Joint Parsing and Segmentation of Articulated Human Bodies From Videos

Zhao Liu[1], Jingrun Sun[2], Chun Chen[3]

[1,2,3]38 Zheda Road · Hangzhou · Zhejiang Province · 310027 · P. R. China

liuzhao@zju.edu.cn, zjusunjingrun@zju.edu.cn, chenc@zju.edu.cn

Keywords: Human pose; Segmentation; Parsing; Grabcut; Articulated Model

Abstract. Human body parsing and segmentation are two fundamental problems in computer vision. In this paper we build an automatic system for solving the two problems together. We reconstruct the mixture-of-part model by adding various features, because of the high relevance constructed by the new model, we are able to execute precise parsing inference on the whole human body poses. For segmentation we replace the user interaction input with the detection boxes, since the detection boxes fit the human body part well, it is easy for the refined segmentation to distinguish the foreground and background parts. Experiment results show apparent improvements compared with former methods.

Introduction

Human body parsing and segmentation have been two long-standing problems in computer vision. While recent work report amazing results on both the parsing and segmentation domain, few works mention of applying the two problems together, especially in challenging video clips with high confusions. Meichner et al. [1] segmented foreground regions before parsing but it was just an intermediate step. And the segmentation results contain too much of the background.

We design a framework aims at jointly solving parsing and segmentation on video clip frames. Our system is based on an temporal-correlated model for the human body pose and detecting each body joint, the detections data greatly constrain every joint's location, so we can estimate the body pose from the connection detections data. On the other hand, it is easy for us to segment the human body pose from the background via the shape prior provided by the parsing method.

The Parsing system establishes on two state-of-art models, while flexible mixture-of-arts model [2] is efficient in generating various kinds of articulated body poses frame by frame, it lacks the ability of predicting multi-frame relationships. The stretchable model [3] provides a more sparse structure for adding various features with rich motion, appearance and contour cues, the new added features construct the relationship between adjacent frames.

Grabcut [4] is useful in segmentation, but it need the interactive data from users, as body paring method predict the joints around the human body, we can use these joints as the prior knowledge for initializing, then we refine the result by applying Grabcut to each of the bounding boxes around all the body joints, thus the user need not do any interaction action during the process.

Related Work

Human Pose parsing. The region-based deformable model in [6] aims at solving the body parsing problem. This work was later used by Vittorio Ferrari et al. in their Progressive Search algorithm [7, 8]. Yi Yang's mixture-of-parts models [2] performs well in static images but it lacks the ability of capturing rich time information, thus is not capable for used in short video clips.

Video Object Segmentation. Segmentation on video objects in an interactive way has been studied for years, traditional techniques includes GraphCut [9], Grabcut [4] and random walks [10] approaches. Veksler implemented the Star Shape Prior [5] on Graph Cuts, and the extension is in [11] by Varun Gulshan, from a single star to multiple stars and from Euclidean rays to Geodesic paths.

Problem Formulation

The whole process of our system is shown in Figure 1. Our system can be decomposed into the human pose parsing on video frames and pose segmentation, with the detection data acting as the link between the two parts. The labeled video clips are first used by the parsing system to construct the human location boxes, the parsing system then learns the body structure from the detection area and inference the body pose in each frame. Finally the segmentation system performs Grabcut [4] based on the detection bounding boxes to separate the human body from the background.

Fig. 1 Overview. (a) Original image (b) Human location boxes built on the training data (c)Body joint locations detected by the parsing process (d)Foreground human body pose separated from the images.

We denote the whole problem as a prediction task similar to the articulated structure model [8]. We write I for an image in a video sequence which contains m frames, P for different body parts and L for different parts' locations. The location can be written as a function based on the frame and the body part L(p, m). So we build the parsing model:

$$S(P, I) = \sum_{x=1}^{m} \sum_{i \in v} L(p_i, I_x) + \sum_{x=1}^{m} \sum_{(i,j) \in \varepsilon} L(p_i, p_j; I_x) \tag{1}$$

The first part in (1) stands for the local scores type i while the second expresses the relationship for the pairwise scores. The pair-wise terms include the edge relation both in a single frame and in between frames, and the configuration score term can be decomposed to the inner product of parameter vector and feature functions.

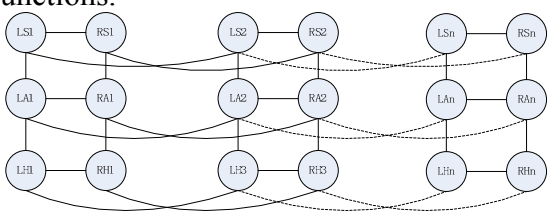

Fig. 2 Time-dependencies model. L and R stands for the left and right part of the human body and S, E, H separately for shoulder, elbow and hand joint.

Human pose parsing. For all the body parts, we compute the local score as:

$$m(p_i, p_j) = \max L(p_i, p_j) + \max s(p_i) = \max \alpha_i^T \alpha_j^T L(p_i - p_j, I_x) + \max s(p_i) \tag{2}$$

Here p_i represents the score in part i, it is the sum of the local score in i and the message its children passes to it. This process continues until the tree reach the root node (i = 1), and $s(p_i)$ represents the maximal score configurations for each node, And every part's location can be determined by backtracking the argmax indices of each root node.

To acquire all the joints' relations through the frame sequence, we extend the single frame inference process to include the time-dependencies, as in Figure 2. By merging the nodes between different frames, the new argmax problem can be coupled through the global constraints over the M (in our experiments 30) sub-models.

$$\arg\max_{I, I_1, I_2, \ldots, I_x} \sum_{x=1}^{M} \alpha_{i:C_p}^T f(p_i, I_x); \tag{3}$$

We solve the dual decomposition problem of (3) by using sub-gradient descent. The inference converges when all I_x reach an agreement.

Body pose segmentation. After the body parsing step, we have the detections data which provide us with the information where the human body pose located in the image, we use these detections to initialize the Grabcut process, and generate the initial foreground/background GMM based on the areas in or out of the detection boxes.

To Refine the results we transfer the detection bounding boxes to Grabcut. While some of the bounding boxes are out of the human pose range, not all of them can be used as the foreground label, so we choose the bounding boxes in lower head and the middle-torso, these parts are located totally in the body pose range, thus are capable for labeling the foreground.

Experiments

Dataset. Our experiment was on the Videopose 2 dataset [3], which contains a series of video clips taken from the TV shows Friends and Lost, and mainly on the upper body which includes head, torso and upper limbs. We choose 12 clips, 470 frame images in our experiments, and we use 200 images for training and 270 images for testing.

TABLE 1 PCP EVALUATIONS ON VIDEOPOSE 2

Methods	***Body Parts Accuracy***					
	Torso	***RLA***	***LLA***	***LHA***	***RHA***	***Total***
Yi et al. 2011[2]	4.0	7.9	2.2	4.2	4.6	4.6
Ours	4.1	8.1	2.5	4.2	4.6	4.7

Human pose Parsing. To compare our method with previous work, we choose the criterions of detection rate and the traditional probability of a correct pose (PCP). We evaluate the two rates both on the whole body pose and on every joints of the body parts, the results is shown in Table 1, our method performs better in the right low arm (RLA) and left low arm (LLA), however, it gets nearly the same results on the right high arm (RHA) and left high arm (LHA). It is because the added features are mainly on localize the lower arms and the upper arms move more frequently, temporal information is more important for relocating the score in a new frame. For evaluation with Ben Sapp's method we evaluate the detection accuracy with a given deviate range. Our method outperforms Ben Sapp's in predicting the joints' location on shoulder and hands parts, while Ben Sapp's method does better on prediction the elbow joints. Comparison of our method and Sapp's are shown in Figure 3.

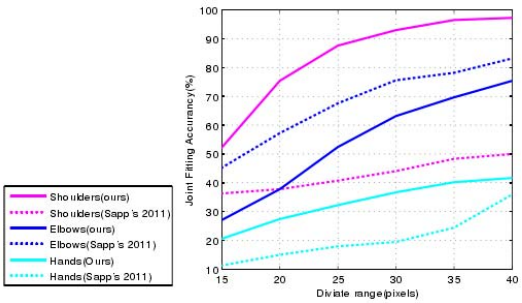

Fig. 3 Quantitative comparations of our method to Ben's strechable model [3], our method significantly exceeds Ben's in parsing shoulder and hand parts.

Body pose segmentation. We compare our work with the original Grabcut [4] algorithm and the Geodesic star convexity [5]. The comparison results are shown in Figure 4, we test the three methods on the VideoPose 2.0 dateset, With the whole 270 test frames with 4860 body parts, our method is able to recognize 90.61% of the data compared with 88.52% of Grabcut and 89.51% of Geodesic star convexity. While our method has significant less use interaction than the two algorithms. Our method can well handle images from different view-points, as well as with confusing background.

Conclusion

We propose a system which solves the human pose parsing and segmentation together. Our method can automatically handle video frame images with high occlusion. By adding various efficient features, the relationship of different frames can be constructed and we can infer the whole video clips parsing model. Since the parsing bounding boxes fit body joints, we use them to refine the Grabcut and get amazing results. In future work, we would consider using motion capture to optimize the parsing process and explore more applications.

Fig. 4 Example images compartion for our method with Geodesic star convexity(the second row) and the original Grabcut.

Acknowledgement

This work was supported by the National High Technology Research and Development Program of China(2013AA040601).

References

[1] Liu, Y., Stoll, C., Gall, J., Seidel, H.P., Theobalt., C.: Markerless motion capture of interacting characters using multi-view image segmentation. In CVPR (2011)

[2] Yang, Y., Ramanan., D.: Articulated pose estimation with flexible mixture-of- parts. In CVPR (2011)

[3] Sapp, B., Weiss, D., Taskar., B.: Parsing human motion with stretchable models. In CVPR (2011)

[4] B. Sapp, A.T., Taskar., B.: Cascaded models for articulated pose estimation. In ECCV (2010)

[5] Veksler., O.: Star shape prior for graph-cut image segmentation. In ECCV (2008)

[6] Ramanan., D.: Learning to parse images of articulated bodies. In NIPS (2006)

[7] Ferrari, V., Marin-Jimenez, M., Zisserman., A.: Progressive search space reduction for human pose estimation. In CVPR (2008)

[8] Eichner, M., Marin-Jimenez, M., Zisserman, A., Ferrari., V.: Articulated human pose estimation and search in (almost) unconstrained still images. ETH Zurich, D-ITET, BIWI, Technical Report (2010)

[9] Boykov, Y., Funka-Lea., G.: Graph cuts and efficient n-d image segmentation. International Journal of Computer Vision 70 (2006) 109–131

[10] Grady., L.: Random walks for image segmentation. PAMI 28 (2006) 1768–1783

[11] Gulshan, V., Rother, C., Criminisi, A., Blake, A., Zisserman., A.: Geodesic star convexity for interactive image segmentation. In CVPR (2010)