Computer Vision and Image Understanding 141 (2015) 95–107

Contents lists available at ScienceDirect

Computer Vision and Image Understanding

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

Structured forests for pixel-level hand detection and hand part labelling

Xiaolong Zhu*, Xuhui Jia, Kwan-Yee K. Wong

Department of Computer Science, The University of Hong Kong, Chow Yei Ching Building, Pokfulam Road, Hong Kong

ARTICLE INFO

Article history:
Received 18 October 2014
Accepted 28 July 2015

Keywords:
Hand detection
Egocentric vision
Random forests
Hand part labelling

ABSTRACT

Hand detection has many important applications in Human-Computer Interactions, yet it is a challenging problem because the appearance of hands can vary greatly in images. In this paper, we present a new approach that exploits the inherent contextual information from structured hand labelling for pixel-level hand detection and hand part labelling. By using a random forest framework, our method can predict hand mask and hand part labels in an efficient and robust manner. Through experiments, we demonstrate that our method can outperform other state-of-the-art pixel-level detection methods in ego-centric videos, and further be able to parse hand parts in details.

1. Introduction

Hand detection has many important applications in Human-Computer Interactions. It enables computers to consider the flexible movement of human hands in 3D space as a new type of high dimensional user input, and to understand the natural interaction of hands with other objects in various scenarios. However, hand detection is a challenging problem because the appearance of hands can vary greatly in images. For instance, the shape of a hand can change dramatically due to the articulation of fingers as well as changes in viewpoint. A hand can be (partially) occluded while interacting with other objects. The colour of a hand can vary greatly under different illuminations, and a hand can even appear to be textureless under extreme illuminations. Traditional Methods [1–4] based on gradients or skin detection often cannot handle practical unconstrained hand images well due to insufficient training data. Furthermore, ego-centric cameras have become more and more popular. Images captured by such cameras often have a dynamic background, which makes hand detection even more difficult. Nonetheless, hands play a major part in these images, and it is of great interest and importance to detect hands in detail robustly for further higher level analysis.

In this paper, our goal is to improve pixel-level hand detection and hand part labelling within the random forest framework. Rather than predicting per-pixel labels independently as in [5], we aim at exploiting the inherent structure from the label output space and predicting a patch region, which corresponds to a binary shape mask in hand detection and a multi-class label patch in hand part labelling. Technically, our approach is inspired by Semantic Texton Forests [27] and recent work on semantic image labelling [28]. During their training process, only limited number of pixels of a patch were considered in the split function. In order to consider more pixels, we propose to use an intermediate mapping, which groups the training patches for each node into certain amount of clusters by means of unsupervised learning methods. As shown in Fig. 1, our method detects hand regions more robustly than previous methods and is able to parse a hand into different parts.

Our proposed approach has the following contributions:

- we explicitly model the labelling of a pixel together with its local neighbourhood as a structured output to better utilise the inherent topological information in the training data and enforce such information as constraints during estimation;
- a novel structured split criterion is proposed to enable an efficient training and consider more pixels of our structured forests by incorporating unsupervised learning methods;
- we extend the binary hand detection to multi-class hand part labelling within our unified framework to solve these problems in an efficient and robust manner;
- throughout the experiments, our method outperforms the state-of-the-art methods. We also present a comprehensive analysis on different factors affecting the performance of our method on both tasks.

Next, we briefly review related work on pixel-level hand detection in Section 2. In Section 3, we describe our proposed structured forests for hand detection. In Section 4, we extend our structured forests to handle more general output and apply them to hand part labelling. In Section 5, we show the experimental results for both hand detection and hand part labelling. Finally, we conclude our method in Section 6.

* Corresponding author.
E-mail address: lucienxlzhu@gmail.com (X. Zhu).

http://dx.doi.org/10.1016/j.cviu.2015.07.008
1077-3142/

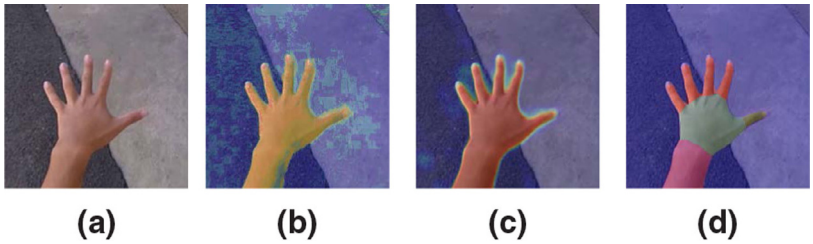

Fig. 1. Introduction to our method. (a) Original image. (b) Pixel-level hand detection by single pixel prediction. (c) Pixel-level hand detection by structured mask prediction. (d) Hand part labelling by structured label prediction.

1.1. Literature review

For many years, hand detection has been studied as a part of gesture analysis and human layout parsing. Early efforts in detecting human hand from a colour image usually considered skin-colour as the major cue [2,3,6] to build a model of the hand region in colour space. Mixture of Gaussians [7] was commonly used to model colours of skin and non-skin regions for hand localisation [8] and hand tracking [9,10]. As these methods often require a priori knowledge of skin colour, extracted either from training data or from face detection, to build the skin model, they cannot obtain robust results when they are applied to a novel scene or when illumination changes cast a large variation in colour.

In the mean time, inspired by the great progress in object detection and recognition, a few works directly modelled the appearance of hands with a generic object detection framework. Features could be extracted from a number of training images to train a Viola & Jones-like boosted detector [1,11,12] or an HOG-SVM detector [4,13], which can be viewed as a hand template representation. A hand template could also be learned as an ensemble of edges [14,15] from a set of 2*D* projections of a 3*D* synthetic hand model. Furthermore, colour information can be used to further create more proposals to improve the detection performance [4]. However, the applications of these methods are limited to a small number of hand configurations. They often need to exploit more training data in order to cover a larger configuration space. Alternatively, hands can be detected as part of a human pictorial structure [16], which may bring more context information and allow inferring hand position via optimisation. This is a common practice for still images, but it usually requires at least the upper body being visible for the inference of human layout.

When it comes to videos, motion-based methods can be used for *ad-hoc* applications, such as activity analysis and gesture recognition. They segment foreground hands from background by motion and appearance cues [17–19]. Hands can usually be tracked easily and they do not require a strong appearance model in most cases. Nevertheless, motion-based methods are often not suitable for moving cameras which produce images with lots of background motion.

Recently, ego-centric cameras, such as Google Glass and GoPro cameras, have become more and more popular. A local-appearance-based pixel labelling method recently proposed by Li and Kitani [5] has shown to be quite successful in dealing with dynamic background and varying appearance of hands in ego-centric videos. However, their method only predicts the label of every pixel independently without considering any shape constraint. To deal with the noisy output, segmentation is required to optimise the shape of hand region [20,21].

Often hand detection is only the first step in hand gesture analysis. It is of great interest to further recognise the hand parts in detail. Following the great success of image labelling for human pose estimation [22], hand part labelling becomes one of the most investigated fields, especially for depth images. In this line of works [23–26], recognizing hand parts are considered as an intermediate step for subsequent articulated hand joints estimation. As it is easy to synthesise hand depth images using graphics techniques, there is a lot of priori knowledge, e.g., hand joint position, hand orientation, that can be used to customise the construction of a per-pixel random forest classifier. Such information, however, is not available in conventional colour images. This makes hand part labelling in colour images not well investigated. On the other hand, the progress of semantic labelling [27,28] enriches us with more possible ways to exploit per-pixel labels for prediction. This encourages us to fill the blank of hand parts labelling in colour images.

2. Random decision forests for hand detection

In this section, we begin with a review of random decision forests for pixel-level hand detection, and introduce some notations used in pixel-level hand detection settings.

Given an image patch $\mathbf{I}_\mathbf{p} \in \mathbb{R}^{w\times w\times 3}$ with a size of $w \times w$ centred at pixel $\mathbf{p} \in \mathbb{Z}^2$ in a colour image $\mathbf{I}$, a feature vector $\mathbf{x}_\mathbf{p} \in \mathcal{X}$ is extracted to encode the colour, gradient and texture information of this patch. A binary decision tree $f_\Theta(\mathbf{x}_\mathbf{p})$, parameterised by Θ, is a tree-structured classifier that maps $\mathbf{x}_\mathbf{p}$ to a binary label $y_\mathbf{p} \in \{0, 1\}$, which indicates whether the pixel $\mathbf{p}$ belongs to a hand (i.e., $y_\mathbf{p} = 1$) or not (i.e., $y_\mathbf{p} = 0$). The feature sample $\mathbf{x}_\mathbf{p}$ is recursively branched left or right down through the tree. In Node j, this process is done according to a split function with parameter $\boldsymbol{\theta}_j$

$$\Phi(\mathbf{x}_\mathbf{p}, \boldsymbol{\theta}_j) = \begin{cases} 1, & \text{if } \boldsymbol{\theta}_j^\top [\mathbf{x}_\mathbf{p}^\top \ \ 1]^\top \le 0 \\ 0, & \text{otherwise} \end{cases}, \tag{1}$$

where 1 means $\mathbf{x}_\mathbf{p}$ belongs to the left child of Node j while 0 means to right. When the sample reaches a leaf node, the posterior distribution $P(y_\mathbf{p})$ stored in that leaf will be associated to the sample for prediction.

A decision forest is an ensemble of T decision trees, each with independent parameters Θ_i. Given the feature sample $\mathbf{x}_\mathbf{p}$, the output of the decision forest $F(\mathbf{x}_\mathbf{p})$ is the final class label $y^*_\mathbf{p}$, which is obtained using an ensemble model of the posterior distributions $P_i(y_\mathbf{p}|\mathbf{x}_\mathbf{p})$ in the leaf node of tree i as,

$$y^*_\mathbf{p} = \arg\max_{y_\mathbf{p}} \frac{1}{T} \sum_{i=1}^{T} P_i(y_\mathbf{p}|\mathbf{x}_\mathbf{p}). \tag{2}$$

2.1. Training decision forests

During the training process, each decision tree is constructed independently from a randomly sampled subset of the training set $S \subseteq \mathcal{X} \times \mathcal{Y}$ in a recursive manner. For a node j with a set of training data $S_j \subset S$, there are several randomly generated candidates of $\boldsymbol{\theta}_j$ and the goal is to find a candidate that maximises the information gain, $\mathbf{G}(\boldsymbol{\theta}_j)$, of the current split test. The information gain is defined as

$$\mathbf{G}(\boldsymbol{\theta}_j) = H(S_j) - \sum_{k\in\{L,R\}} \frac{|S_j^k|}{|S_j|} H(S_j^k), \tag{3}$$