Deep Learning for Fluoroscopic Feature Detection

Seminar Report

Liujiang Yan, Mentor: Robert Grupp

1. Introduction

1.1. Reviewed Papers

- Wei, Shih-En, et al. "Convolutional pose machines." Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2016.
- Xie, Saining, and Zhuowen Tu. "Holistically-nested edge detection." Proceedings of the IEEE international conference on computer vision. 2015.

1.2. Motivation

The ultimate goal of the project is to construct a pipeline for automatically retrieving the projective transformation from intra-operative 2D X-ray image to pre-operative 3D model in real time. One of the feasible method is feature driven and has following steps:

- Find features in both 2D X-ray images and 3D models.
- Perform data association.
- Solve for projective transformation by minimizing reconstruction error.

However, from the steps listed here we shall see that, it is not trivial to have a good representation for both detection and description that works in both 2D X-ray images and 3D-models.

An alternative choice is to use anatomical landmarks well-defined in 3D model and try to detect them in 2D X-Ray images. Through such methods, we explicitly know the data association and it leads to straight forward optimization step in following solving for the transformation. Apparently, it is not a trivial task to recognize the anatomical landmarks in 2D X-Ray images even for experts, therefore it is not possible to perform the whole process in real time manner.

Another choice is to use contours as feature. There are well-experimented registration methods using contours as features for 2D-3D projective transformation use. However, since we are working on X-Ray image, the traditional edge detection like canny edge detector does not perform well in our task.

The goal of this project is to utilize learning based method, especially deep neural networks, for those feature detection tasks. The papers discussed here both perform end to end learning and inference for feature detection, using convolutional layer as basic block. Though the papers use natural images for experiments and evaluation, we shall see that there is little gap to adapt such methods for our tasks.

2. Convolutional Pose Machines

2.1. Introduction

This paper introduces a systematic design that incorporates convolutional networks into the pose machine framework for human pose estimation task, which is to detect a fixed number of key points (joints) of human body in given image., in an end to end manner. As for the task, we shall see that, in order to capture the key points, we need to capture the long-range dependencies in the image. The contribution of this paper consists of several aspects. In order to capture long-range dependencies, the network architecture composes of convolutional networks that keep refining the estimates given the result from previous stage, without the need to perform explicit inference using hand-designed graphical model. The other contribution is that, in order to alleviate the gradient vanishing for deep networks, the objective function is designed to sum up the loss by each stage.

2.2. Methods

Convolutional Pose Machine

The general pose machine method consists of a sequence of multi class predictors that are trained to predict the location of each key point in each stage. In each stage, the predictor predicts the beliefs based on the features extracted from the image as well as the contextual information from preceding stage. The belief represented for each key point is given by a heatmap here, where each pixel stands for the confidence of the key point.

$$b_t^p(u,v) = b_t^p(Y_p = z)$$

In subsequent stages from the first stage, the predictor combines information from both feature extracted and contextual information to predict a refined belief map, given as.

$$g_t(x, b_{t-1}) \longrightarrow \{b_t^p(Y_p = z)\}$$

Therefore, intuitively, in each stage the computed beliefs provide an increasingly refined estimate for the location of each part.

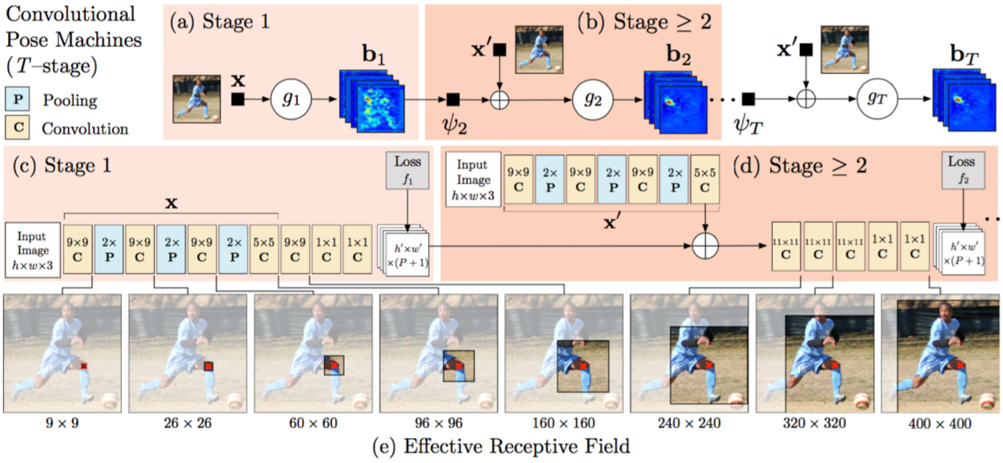

Figure 1. Architecture and receptive fields of Convolutional Pose Machines.

More specifically, the paper discussed here utilizes convolution layer based block for performing the feature extraction and prediction. In such setting, the pose machine is completely differentiable that could be jointly trained in all stages and could be trained globally.

The kernel size of each convolution layer determines the size of local receptive field that this layer could capture. Generally, through composing multi convolution layers we are able to enlarge the receptive field. Therefore, with stages going deeper, the network could capture local to global features and utilize them for refining the estimates, especially when there might be ambiguity given only local information. Further experiments are also performed and show that, accuracy improves as the effective receptive field increases and will saturate.

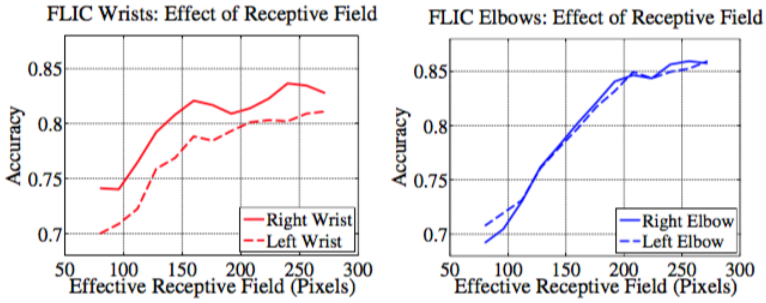

Figure 2 Large receptive fields for spatial context.

Learning Scheme

The pose machine with convolution layer as basic blocks forms a deep architecture that have a large number of layers, and therefore prone to the problem of vanishing gradients, that the magnitude of back propagated gradients decreases in strength from output layer to input layer.

There exist several techniques that could alleviate the vanishing gradients problem. The sequential prediction framework of pose machines provides the possibility to utilize intermediate supervision that we could define the loss at the output of each stage to minimize the distance measure between the predicted belief map and the stage-wise target belief map. Mathematically, the cost function for each stage is given as,

$$f_t = \sum_{p}^{P+1} \sum_{z} ||b_t^p(z) - b_*^p(z)||_2^2$$

The overall objective for the network is given as,

$$F = \sum_{t}^{T} f_t$$

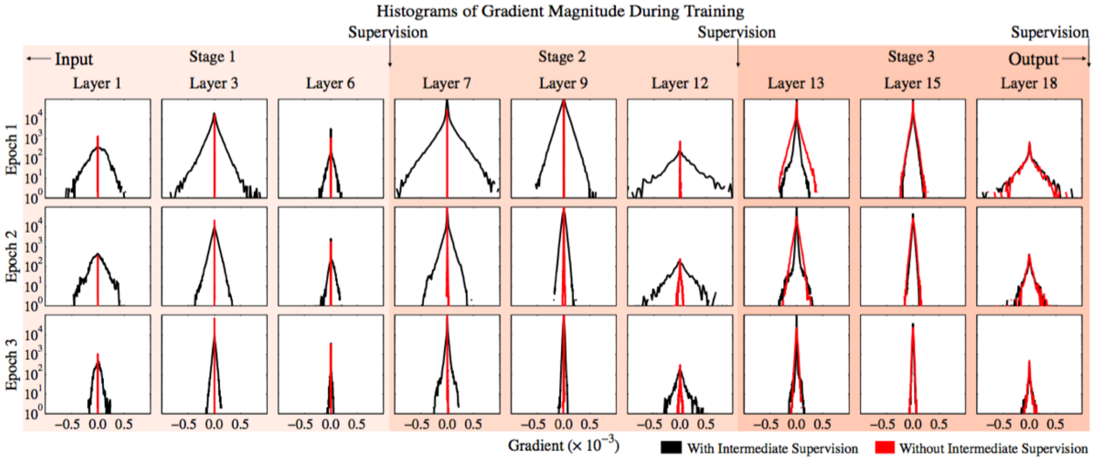

Figure 3 Intermediate supervision addresses vanishing gradients.

2.3. Results and Conclusion

The proposed method is evaluated in several human pose datasets including MPII, LSP, and FLIC datasets. The method achieves state of the art performance and improves the accuracy the accuracy in all parts, over all precisions, across all view angles.

In summary, the convolutional pose machines provide an end to end network architecture for structured prediction problems in computer vision, without the need to utilize graphical model for capturing long range dependencies, which could be instead implicitly done by composing deep convolution layers.

2.4. Application and Discussion

The task of human pose detection is a key point regression problem, which is similar to our landmark detection task, and therefore we could utilize such architecture as a start. Furthermore, the landmarks detection task is simpler since the viewpoint has less variance and the transformation among landmarks are nearly rigid. Also, the belief map introduces uncertainty measure for the corresponding key point. Though such uncertainty measure has little discussion in the paper for the human pose detection task, in our circumstance it could help threshold poor predictions for following optimization problem in our task.

3. Contour Detection

3.1. Introduction

The second paper is about edge detection. Traditional edge detection method is performed locally and single scale, and the feature involved might not be trivial to adapt to other types of images other than natural images. This paper utilizes deep neural networks and is able to perform: (1) holistic image training and prediction and (2) multi scale and multi level feature learning. The network is called holistically nested networks, and resembles deeply supervised net for feature extraction, which previously works well in other visual tasks like image classification.

3.2. Methods

We denote the training data set by S = {X, Y}, where X the input image and Y the binary edge image. The basic network is the VGG net originally used for image classification task, and here the paper uses for feature extraction. VGG net has multi-stage, with representation captured locally to globally determined by the size of receptive field.

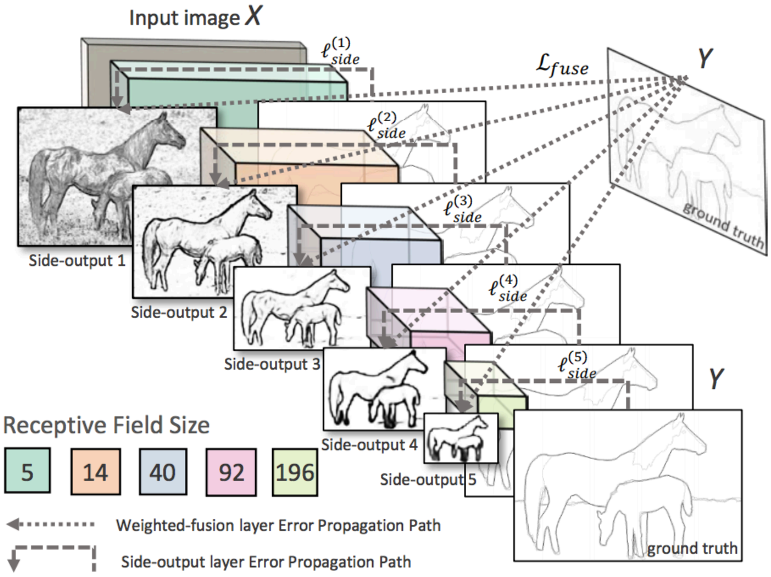

Figure 4 Illustration of HED architecture for edge detection.

Then, in each stage (multiple convolution layer and pooling layer) the network attach a classifier to predict a side output for edge belief map and compute the loss accordingly. Since edge detection is essentially a pixel-wise classification problem, here the paper uses modified binary cross entropy loss, taking the difference of distributions of edge and non-edge into account.

$$L = -\beta \sum_{Y_+} logP(y = 1|X; w) - (1 - \beta) \sum_{Y_-} logP(y = 0|X; w)$$

Each side output from different stage could rescale to original image size through upsampling (bilinear interpolation), and represents result from different scale. In order to combine all the side output results, the network utilizes a parameterized convolution layer to fuse all side outputs. Similarly, as the Convolutional Pose Machine purposes, the overall subjective here also takes loss from each stage into account such that alleviates the gradient vanishing problem.

The prediction phase is straight forward in human pose detection to take the pixel location with maximum value. Here in edge detection, there are some post processing needed to be done to have the final prediction result.

3.3. Results and Conclusion

The performance of HED architecture is evaluated on the Berkeley Segmentation Dataset and Benchmark, and NYU Depth dataset. The paper also discusses several variants and implementation details that might improve the performance. General data augmentation methods as random rotation, random crop, and flip are also introduced to improve the generalization. In summary, the holistically nested networks based edge detection pipeline demonstrates state of the art performance on natural images. Since the basic block is VGG net, the network could start by pretrained model. By implicitly modeling long range dependencies by deep network, it does not need explicit contextual modeling using graphical model.

3.4. Application

Edges and contours are also features that could be used for 2D-3D registration. The comparative advantages of using edges and contours as feature is that they are wide spread in the images and alleviate the effort needed to annotate landmark in 3D model. Also, the network itself has great flexibility. For example, the basic block could be replaced by more powerful architecture as ResNet.

4. Conclusion

Generally speaking, deep neural networks in visual tasks discussed above aims to extract representation within multi-scale, determined by the corresponding size of receptive field from the convolution operation. Then, in order to capture the response in different scale, the methods discussed above both utilize intermediate supervision, and sum up the loss from mutli-stage. In order to adapt the stage results from the deep network, some other post-processing should be performed. As for our landmark detection project, the abstract task is similar though we shall make some revises on the architecture and also carry out the specific validation criteria for our own performance evaluation. As a conclusion, these two papers reach the state of art performance on pose estimation and edge detection, and the architectures are good starts for our own pipeline development.