly due to the more precise localization of the labels. Some typical retrieval results are shown in Fig. 14(a), (b) and (c) for dress, upper and lower body respectively.

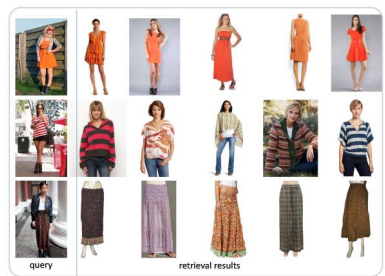


Fig. 14. Three examples of clothing retrieval results: (a) dress, (b) top, and (c) skirt. The leftmost images are the query images. The right columns show some retrieval results.

X. Conclusion and Future Work

In this paper, we proposed a semi-supervised framework for human parsing which leverages video contexts without extra annotation. It contains two components: the contextual video parsing and the non-parametric human parsing. Extensive experiments on two benchmark human parsing datasets as well as a newly collected FI dataset well demonstrate the effectiveness of our proposed framework. We can optionally label more images to train a better human parser. However, the great burden of human labelling may considerably limit the scalability of the human 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. If the human labeling are available, we can requires labelers to check the video pose co-estimation results and then train a better human parser.

Two possible research directions can be considered in future. First, we plan to develop a mobile application, which can parse images uploaded by users in the server-side. Second, inspired by the great advantage of deep learning in classification [19], [42] and detection [12], we will try to solve the human parsing problem in the Convolutional Neural Network (CNN) architecture. Current deep learning based image parsing models [8], [28], [12] require a large amount of training data. We believe that the framework introduced in this paper can greatly enrich the training dataset by exploring the video and finally make training the CNN possible.

REFERENCES

[1] V. Badrinarayanan, F. Galasso, and R. Cipolla, "Label propagation in video sequences," in *Proc. IEEE Conf. Comput. Vis. Pattern Recog.*, 2010, pp. 3265–3272.

[2] D. Batra, A. Kowdle, D. Parikh, J. Luo, and T. Chen, "icoseg: Interactive co-segmentation with intelligent scribble guidance," in *Proc. IEEE Conf. Comput. Vis. Pattern Recog.*, 2010, pp. 3169–3176.

[3] Y. Bo and C. C. Fowlkes, "Shape-based pedestrian parsing," in *Proc. IEEE Conf. Comput. Vis. Pattern Recog.*, 2011, pp. 2265–2272.

[4] J. Carreira, R. Caseiro, J. Batista, and C. Sminchisescu, "Semantic segmentation with second-order pooling," in *Proc. Comput. Vis.*, 2012, pp. 430–443.

[5] H. Chen, Z. Xu, Z. Liu, and S. C. Zhu, "Composite templates for cloth modeling and sketching," in *Proc. IEEE Conf. Comput. Vis. Pattern Recog.*, 2006, pp. 943–950.

[6] A. Cherian, J. Mairal, K. Alahari, and C. Schmid, "Mixing body-part sequences for human pose estimation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recog.*, 2014, pp. 2361–2368.

[7] J. Dong, Q. Chen, W. Xia, Z. Huang, and S. Yan, "A deformable mixture parsing model with parselets," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2013, pp. 3408–3415.

[8] C. Farabet, C. Couprie, L. Najman, and Y. LeCun, "Learning hierarchical features for scene labeling," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 35, no. 8, pp. 1915–1929, Aug. 2013.

[9] P. F. Felzenszwalb and D. P. Huttenlocher, "Efficient graph-based image segmentation," *Int. J. Comput. Vis.*, vol. 59, no. 2, pp. 167–181, 2004.

[10] V. Ferrari, M. Marin-Jimenez, and A. Zisserman, "Progressive search space reduction for human pose estimation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recog.*, 2008, pp. 1–8.

[11] B. Fulkerson, A. Vedaldi, and S. Soatto, "Class segmentation and object localization with superpixel neighborhoods," in *Proc. IEEE Conf. Comput. Vis. Pattern Recog.*, 2009, pp. 670–677.

[12] R. Girshick, J. Donahue, T. Darrell, and J. Malik, "Rich feature hierarchies for accurate object detection and semantic segmentation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recog.*, 2014, pp. 580–587.

[13] R. B. Girshick, P. F. Felzenszwalb, and D. A. McAllester, "Object detection with grammar models," in *Proc. NIPS*, 2011, pp. 442–450.

[14] A. Guzman-Rivera, P. Kohli, D. Batra, and R. Rutenbar, "Efficiently enforcing diversity in multi-output structured prediction," in *Proc. AI Statist.*, 2014, pp. 284–292.

[15] S. Hare, A. Saffari, and P. HS. Torr, "Struck: Structured output tracking with kernels," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2011, pp. 263–270.

[16] A. Joulin, F. Bach, and J. Ponce, "Discriminative clustering for image co-segmentation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recog.*, 2010, pp. 1943–1950.

[17] P. Kohli, J. Rihan, M. Bray, and P. HS. Torr, "Simultaneous segmentation and pose estimation of humans using dynamic graph cuts," *Int. J. Comput. Vis.*, vol. 79, no. 3, pp. 285–298, 2008.

[18] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," in *Proc. NIPS*, 2012, pp. 1106–1114.

[19] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," in *Proc. NIPS*, 2012, pp. 1097–1105.

[20] C. Liu, J. Yuen, and A. Torralba, "Sift flow: Dense correspondence across scenes and its applications," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 33, no. 5, pp. 978–994, May 2011.

[21] H. Liu, X. Yang, L. J. Latecki, and S. Yan, "Dense neighborhoods on affinity graph," *Int. J. Comput. Vis.*, vol. 98, no. 1, pp. 65–82, 2012.

[22] S. Liu, J. Feng, C. Domokos, H. Xu, J. Huang, Z. Hu, and S. Yan, "Fashion parsing with weak color-category labels," *IEEE Trans. Multimedia*, vol. 16, no. 1, pp. 253–265, Jan. 2014.

[23] S. Liu, X. Liang, L. Liu, K. Lu, L. Lin, and S. Yan, "Fashion parsing with video context," in *Proc. ACM Multimedia*, 2014, pp. 467–476.

[24] S. Liu, H. Liu, L. J. Latecki, S. Yan, C. Xu, and H. Lu, "Size adaptive selection of most informative features," in *Proc. AAAI*, 2011, pp. 392–397.

[25] S. Liu, Z. Song, G. Liu, C. Xu, H. Lu, and S. Yan, "Street-to-shop: Cross-scenario clothing retrieval via parts alignment and auxiliary set," in *Proc. IEEE Conf. Comput. Vis. Pattern Recog.*, 2012, pp. 3330–3337.

[26] X. Liu, D. Tao, M. Song, J. Bu, and C. Chen, "Discriminative segment annotation in weakly labeled video," in *Proc. Comput Vis. Pattern Recog. Workshop*, 2014, pp. 512–519.

[27] D. Park and D. Ramanan, "N-best maximal decoders for part models," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2011, pp. 2627–2634.

[28] P. H. O. Pinheiro and R. Collobert, "Recurrent convolutional neural networks for scene labeling," in *Proc. ICML*, 2014, pp. 82–90.

[29] A. S. Razavian, H. Azizpour, J. Sullivan, and S. Carlsson, "CNN features off-the-shelf: An astounding baseline for recognition," in *Proc. Comput. Vis.Pattern Recog. Workshops*, 2014, pp. 512–519.

[30] B. Sapp, D. Weiss, and B. Taskar, "Parsing human motion with stretchable models," in *Proc. IEEE Conf. Comput. Vis. Pattern Recog.*, 2011, pp. 1281–1288.

[31] P. Sermanet, D. Eigen, X. Zhang, M. Mathieu, R. Fergus, and Y. LeCun, "Overfeat: Integrated recognition, localization and detection using convolutional networks," in *Proc. Int. Conf. Learning Represent.*, Apr. 2014.

[32] J. Shotton, M. Johnson, and R. Cipolla, "Semantic texton forests for image categorization and segmentation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recog.*, 2008, pp. 1–8.

[33] A. Shrivastava, S. Singh, and A. Gupta, "Constrained semi-supervised learning using attributes and comparative attributes," in *Proc. ECCV*, 2012, pp. 369–383.

[34] A. Toshev and C. Szegedy, "Deeppose: Human pose estimation via deep neural networks," in *Proc. Comput. Vis. Pattern. Recog.*, 2014, pp. 1653–1660.

[35] D. Tran and D. Forsyth, "Improved human parsing with a full relational model," in *Proc. ECCV*, 2010, pp. 227–240.

[36] S. Vijayanarasimhan and K. Grauman, "Active frame selection for label propagation in videos," in *Proc. ECCV*, 2012, pp. 496–509.

[37] N. Wang and H. Ai, "Who blocks who: Simultaneous clothing segmentation for grouping images," in *Proc. IEEE Int. Conf. Comput. Vis.*, Nov. 2011, pp. 1535–1542.

[38] W. Yang, P. Luo, and L. Lin, "Clothing co-parsing by joint image segmentation and labeling," in *Proc. IEEE Conf. Comput. Vis. Pattern Recog.*, 2014, pp. 3182–3189.

[39] K. Yamaguchi, M. H. Kiapour, and T. L. Berg, "Paper doll parsing: Retrieving similar styles to parse clothing items," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2013, pp. 3519–3526.

[40] K. Yamaguchi, M. H. Kiapour, L. E. Ortiz, and T. L. Berg, "Parsing clothing in fashion photographs," in *Proc. IEEE Conf. Comput. Vis. Pattern Recog.*, 2012, pp. 3570–3577.

[41] Y. Yang and D. Ramanan, "Articulated pose estimation with flexible mixtures-of-parts," in *Proc. IEEE Conf. Comput. Vis. Pattern Recog.*, 2011, pp. 1385–1392.

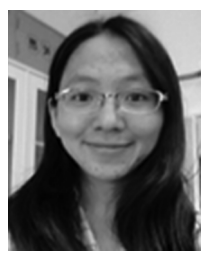
[42] M. D. Zeiler and R. Fergus, "Visualizing and understanding convolutional networks," in *Proc. ECCV*, 2014, pp. 818–833.

[43] T. Zhang, B. Ghanem, S. Liu, and N. Ahuja, "Low-rank sparse learning for robust visual tracking," in *Proc. ECCV*, 2012, pp. 470–484.

[44] T. Zhang, B. Ghanem, S. Liu, and N. Ahuja, "Robust visual tracking via multi-task sparse learning," in *Proc. IEEE Conf. Comput. Vis. Pattern Recog.*, 2012, pp. 2042–2049.

Si Liu (M'13) received the Ph.D. degree from the National Laboratory of Pattern Recognition, Institute of Automation, Chinese Academy of Sciences, Beijing, China, in 2012.

She is an Associate Professor with the Institute of Information Engineering, Chinese Academy of Sciences, Beijing, China. She was previously a Research Fellow with the Learning and Vision Group, National University of Singapore, Singapore. Her research interests includes computer vision and multimedia.



Xiaodan Liang is currently working toward the Ph.D. degree at the School of Information Science and Technology, Sun Yat-sen University, Guangzhou, China.

She is currently a Research Intern with the National University of Singapore, Singapore. Her research interests mainly include semantic segmentation, object/action recognition, and medical image analysis.

Luoqi Liu is currently working toward the Ph.D. degree at the Department of Electrical and Computer Engineering, National University of Singapore, Singapore.

His research interests include computer vision, multimedia, and machine learning.

Ke Lu received the M.S. and Ph.D. degrees from Northwest University, Shanxi Xian, China, in 1998 and 2003, respectively.

He was a Postdoctoral Fellow with the Institute of Automation, Chinese Academy of Sciences, Beijing, China, from July 2003 to April 2005. He is currently a Professor with the University of the Chinese Academy of Sciences. His current research interests include computer vision, 3-D image reconstruction, and computer graphics.

Liang Lin received the Ph.D. degree from the Beijing Institute of Technology, Beijing, China.

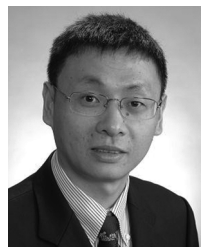
He is currently a Full Professor with the School of Advanced Computing, Sun Yat-Sen University, Guangzhou, China. He was previously a Post-Doctoral Research Fellow with the University of California at Los Angeles, Los Angeles, CA, USA. His research interests include new models, algorithms, and systems for intelligent processing and an understanding of visual data such as images and videos.

Prof. Lin was the recipient of the Runner-Up Best Paper Award in ACM NPAR 2010, the Google Faculty Award in 2012, and the Best Student Paper Award at IEEE ICME 2014.

Xiaochun Cao (M'14) received the B.E. and M.E. degrees from Beihang University, Beijing, China, and the Ph.D. degree in computer science from the University of Central Florida, Orlando, FL, USA.

After graduation, he was with ObjectVideo Inc., Beijing, China, for about three years. From 2008 to 2012, he was a Professor with Tianjin University, Tianjin, China. He is currently a Professor with the Institute of Information Engineering, Chinese Academy of Sciences, Beijing, China.

Prof. Cao is a Fellow of the IET. He was the recipient of the Piero Zamperoni Best Student Paper Award at the ICPR in 2004 and 2010.



Shuicheng Yan (M'06–SM'09) received the Ph.D. degree from the School of Mathematical Sciences, Peking University, Beijing, China, in 2004.

He is currently an Associate Professor with the Department of Electrical and Computer Engineering, National University of Singapore, Singapore, and the founding lead of the Learning and Vision Research Group, National University of Singapore. His research interests include machine learning, computer vision, and multimedia.

Prof. Yan is an ISI Highly-Cited Researcher of 2014, and an IAPR Fellow of 2014. He was the recipient of the Best Paper Awards from ACM MM'13 (Best Paper and Best Student Paper) and ACM MM12 (Best Demo).