POSE ESTIMATION AND BODY SEGMENTATION BASED ON HIERARCHICAL SEARCHING TREE

Shifeng Li, Huchuan Lu
Department of Electronic Engineering
Dalian University of Technology, CHINA

ABSTRACT

In this paper, we propose a novel method for pose estimation and body segmentation. We estimate the partial configuration of adjacent parts instead of detecting each single part, which makes our method more robust and accurate. Further, we develop a general model to calculate the partial configuration. Besides, we present a tree-based hierarchical probabilistic method to derive the global optimal pose. Additionally, the coarse-to-fine strategy is employed to speed up the pose estimation in the whole procedure. After finishing pose estimation, the estimated pose is used to guide body segmentation. Experiments suggest that our method is efficient and effective for pose estimation and body segmentation simultaneously.

Index Terms— pose estimation, body segmentation, tree-based searching

1. INTRODUCTION

For the next generation cameras, it is greatly useful to obtain precise pose and segmentation of human body to facilitate photo edition, content based compression and re-rendering in future. However, it remains a challenging task due to the unknown image background, partial occlusion and the high dimensional parameter space.

The literature on human pose estimation is vast. The well-known pictorial structures (PS) method was proposed by [1], and extended by [2, 3, 4]. This method can obtain the global optimal pose with fast pose inference. However, the calculation of feature responses for all directions is time-consuming, making it impractical in practice.

Hierarchical method is popular in pose estimation. [5] detected head, main body, legs and arms sequentially for tracking. Navaratanam et al. [6] used 2D templates to detect upper body pose hierarchically. Lee and Nevatia [7] developed a hierarchical method to track human poses. Hu et al. [8] employed the face as guidance to estimate upper body pose based on Data-Driven Markov Chain Monte Carlo. The hierarchical method can efficiently finish pose estimation. However, the heavily dependency on part detectors and the hierarchy of searching scheme leads to inaccurate results.

To overcome these problems, we propose an efficiently hierarchical method for pose estimation and body segmentation. Our contribution is as follows: (i) To enhance the accuracy of pose estimation, we propose a partial configuration strategy where the configuration of whole body pose is decomposed into three main partial configurations. (ii) A general model is designed to compute partial configuration. (iii) We have developed a hierarchical probabilistic searching scheme, which can effectively avoid the heavy dependence between parts caused by hierarchy.

*College of Information Science and Engineering, Ritsumeikan University, Kusatsu, JAPAN

2. POSE ESTIMATION OF HUMAN BODY

Considering our special application, digital camera, it is wise and reasonable to detect the head with the assumption that people in images are upright. Since any body part likelihood may result in false positives, it is important to encode the higher order relationship between body parts to improve the discrimination. In our work, we use the partial configuration as pose estimation unit, as shown in fig. 1. Our method begins with the partial configuration of head and torso, then estimates the partial configuration of arms and legs based on the detected torso, which can be typically viewed as traditional hierarchical part-based method. To reduce the heavy dependencies between parts, we develop a three-level tree based searching scheme to efficiently obtain the global optimal pose in images. As shown in fig. 2, we keep several candidate nodes in layer L1, then explore other parts along each tree to find the global pose. In the following, we will introduce the details of pose estimation.

2.1. Partial configuration of our method

To calculate the partial configuration, a general model is designed. Formally, let the partial configuration be \( L = l_1, l_2, \ldots, l_n \), where \( l_i \) specifies one part. Every part \( l_i = (s_i, \theta_i) \) is determined by scale \( s_i \) and orientation \( \theta_i \), and the part to which it is connected (for simplicity, the size of \( l_i \) is fixed). Given an image \( I \), the partial configuration of parts can be formulated as follows:

\[
p(L|L_{pri}, I) \propto p(L_{pri}) \prod_i p(l_i|l_{par(i)}) \prod_i p(I|l_i)
\]

(1)

where \( p(L_{pri}) \) is the prior knowledge of partial configuration; \( p(l_i|l_{par(i)}) \) stands for the conditional likelihood between parts, and \( p(I|l_i) \) is the likelihood of a single part, representing the likelihood of seeing a particular image \( I \) given a special part configuration \( l_i \).
Hierarchical probabilistic searching tree. 'H', 'T', 'A' and 'L' stand for the configuration of head, torso, arms and legs; 'HT', 'HTA' and 'HTL' define the partial configuration of head and torso, torso and arms (arms), and torso and legs (legs).

2.1.1. Partial configuration of head and torso

In our implementation, the pose estimation starts from the configuration of head and torso (HT). Based on the general partial configuration model, the state vector L is the concatenation of sub-state vectors of head and torso, \( L_{ht} = \{ l_h, l_t \} = \{ l_t \} \). The partial configuration of HT \( p(L_{ht} | I) \) can be specified as:

\[
p(L_{ht} | I) \propto p(l_h)p(l_t) \prod_i p(I | l_i)
\]

where \( p(l_h) \) is the head prior, estimated by head detection; \( p(l_t | l_h) \) is the conditional distribution modeled as uniform distribution, and \( p(I | l_i) \) is specified by \( p(l_{ht} | l_i) \) and \( p(l_t | l_i) \), representing the likelihood of seeing the image \( I \) given the head and torso part location respectively. The specified model of HT is shown in fig. 3(a).

2.1.2. Partial configuration of arms

Given the estimated partial configuration of HT, we can further estimate the partial configuration of arms. Based on the general pose model, the state vector of arm configuration \( L_{arm} \) can be defined as \( L_{arm} = \{ l_t, l_{ua}, l_{la}, l_{ua}, l_{ra} \} = \{ l_t \} \). The posterior probability of a pose \( p(L_{arm} | L_{ht}, I) \) is formulated as

\[
p(L_{arm} | L_{ht}, I) \propto p(l_t) \prod_i p(l_{ia} | l_{par(i)}) \prod_i p(I | l_i)
\]

where \( p(l_t) \) is the prior of detected torso, obtained from the estimation of head and torso configuration \( p(l_t) \) is considered as the spatial prior for arms); The conditional prior \( p(l_{ia} | l_{par(i)}) \) stands for \( p(l_{ua} | l_t), p(l_{ua} | l_t), p(l_{ua} | l_{ra}), p(l_{ua} | l_{ra}) \) and they are also modeled as uniform distributions.

2.1.3. Partial configuration of legs

The partial configuration of legs is similar with that of arms. The general model can be easily applied to lower body, just as shown in fig. 3(c). The posterior probability of lower body pose is identical to Eq. 3, where \( p(l_t) \) is also the detected torso prior, the conditional prior \( p(l_{ia} | l_{par(i)}) \) represents \( p(l_{ua} | l_t), p(l_{ua} | l_t), p(l_{ua} | l_{ra}), p(l_{ua} | l_{ra}) \). The related likelihood \( p(I | l_i) \) used here is also the same as those in arm pose estimation.

2.2. Likelihood \( p(I | l_i) \) of our models

In our models, there are two kinds of appearance. One is the histogram of oriented gradients (HOG) [9], and the other is the segmentation cue [10], obtained from category-specific shape mask.

\[1\] Here, 'lua' stands for left upper arm, 'rua' represents right upper arm, 'lla' is left lower arm and 'lla' defines right lower arm.

2.3. Coarse-to-fine strategy

To accelerate part detection, coarse-to-fine technology is used. We first detect several top part candidates using one feature (e.g. HOG). Based on the detected candidates, we then add segmentation cue to find the best candidate.

3. HIERARCHICAL PROBABILISTIC DETECTION

Our method starts from configuration of head and torso, then estimates the other partial configurations based on the estimated part, which can be viewed as a three level tree (fig. 2). Motivated by [11], we formulate the combination of partial configuration as maximization of a global pose estimation likelihood \( P \). In order to quickly evaluate the likelihood for a global detection, we model \( P \) as a summation of detection scores of each partial configuration in all layers. We can think of \( P \) as log-likelihood of the summation of the detection scores over different partial configurations, which is equivalent to the multiplication of probabilities. Let a candidate pose estimation model be \( P = \{ P(l_i) \} \). Based on the tree structure, the likelihood \( P \) is decomposed into conditional likelihood as follows:

\[
P(L | I) = P(L_{ht} | I) + P(L_{arm} | L_{ht}, I) + P(L_{leg} | L_{ht}, L_{arm}, I)
\]

where the decomposition is performed in a top-to-bottom order of the layers, and independence is assumed between the two non-jointing parts, i.e. arms and legs. \( P(L_{ht} | I) \) is head and torso partial configuration given the image \( I \) (Sec. 2.1.1); \( P(L_{arm} | L_{ht}, I) \) defines arms configuration score (Sec. 2.1.2) and \( P(L_{leg} | L_{ht}, I) \) is the leg configuration score (Sec. 2.1.3). Algorithm 1 shows the tree based searching scheme.

More precisely, in step 3, we can keep several top partial configuration candidates of arms \( P(L_{arm} | L_{ht}, I), i = 1, 2, \cdots , K_{arm} \) in layer L2 and obtain the optimal configuration \( P(L_{leg} | L_{ht}, I) \) in layer L3, then compute the likelihood based on Eq. 5. Besides, we
can estimate two corresponding thresholds $t_i, i = 1, 2$ for top two layers to decide whether further to explore in next layer. In our work, we keep the first 5 top nodes at Layer L1 and 3 at Layer L2, and use greedy searching at Layer L3.

Algorithm 1. Probabilistic Hierarchical detection
1: Keeping $K_{ht}$ candidates and the corresponding detection scores $P(L_{ht}^i), k = 1, 2 \ldots, K_{ht}$.
2: For $k = 1 \ldots K_{ht}$, repeat the following steps (3)-(4), and select $k^* = k^*$ and $L = L^*$ with the maximum $P(L, I)$.
3: According to the candidates of head and torso configuration $L_{ht}$ of layer L1, estimating the maximum partial configuration of arms $P(L_{arm}^i|L_{ht}, I)$ in layer L2 and legs configuration $P(L_{leg}^i|L_{ht}, I)$ in layer L3 using a greedy search algorithm.
4: Given the above part model estimates, computing the current global likelihood based on Eq. 5.
5: Return the global detection estimate $L^* = \{L_{ht}^*, L_{arm}^*, L_{leg}^*\}$ and the corresponding parameters of models.

4. POSE ESTIMATION GUIDING SEGMENTATION

After finishing the pose estimation, we can use the pose to guide the body segmentation. Given an image $I$, $\tau = \{\tau_i | i = 1, 2, \ldots, n\}$ denotes the set of binary class labels for all pixels ($\tau_i = 0$ for background and $\tau_i = 1$ for foreground) where $n$ is the number of pixels in image $I$. Following the segmentation energy definition from [12]

$$E(\tau) = \lambda \sum_{i=1}^{n} R(\tau_i) + \sum_{i \neq j} V_{i,j} \sigma(\tau_i, \tau_j)$$

where $R(\tau_i)$ is the region term, relating to the posterior probability of pixel $i$ belonging to class $\tau_i; V_{i,j}$ is the boundary term, which represents the consistence between two pixels $i$ and $j; \sigma(\tau_i, \tau_j) = 1$ when $\tau_i \neq \tau_j$, otherwise $\sigma(\tau_i, \tau_j) = 0; \lambda$ specifies the relative importance between two terms.

After pose estimation, employing the strategy of soft-map [13], we use logistic regression to compute region term $R(\tau_i)$

$$R(\tau_i = 1) = -\log(1/(1 + e^{-\beta x}))$$

$$R(\tau_i = 0) = -\log(1 - 1/(1 + e^{-\beta x}))$$

where $\beta$ is the parameter obtained from related parts; $x$ is the R, G, B color features. The boundary term $V_{i,j}$ is defined in the same way as in [12]. To enhance the accuracy of segmentation, the segmentation task is divided into upper and lower body segmentation.

5. EXPERIMENTAL RESULTS

To evaluate our method, 200 images with size 208*156 are collected. The images cover various poses, backgrounds, clothing, individuals and lighting conditions. 50 images are used to train the parameters of our method. The classifier used in our experiments is Libsvm, which is used as probability estimates with linear kernel.

5.1. Training classifiers and shape prior

This subsection introduces the training of two classifiers (HOG and segmentation based classifiers) and the category-specific shape prior. Shape prior training: Our category-specific shape prior for a foreground/background segmentation is a binary value mask where ‘1’ stands for foreground and ‘0’ represents background. They are obtained from the pixel-based independent Bernoulli model in a canonical coordinate frame [10, 3]. Suppose each pixel $x$ has a prior probability $p_{fg}(x)$ of being foreground. Given $T$ training windows with ground-truth segmentation $g_{fg}$, one could fit $p_{fg}$ by its maximum likelihood estimate (MLE): $p_{fg}(x) = \frac{1}{T} \sum_t g_{fg}^t$.

Classifiers training: Our classifiers are based on local regions, which can characterize the classifiers with distinctive properties. For HOG classifier training, positive samples are first cropped from the training images, aligned to a uniform orientation, and resized to a fixed size (head 32*32, torso 80*64, arms 24*16 and legs 32*24), and the negative samples are randomly derived around the positive samples in the same manner. Based on the category-specific shape prior, the segmentation-based classifier training is similar to that of HOG-based classifier.

5.2. Evaluation and comparison of pose estimation

The top of fig. 4 lists some pose estimation samples in our dataset and the bottom shows the corresponding pose estimation samples using the method of [2] (we run the code obtained from the author’s home page on our dataset). For our results, the torso is labeled with a blue rectangle as the first level pose, the second level poses including head, upper arms and legs are labeled with red rectangles and the rest parts are drawn with green rectangles as the last level. As shown in fig. 4, the most poses estimated by our approach are better than those estimated by [2]. Additionally, to further evaluate pose estimation in our dataset, the number of correctly detected parts containing head, torso, upper and lower arms and legs (if the overlap between the detected part and the corresponding part groundtruth is larger than 0.5) is counted and reported in Fig. 5. The x-axis is the testing images and the y-axis...
is the correctly detected part number. The red line stands for our method while the blue one represents the approach of [2]. In most cases, our method is better than [2] (about 142/150). Moreover, we find that the torso and head detection in both methods (our method and [2]) is better than other parts. For our approach, the torso and head detection is the best, and the legs detection is better than arms. It seems that the parts having little variety are easily detected.

5.3. Evaluation and comparison of segmentation

Fig. 6 shows some examples of body segmentation, which consists of eight group images. Each group has two images, the original image and the segmentation result. As shown in fig. 6, the poses and backgrounds in images are different from each other; the background color of some images is highly similar to the foreground (torso color in (a), trousers color in (b)); the illumination in some images changes drastically (e, f, g) and the clothing varies from coat (d, g), sweater (a, h), shirt (b, I) to T-shirt (c). Nevertheless, all students are successfully segmented from the images by our approach.

Comparison of segmentation: in order to test the accuracy of our method and compare with other method, we use human-labeled images as the ground truth, and use $e = \sum_x (R(x) \otimes G(x))/\sum_x (R(x) \cup G(x))$ to evaluate the accuracy performance, where $R(x)$ and $G(x)$ denote the value of pixel $x$ for segmentation and ground truth. $\otimes$ and $\cup$ perform pixel-wise XOR and OR.

According to the above evaluation criterion, it is obvious that the smaller of the ratio is, the better result will be. Based on the criterion, a comparison with grab cut [14] is illustrated in fig. 7 where we labeled the rectangle prior knowledge of grab cut when segmenting the testing images. As shown in fig. 7, our method (red line) is obviously better than grab cut and most evaluation results of our method range from 0.1 to 0.3. It suggests that our method can segment the human body from images more accurately.

In our experiment, torso detection is relatively accurate and the region of torso is larger than other parts, which determines its main influence to region term calculation of Eq. 6 and leads to the precision of body segmentation. Besides, we find that grab cut is sensitive to pose variety. If the pose varies little (the arms or legs limited near the main body), the result of grab cut will be better, otherwise the segmentation will contains more background.

6. CONCLUSION

In this paper, we propose a novel method for pose estimation and body segmentation. Using partial configuration as detection unit, we first detect the head and torso configuration, then estimate the partial configuration of arms and legs based on the detected torso with the designed general model. Furthermore, employing the hierarchy of the body parts, a tree searching strategy is proposed to derive the global optimal pose configuration. Moreover, the achieved pose can further guide body segmentation. Experiments on our dataset suggest that our method is efficient and effective.

7. ACKNOWLEDGEMENT

The work was supported by the Fundamental Research Funds for the Central Universities, No. DUT10JS10, the National Natural Science Foundation of China (NSFC), No.61071209 and Omron Corporation’s grant No.DUT09001.

8. REFERENCES