Master's Thesis

Human Pose Estimation using Part-based Region Matching

Sueyoung Oh (오 수 영)

Department of Computer Science and Engineering

Pohang University of Science and Technology

2016

파트 기반 영역 매칭을 이용한 사람 자세 추정

Human Pose Estimation using Part-based Region Matching

Human Pose Estimation using Part-based Region Matching

by

Sueyoung Oh
Department of Computer Science and Engineering
Pohang University of Science and Technology

A thesis submitted to the faculty of the Pohang University of Science and Technology in partial fulfillment of the requirements for the degree of Master of Science in the Computer Science and Engineering

Pohang, Korea
12. 24. 2015
Approved by
Joon Hee Han
Academic advisor

Human Pose Estimation using Part-based Region Matching

Sueyoung Oh

The undersigned have examined this thesis and hereby certify that it is worthy of acceptance for a master's degree from POSTECH

12. 24. 2015

Committee Chair Joon Hee Han

Member Daijin Kim

Member Ki-Sang Hong

MCSE 오 수 영. Sueyoung Oh
20130732 Human Pose Estimation using Part-based Region Matching,
파트 기반 영역 매칭을 이용한 사람 자세 추정
Department of Computer Science and Engineering , 2016, 16p, Advisor : Joon Hee Han. Text in English.

ABSTRACT

In this thesis, a part-based region matching algorithm is proposed for human pose estimation in 2D images. A new notion of part, named a semantic part is introduced. A semantic part is represented as a combination of classic rigid parts and contains partial semantic information of body pose. Region proposals are used to form a set of candidate bounding boxes for semantic parts. These regions are matched between target and source images and their confidences are evaluated by computing the matching score. The regions with high confidence is the final semantic parts which form a body pose. Based on a data-driven approach, the final pose is estimated by getting information of joint positions from the source correspondences of the final semantic parts. Using semantic information catches more meaningful pose information and the part-based region matching has simple algorithm and performs effectively for a large number of data.

Contents

I. Introduction **1**

II. Related Work **2**

2.1 Human Pose Estimation 2

III. Proposed Method **3**

3.1 Overview 3

3.2 Semantic Part 5

3.2.1 Definition 5

3.2.2 Criteria of Semantic Part 6

3.2.3 Semantic Part Selection Algorithm 7

IV. Experiment **8**

4.1 Datasets 8

V. Result **9**

5.1 Quantitative Result 9

VI. Conclusion **10**

Summary (in Korean) **11**

References **12**

I. Introduction

Human pose estimation from a single static image is a challenging problem in computer vision. Various models for pose estimation have been proposed over the last decade and estimated human pose is applied to diverse high level vision tasks such as image understanding [1] and action recognition [2].

Human pose is represented as configuration of multiple body parts, which generally are parameterized by pixel location and orientation. The estimation of human pose can be considered as a part-based object detection problem, specifically for the case of articulated objects. This means an object is modeled by a collection of parts arranged in a deformable configuration. In terms of a problem of localization of body parts, pose estimation is more difficult problem than other general object detection because human body is highly articulated and has a large number of degrees of freedom to be estimated. The large pose variations, cluttered background, self-occlusions are challenging aspects of human pose estimation.

II. Related Work

2.1 Human Pose Estimation

Pictorial structure Part-based representation is a classic framework for human pose estimation. A part-based model represents the human body as a constellation of a set of rigid parts constrained in some fashion. Most work [14, 15, 16] on human pose estimation are graphical model-based for these body parts. In terms of part-based object detection, pose estimation is mainly represented by the pictorial structure models which is related to mixtures of parts [3, 17]. The pictorial structure model [3] is efficient to solve object recognition and model learning problems. The pictorial structure decomposes the appearance of objects into local part templates, together with geometric constraints between pairs of parts. This models the spatial relations of rigid parts using usually a tree model which allows for efficient inference.

III. Proposed Method

3.1 Overview

The task of human pose estimation is examined in static images. Human pose estimation from a single static image is a challenging problem in computer vision. Various models for pose estimation have been proposed and a working technology is applied to diverse high level vision tasks such as image understanding [1] and action recognition [2].

In general, human body pose is represented by configuration of multiple rigid parts located at joints. Body parts usually are parameterized by pixel location and orientation. The appearance of a body is decomposed into local parts, and the deformable configuration is defined with geometric constraints between a pair of parts. The goal of human pose estimation is to find every pixel location of body joints. That means to find the best configuration of multiple body parts.

The rigid parts are based on human anatomy. A human pose is labeled joint positions and they provide part positions as limbs. A body limb is represented by the joint points as endpoints of its region. In this sense, the word "rigid parts" appears to be used interchangeably with the expression "joints" and "limbs" here. Figure 3.1 shows the human body model based on rigid parts.

Let $p_i \in \{p_1, \ldots, p_K\}$ denote the i-th joint point, where K is the total number of joint points. Note that we are using the expression "rigid parts" distinctively from the notion "semantic part", introduced next.

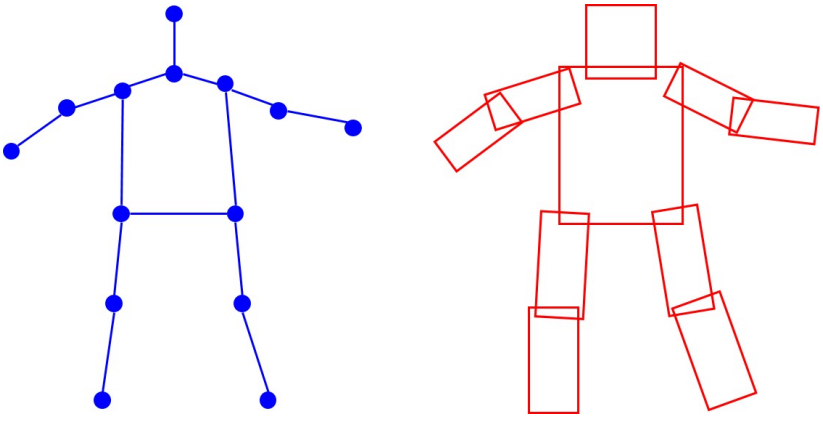

Figure 3.1: The graph model of the articulated human body. **left** : Generally, human body is modeled with joint points based on human anatomy. It is parameterized by K body points located at joints ($K = 14$). **right** : Rigid parts are represented as limbs by using the joints as endpoints of the limbs. Most previous approaches build detectors for these rigid parts. In this thesis, the final estimation is based on this human body model as the configuration of joint points, but a new approach is proposed using mid-level representation named a 'semantic part', which is different from a rigid part.

3.2 Semantic Part

In most previous part-based models [4], rigid body parts such as head, torso, half limbs are guided by human anatomy. This representation of parts are difficult to detect accurately because single rigid part which is represented by parallel lines is not discriminative in appearance and does not contain any contextual information. This may commonly occur in clutter and induce self-occlusion. To handle large variance in appearance of human pose, the use of larger parts with more context needs to be considered. Some researches [24, 7, 8] used larger or combined parts such as poselet [9].

3.2.1 Definition

Our notion of "parts" can range from basic rigid parts (e.g. torso, head, half-limb), to large pieces of bodies covering more than one rigid part. In the extreme case, the "parts" corresponding to the whole body. This "part" is named as a *semantic part*. In other words, a semantic part spans multiple rigid parts (or joint points) on the body. For example, a semantic part is 'head + torso', 'torso + left arm' and 'left leg + right leg'.

A semantic part is described as a set of one or more body limbs, in other words, subset of whole body parts. Let S_i be the i-th semantic part, where i is written as $i \in \{1, ..., N\}$ and N is the number of semantic parts which form a human body. We define a semantic part as:

Definition

$$S_i \in \mathcal{P}(p_1, \ldots, p_K) \backslash \{\emptyset\}, \qquad \text{for } i = 1, \ldots, N \tag{3.1}$$

where $\mathcal{P}(\cdot)$ denotes the power set, the set of all subsets of a set. A semantic part is one of elements of the power set of all joint points $\{p_1, \ldots, p_K\}$, excluding the empty set. This involves that in the extreme case, a semantic part can even be

a rigid part or the whole body.

However, Definition 3.1 occurs disconnected composition of joints. The disconnected composition cannot be considered semantically meaningful and this cannot even be represented within a image patch. Accordingly, more strict definition for a semantic part is necessary. The new definition of a semantic part is written as follows.

Definition

$$S_i \in \{G' | G' \text{ is a connected component of graph } G\}, \quad \text{for } i = 1, \ldots, N \quad (3.2)$$

3.2.2 Criteria of Semantic Part

The term semantic part is used to describes a part of a human pose. According to our intention, we argue that a "good" semantic part must satisfy the criteria as in the following.

1. It should be easy to find semantic part given the input image. This suggests that the semantic part must be discriminative in appearance.
2. A semantic part should be contain sufficiently semantic information. This means that the semantic part muse be semantically discriminative with the appearance.

The criteria above involves the relation between semantics and appearance of the semantic part. In brief, the semantic part must be discriminative in both semantics and appearance. Semantic information of human pose in an image is interpreted as the geometry relation of joint positions. In next section, how the semantic parts are selected with consideration of the relation between appearance and geometry information in human poses is described.

Semantic part	left legs/arms	right legs/arms	**head+torso+arms**	**legs**
Correlation	0.408	0.399	0.391	0.383

Table 3.1: Semantic parts with high correlation. Bold : the semantic parts used in our experiments.

Semantic part	right arms	arms	left lower arm	right lower arm
Correlation	0.163	0.119	0.071	0.067

Table 3.2: Semantic parts with low correlation

3.2.3 Semantic Part Selection Algorithm

Algorithm 1 Semantic part selection

1: Set candidates of semantic parts as combinations of rigid parts.
2: Sample patches of semantic parts, from each image of dataset.
3: Extract appearance/geometry features from the patches.
4: Compute similarities between every pair of the patches, for each appearance/geometry feature.
5: Compute correlation between appearance and geometry features, for each semantic part.
6: Select semantic parts which have high correlation.

IV. Experiment

4.1 Datasets

Experiments are performed on two standard pose estimation benchmarks: the Leed Sports Pose (LSP) dataset [12] and the Image Parse dataset [13]. The LSP dataset contains 2000 pose annotated images, that contains 1000 training and 1000 test images from sport activities with annotated full-body human poses. The Parse dataset contains 305 pose-annotated images of highly articulated full body images of human poses. The Parse dataset is not a good dataset for the proposed data-driven approach, because the number of the images is too small. Experiments in this thesis are performed mostly on LSP dataset which contains relatively enough number of image data among human pose datasets. As the images have low resolution and contain partially occluded people with all kinds of poses and viewpoints, our test is challenging. Both datasets included a standard train/test split. For the proposed model, the training set are used for source images and the test set for target images.

V. Result

5.1 Quantitative Result

Proposed method is compared to the state-of-the-art method [4] that uses a flexible mixture of parts modeled by linear SVMs. The available source code published by [4] are used. To compare with other previous work using part detectors, a joint representation is converted into a limb representation by using the joints as endpoints of the limbs.

VI. Conclusion

A part-based region matching method is proposed to estimate human poses in 2D images. As the elements of matching, a semantic part, a new part representation which is a combination of classic rigid parts is introduced. A semantic part is expected to contain sufficient semantic information of human pose. They are easier to be detected rather than detecting rigid parts which is represented by parallel lines and not discriminative in appearance. A part-based matching between source and target regions is performed for human pose estimation. A set of candidate bounding boxes are generated from the target image by extracting object proposals, which restrict the search space. The matching algorithm evaluates matches between two sets of regions by considering both appearance and spatial consistency. The matching algorithm leads to maximize the total score for the final pose. The total score is computed by region confidences on each joint point. Finally, several best matches of regions are selected as semantic parts to form final human pose. By inferring the pose information from the source regions of the best match, every joint positions for whole body pose can be estimated.

요 약 문

본 논문은 2차원 영상에서 사람 자세 추정(human pose estimation)을 위해 1) 신체 파트를 관절 단위가 아닌 의미적 영역으로 구분하고, 2) 파트 검출기를 학습하는 대신 파트 기반의 영역 매칭 알고리즘(part-based region matching)을 제안한다.

References

[1] Nam-Gyu Cho, Alan L. Yuille, and Seong-Whan Lee. Adaptive occlusion state estimation for human pose tracking under self-occlusions. *Pattern Recognition*, 46(3):649–661, 2013.

[2] Chunyu Wang, Yizhou Wang, and Alan L. Yuille. An approach to pose-based action recognition. *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 0:915–922, 2013.

[3] Pedro F. Felzenszwalb and Daniel P. Huttenlocher. Pictorial structures for object recognition. *International Journal of Computer Vision (IJCV)*, 61(1):55–79, January 2005.

[4] Yi Yang and D. Ramanan. Articulated pose estimation with flexible mixtures-of-parts. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1385–1392, 2011.

[5] Xianjie Chen and Alan Yuille. Articulated pose estimation by a graphical model with image dependent pairwise relations. In *Advances in Neural Information Processing Systems (NIPS)*, 2014.

[6] Norimichi Ukita. Articulated pose estimation with parts connectivity using discriminative local oriented contours. In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3154–3161, 2012.

[7] Min Sun and S. Savarese. Articulated part-based model for joint object detection and pose estimation. In *IEEE International Conference on Computer Vision (ICCV)*, pages 723–730, 2011.

[8] Leonid Pishchulin, Mykhaylo Andriluka, Peter Gehler, and Bernt Schiele. Poselet conditioned pictorial structures. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2013.

[9] Lubomir Bourdev and Jitendra Malik. Poselets: Body part detectors trained using 3d human pose annotations. In *International Conference on Computer Vision (ICCV)*, 2009.

[10] Minsu Cho, Suha Kwak, Cordelia Schmid, and Jean Ponce. Unsupervised Object Discovery and Localization in the Wild: Part-based Matching with Bottom-up Region Proposals. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2015.

[11] C. Lawrence Zitnick and Piotr Dollár. Edge boxes: Locating object proposals from edges. In *European Conference on Computer Vision (ECCV)*, 2014.

[12] Sam Johnson and Mark Everingham. Clustered pose and nonlinear appearance models for human pose estimation. In *Proceedings of the British Machine Vision Conference (BMVC)*, 2010.

[13] Deva Ramanan. Learning to parse images of articulated bodies. In B. Schölkopf, J. Platt, and T. Hoffman, editors, *Advances in Neural Information Processing Systems (NIPS)*, pages 1129–1136. 2007.

[14] Yang Wang and Greg Mori. Multiple tree models for occlusion and spatial constraints in human pose estimation. In David Forsyth, Philip Torr, and Andrew Zisserman, editors, *European Conference on Computer Vision (ECCV)*, volume 5304, pages 710–724. 2008.

[15] Mykhaylo Andriluka, Stefan Roth, and Bernt Schiele. Pictorial structures revisited: People detection and articulated pose estimation. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2009.

[16] Fang Wang and Yi Li. Beyond physical connections: Tree models in human pose estimation. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 596–603. IEEE, 2013.

[17] P.F. Felzenszwalb, R.B. Girshick, D. McAllester, and D. Ramanan. Object detection with discriminatively trained part-based models. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 32(9):1627–1645, 2010.

[18] Deva Ramanan and Cristian Sminchisescu. Training deformable models for localization. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 206–213, 2006.

[19] Leonid Sigal and Michael J. Black. Measure locally, reason globally: Occlusion-sensitive articulated pose estimation. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2041–2048, 2006.

[20] Lubomir Bourdev, Subhransu Maji, Thomas Brox, and Jitendra Malik. Detecting people using mutually consistent poselet activations. In *European Conference on Computer Vision (ECCV)*, 2010.

[21] Subhransu Maji, Lubomir Bourdev, and Jitendra Malik. Action recognition from a distributed representation of pose and appearance. In *IEEE International Conference on Computer Vision and Pattern Recognition (CVPR)*, 2011.

[22] Lubomir Bourdev, Subhransu Maji, and Jitendra Malik. Describing people: Poselet-based attribute classification. In *International Conference on Computer Vision (ICCV)*, 2011.

[23] Weilong Yang, Yang Wang, and G. Mori. Recognizing human actions from still images with latent poses. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2030–2037, June 2010.

[24] Yang Wang, Duan Tran, and Zicheng Liao. Learning hierarchical poselets for human parsing. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1705–1712, 2011.

[25] M. Dantone, J. Gall, C. Leistner, and L. van Gool. Human pose estimation using body parts dependent joint regressors. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3041–3048, june 2013.

[26] Jonathan Tompson, Ross Goroshin, Arjun Jain, Yann LeCun, and Christoph Bregler. Efficient object localization using convolutional networks. *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, abs/1411.4280, 2014.

[27] Gunhee Kim and Antonio Torralba. Unsupervised detection of regions of interest using iterative link analysis. In Y. Bengio, D. Schuurmans, J.D. Lafferty, C.K.I. Williams, and A. Culotta, editors, *Advances in Neural Information Processing Systems (NIPS)*, pages 961–969. 2009.

[28] Thomas Deselaers, Bogdan Alexe, and Vittorio Ferrari. Localizing objects while learning their appearance. In *European Conference on Computer Vision (ECCV)*, pages 452–466, 2010.

[29] Kevin Tang, Armand Joulin, Li-Jia Li, and Li Fei-Fei. Co-localization in real-world images. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2014.

[30] Bogdan Alexe, Thomas Deselaers, and Vittorio Ferrari. What is an object? In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 73–80, 2010.

[31] Santiago Manén, Matthieu Guillaumin, and Luc Van Gool. Prime object proposals with randomized prim's algorithm. In *International Conference on Computer Vision (ICCV)*, December 2013.

[32] N. Dalal and B. Triggs. Histograms of oriented gradients for human detection. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, volume 1, pages 886–893, 2005.

[33] V. Ferrari, M. Marin-Jimenez, and A. Zisserman. Progressive search space reduction for human pose estimation. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2008.

[34] Piotr Dollár and C. Lawrence Zitnick. Structured forests for fast edge detection. In *International Conference on Computer Vision (ICCV)*, December 2013.

Acknowledgements

때로는 엄하고 때로는 부드러운 모습으로 지도 해주신 한준희 교수님, 감사드립니다.

Curriculum Vitae

Name : Sueyoung Oh

Education

2009. 3. – 2013. 2. Department of Computer Science and Engineering, Inha University (B.S.)

2013. 3. – 2016. 2. Department of Computer Science and Engineering, Pohang University of Science and Technology (M.S.)

Experience

2013. 4. – 2013. 12. Developed a finger sign detection system (LG Electronics Inc.)