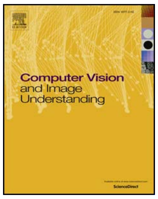
Contents lists available at ScienceDirect

Computer Vision and Image Understanding

journal homepage: www.elsevier.com/locate/cviu



Semi- and weakly-supervised human pose estimation

Norimichi Ukita*,[a], Yusuke Uematsu[b]

[a] Toyota Technological Institute, Japan
[b] Nara Institute of Science and Technology, Japan

ARTICLE INFO

Keywords:
Human pose estimation
Semi-supervised learning
Weakly-supervised learning
Pose clustering

ABSTRACT

For human pose estimation in still images, this paper proposes three semi- and weakly-supervised learning schemes. While recent advances of convolutional neural networks improve human pose estimation using supervised training data, our focus is to explore the semi- and weakly-supervised schemes. Our proposed schemes initially learn conventional model(s) for pose estimation from a small amount of standard training images with human pose annotations. For the first semi-supervised learning scheme, this conventional pose model detects candidate poses in training images with no human annotation. From these candidate poses, only true-positives are selected by a classifier using a pose feature representing the configuration of all body parts. The accuracies of these candidate pose estimation and true-positive pose selection are improved by action labels provided to these images in our second and third learning schemes, which are semi- and weakly-supervised learning. While the first and second learning schemes select only poses that are similar to those in the supervised training data, the third scheme selects more true-positive poses that are significantly different from any supervised poses. This pose selection is achieved by pose clustering using outlier pose detection with Dirichlet process mixtures and the Bayes factor. The proposed schemes are validated with large-scale human pose datasets.

1. Introduction

Human pose estimation is useful in various applications including context-based image retrieval, etc. The number of given training data has a huge impact on pose estimation as well as various recognition problems (e.g, general object recognition Krizhevsky et al., 2012 and face recognition Taigman et al., 2014), Although the scale of datasets for human pose estimation has been increasing (e.g., 305 images in the Image Parse dataset in 2006 (Ramanan, 2006), 2K images in the LSP dataset (Johnson and Everingham, 2010) in 2010, and around 40K human poses observed in 25K images in the MPII human pose dataset (Andriluka et al., 2014), it is difficult to develop a huge dataset for human pose estimation in contrast to object recognition (e.g., over 1,430K images in ISVRC2012–2014 Russakovsky et al., 2014). This is because human pose annotation is much complicated than weak label and window annotations for object recognition.

To increase the number of training images with less annotation cost, semi- and weakly-supervised learning schemes are applicable. Semi-supervised learning allows us to automatically provide annotations for a large amount of data based on a small amount of annotated data. In weakly-supervised learning, only simple annotations are required in training data and are utilized to acquire full annotations in learning.

We apply semi- and weakly-supervised learning to human pose estimation, as illustrated in Fig. 1. In our method with all functions proposed in this paper, fully-annotated images each of which has a pose annotation (i.e., skeleton) and an action label are used to acquire initial pose models for each action (e.g., “Baseball” and “Tennis” in Fig. 1). These action-specific pose models are used to estimate candidate human poses in each action-annotated image. If a candidate pose is considered true-positive, the given pose with its image is used for re-learning the corresponding action-specific pose model.

The key contributions of this work are threefold:

- True-positive poses are selected from candidate poses based on a pose feature representing the configuration of all body parts. This is in contrast to a pose estimation step in which only the pairwise configuration of neighboring/nearby parts is evaluated for efficiency.
- The action label of each training image is utilized for weakly-supervised learning. Because the variation of human poses in each action is smaller, pose estimation in each action works better than that in arbitrary poses.
- A large number of candidate poses are clustered by Dirichlet process mixtures for selecting true-positive poses based on the Bayes factor.

* Corresponding author at: 2-12-1, Tempaku, Nagoya 468-8511, Japan. He worked at Nara Institute of Science and Technology formerly.
E-mail address: ukita@ieee.org (N. Ukita).

https://doi.org/10.1016/j.cviu.2018.02.003
Received 1 August 2017; Received in revised form 13 December 2017; Accepted 5 February 2018

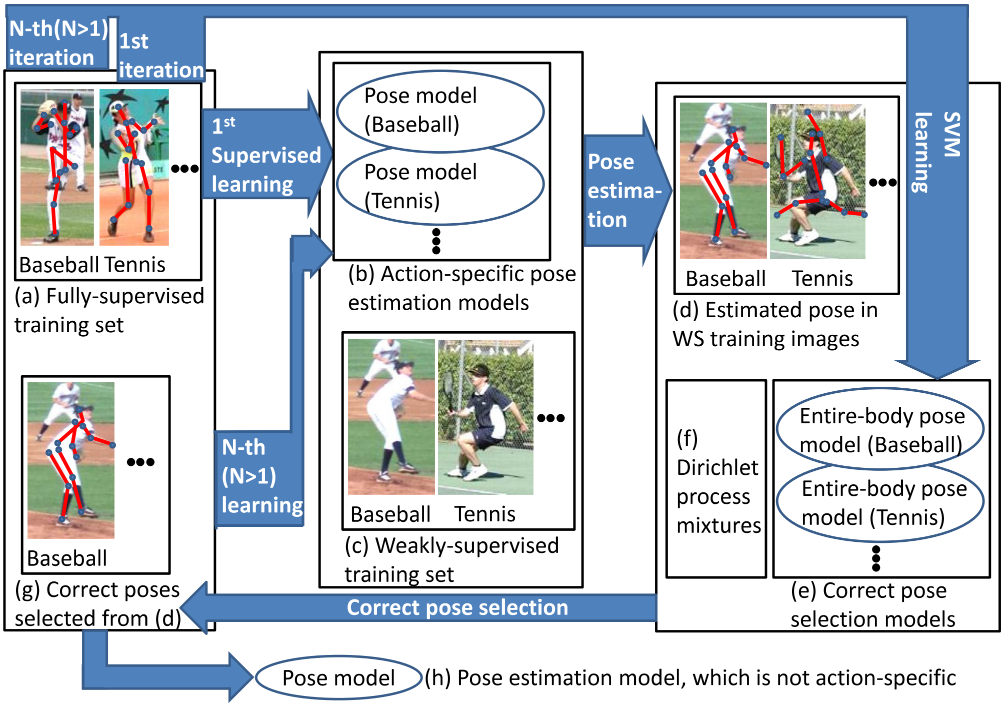

N-th(N>1) iteration
1st iteration
Baseball Tennis
(a) Fully-supervised training set
1st Supervised learning
Pose model (Baseball)
Pose model (Tennis)
(b) Action-specific pose estimation models
Pose estima-tion
Baseball Tennis
(d) Estimated pose in WS training images
SVM learning
N-th (N>1) learning
Baseball Tennis
(c) Weakly-supervised training set
Baseball
(g) Correct poses selected from (d)
(f) Dirichlet process mixtures
Entire-body pose model (Baseball)
Entire-body pose model (Tennis)
(e) Correct pose selection models
Correct pose selection
Pose model
(h) Pose estimation model, which is not action-specific

Fig. 1. Overview of the proposed method. Images each of which has a pose annotation and an action label (i.e., Fig. 1 (a)) are used to acquire initial action-specific pose models (i.e., (b)). Each model is acquired from images along with its respective action label. *a*-th action-specific pose model is used for estimating human poses in training images with the label of *a*-th action (i.e., (c)). Each estimated pose (i.e., (d)) is evaluated whether or not it is true-positive. This evaluation is achieved by a pose feature representing the configuration of all body parts (i.e., (e)). True-positive poses are selected also by outlier detection using Dirichlet process mixtures (i.e., (f)). These true-positive poses are employed as pose annotations (i.e., (g)) for re-learning the action-specific pose models. After iterative re-learning, all training images with pose annotations (i.e., (a) and (g)) are used for learning a final pose model (i.e., (h)). While this paper proposes three learning schemes, described in Sections 4, 5, and 6, this figure illustrates the third one, which contains all functions proposed in Sections 4, 5, and 6.

2. Related work

A number of methods for human pose estimation employed (1) deformable part models (e.g., pictorial structure models Felzenszwalb and Huttenlocher, 2005) for globally-optimizing an articulated human body and (2) discriminative learning for optimizing the parameters of the models (Felzenszwalb et al., 2010). In general, part connectivity in a deformable part model is defined by image-independent quadratic functions for efficient optimization via distance transform. Image-dependent functions (e.g., Sapp et al., 2010; Ukita, 2012) disable distance transform but improve pose estimation accuracy. In Chen and Yuille (2014), on the other hand, image-dependent but quadratic functions enable distance transform for representing the relative positions between neighboring parts.

While global optimality of the PSM is attractive, its ability to represent complex relations among parts and the expressive power of hand-crafted appearance feature are limited compared to deep neural networks. Recently, deep convolutional neural networks (DCNNs) improve human pose estimation as well as other computer vision tasks. While DCNNs are applicable to the PSM framework in order to represent the appearance of parts as proposed in Chen and Yuille (2014), DCNN-based models can also model the distribution of body parts. For example, a DCNN can directly estimate the joint locations (Toshev and Szegedy, 2014). In Tompson et al. (2014), multi-resolution DCNNs are trained jointly with a Markov random field. Localization accuracy of this method (Tompson et al., 2014) is improved by coarse and fine networks in Tompson et al. (2015). Recent approaches explore sequential structured estimation to iteratively improve the joint locations (Carreira et al., 2016; Ramakrishna et al., 2014; Singh et al., 0000; Wei et al., 2016). One of these methods, Convolutional pose machines (Wei et al., 2016), is extended to real-time pose estimation of multiple people (Cao et al., 2017) and hands (Simon et al., 2017). Pose estimation using DCNNs is also extended to a variety of scenarios such as personalized pose estimation in videos (Charles et al., 2016) and 3D pose estimation with multiple views (Pavlakos et al., 2017). As well as DCNNs accepting image patches, DNNs using multi-modal features are applicable to human pose estimation; multi-modal features extracted from an estimated pose (e.g., relative positions between body parts) are fed into a DNN for refining the estimated pose (Ouyang et al., 2014).

While aforementioned advances improve pose estimation demonstrably, all of them require human pose annotations (i.e., skeletons annotated on an image) for supervised learning. Complexity in time-consuming pose annotation work leads to annotation errors by crowd sourcing, as described in Johnson and Everingham (2011). For reducing the time-consuming annotations in supervised learning, semi- and weakly-supervised learning are widely used.

Semi-supervised learning allows us to utilize a huge number of non-annotated images for various recognition problems (e.g., human action recognition (Jones and Shao, 2014), human re-identification (Liu et al., 2014), and face and gait recognition (Huang et al., 2012). In general, semi-supervised learning annotates the images automatically by employing several cues in/with the images; for example, temporal consistency in tracking (Li et al., 2011), clustering (Mahmood et al., 2014), multimodal keywords (Guillaumin et al., 2010), and domain adaptation (Jain and Learned-Miller, 2011).

For human pose estimation also, several semi-supervised learning methods have been proposed. However, these methods are designed for limited simpler problems. For example, in Navaratnam et al. (2006) and Kanaujia et al. (2007), 3D pose models representing a limited variation of human pose sequences (e.g., only walking sequences) are trained by semi-supervised learning; in Navaratnam et al. (2006) and Kanaujia et al. (2007), GMM-based clustering and manifold regularization are employed for learning unlabeled data, respectively. For semi-supervised learning, not only a small number of annotated images but also a huge amount of synthetic images (e.g., CG images with automatic pose annotations) are also useful with transductive learning (Tang et al., 2013).

In weakly-supervised learning, only part of full annotations are given manually. In particular, annotations that can be easily annotated are given. For human activities, full annotations may include the pose, region, and attributes (e.g., ID, action class) of each person. Since it is easy to provide the attributes rather than the pose and region, such attributes are often given as weak annotations. For example, only an action label is given to each training sequence where the regions of a person (i.e. windows enclosing a human body) in frames are found automatically in Shapovalova et al. (2012). Instead of the manually-given action label, scripts are employed as weak annotations in order to find correct action labels of several clips in video sequences in