A Thesis for Master Degree

A Multi-stage Convolution Machine with Scaling and Dilation for Human Pose Estimation

사람 자세 추정을 위한 스케일링 및 확장 기반 다단 콘볼루션 머신

February 22, 2018

Graduate School of Chonbuk National University

Department of Electronic Engineering

Nie Yali

A Multi-stage Convolution Machine with Scaling and Dilation for Human Pose Estimation

사람 자세 추정을 위한 스케일링 및 확장 기반 다단 콘볼루션 머신

2018 년 02 월 22 일

전북 대학교 대학원

전자.정보공학부

섭 아 리

A Multi-stage Convolution Machine with Scaling and Dilation for Human Pose Estimation

지도교수 **박 동 선**

이 논문을 공학 석사 학위논문으로 제출함

2017년 11월 06일

전북 대학교 대학원

전자.정보공학부

Nie Yali

Nie Yali 의 석사학위논문을 인준함

위원장	전북대학교	교수	송상섭	(인)
부위원장	목포대학교	교수	윤 숙	(인)
위원	전북대학교	교수	박동선	(인)

2017 년 **12**월 **15**일

전 북 대 학 교 대 학 원

Dedication

I would like to dedicate this thesis to my family and friends.

-Author

Table of Contents

List of Figures

List of Tables

List of Acronyms

Acronym	Definition
CNN	Convolutional Neural Network
CPM	Convolutional Pose Machines
DPN	Dual Path Network
FC	Fully Connected
GAN	Generative Adversarial Networks
HAR	Human Action Recognition
HCI	Human-computer interaction
HOG	Histogram of Oriented Gradients
HPE	Human Pose Estimation
ILSVRC	Large-Scale Visual Recognition Challenge
PS	Pictorial Structure
PCK	Probability of Correct Keypoint
PCP	Probability of Correct Part
PoseNet	Pose Network
ReLu	Rectified Linear Units
RGB	Red, Green, Blue
SENets	Squeeze-and-Excitation Networks
SGD	Stochastic Gradient Descent
SIFT	Scale Invariant Feature Transform
SVM	Support Vector Machine

A Multi-stage Convolution Machine with Scaling and Dilation for Human Pose Estimation

Nie Yali

Electronic Engineering

Graduate School

Chonbuk National University

Abstract

Human pose estimation (HPE) is a challenging research problem in computer vision, since its problems may include confounding background clutter, the diversity of appearances and changes in scene illumination. HPE has benefits in various applications ranging from biomechanics studies, to physical therapy and exoskeleton control. Powerful CNNs architectures have been developed to capture the important evidences and cues of human pose.

In this thesis, we propose an end-to-end trainable CNN architecture that contains one pre-stage and another four stages with a modified multi-layer convolutional network architecture, for estimating human poses from still images. The main contributions of this thesis are as follows. First, we use

five gates for weighted control of feature maps; one big gate to control the global information and a small gate for each stage to control local information. Second, we use the concept of a fire module in the squeeze net instead of using a single convolution layer to reduce the number of parameters in our model. And skip connections are used in the module to maintain the information and to integrate global and local context concurrently so that features at each resolution can be better preserved. Third, we take pyramid dilated convolutions to learn multi-scale representation for each body part. Dilated convolution is able to control the resolution at which feature responses are computed from the convolutional network without requiring learning extra parameters. It can also allow us enlarge the field of view effectively. Here we use parallel dilated convolution layers with different dilation factors to capture multi-scale information from feature maps. Combining these technologies, we make the receptive field large enough for learning the long-range spatial relationships. Also in the multi-stage network, intermediate supervision is used to produce intermediate confidence maps and refine them through different stages.

Our work shows that it is effective to resample features at different dilation with scaling for accurately and efficiently meeting the structural complexity of limb parts for HPE. We conduct experiments on two benchmark pose datasets, LSP and MPII datasets and demonstrate improved performance compared to other methods based on CNN architecture.

Keywords: Human pose estimation, CNN, Multi-stage, Squeeze, Gate (scaling), Pyramid dilated convolution.

Chapter 1 Introduction

To understand human pose is a long standing requirement with a variety of applications, such as action recognition, human tracking. Despite human pose estimation (HPE) has been studies extensively during the last decade, it still remains one of very challenging tasks in computer vision.

1.1 Definition of Human Pose Estimation

Human pose estimation in computer vision is to determine approximate locations of keypoints (joints) of persons in images, which is still a hard task since some parts of a human body are often strongly articulated and not visible.

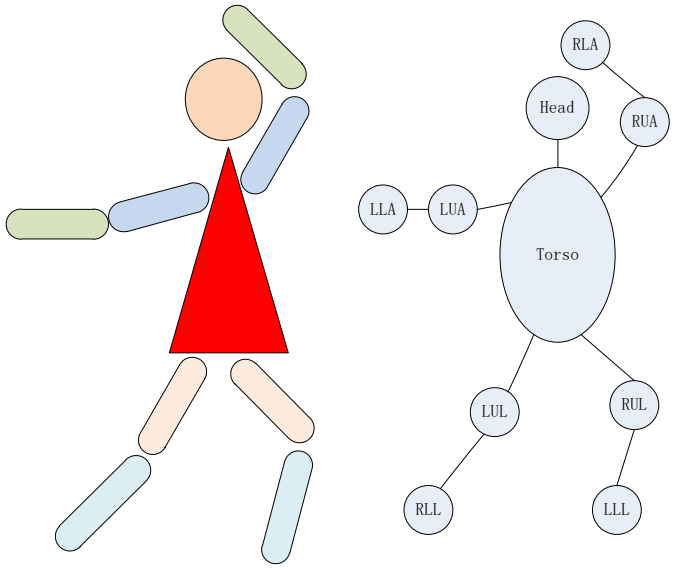

Figure 1.1 Tree-structured relation graph.

As shown in Figure 1.1, human pose can be defined as a tree structure graph, where each node denotes the position of each part, and the edges denote the pairwise spatial relationships. Similarly,

Figure 1.2 shows an input image (left) and its output (right) of a pose estimation system.

Figure 1.2 Articulated body pose estimation.

1.2 Classic Applications of Human Pose

Human pose estimation serves as a foundation for other computer vision tasks such as clothing parsing, human re-identification, movie making, human-computer interaction (HCI), gesture recognition, activity understanding and human tracking, etc.

Fashion is primarily a visual art form. In order to achieve this goal it is necessary to develop a way to interpret the style within images of clothing. The impact of fashion and clothing is tremendous day by day in our society. Being able to automatically parse clothing is one key to conduct large-sociological studies related to family urban groups and familiar income. Pose estimation can help to predict main keypoints such as head, wrists and knees, etc. Then it will exploit these joint locations to bias the clothing labeling in a plausible way. Figure 1.3 is one example of clothing parsing. Actually, clothing a person wears is one semantic segmentation task, which can apply 2D body pose to reduce ambiguities. [38], [39] took pose into clothing parsing work and got good results.

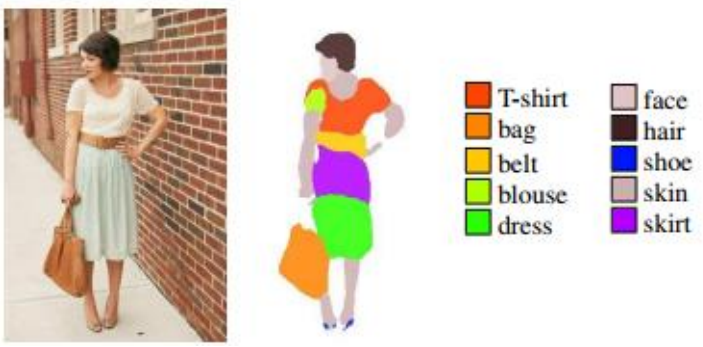

Figure 1.3 Visualization of clothing parsing [39].

Recognizing the same human as he or she moves through a camera is an import problem in security and surveillance system. Once a target has been identified in a camera, we will learn the appearance of the target and want our program can recognize him or her by other cameras. Some researchers applied pose estimation to human re-identification systems [40], [41], [42].

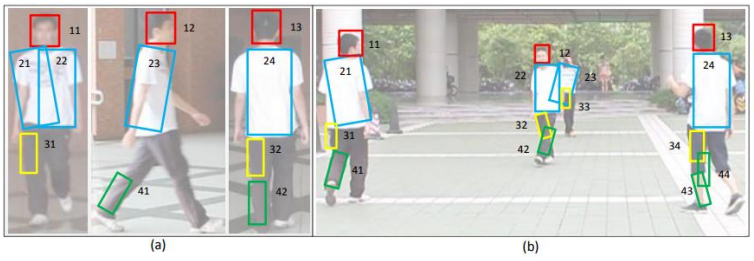

Figure 1.4 A re-identification example used to illustrate [42] inference algorithm. Given (a) reference images and (b) a scene shot, proposals of four parts: head, torso, left thigh, left calf are drawn and numbered in the image. Note that here omits the other parts and only keep few proposals for clear specification.

They built a model for human appearance with pose priors and made matching to identify more robust. With referenced part images representing the body as an articulated assembly of compositional and alternative parts, they gathered person-specific features learned

over the course of tracking to improve the performance. Figure 1.4 shows an example.

Human pose estimation is also well-known use for capturing actor performance in movie making. Motion capture is the process of recording human movement through special cameras, and mapping them onto a character model. In order to make them look realistic, actors will be captured and drive the movement of the artificial character. Utilize sensors placed on the body to measure the low frequency magnetic field generated by a transmitter source. Thus the sensors can report position and rotational information. Figure 1.5 shows Gollum which is a fictional character in the action movie T*he Lord of the Rings*, the actor takes sensors in his body to capture the pose information. We also can see 3D computer animations and video games by electronic arts, Gremlin, Square, Konami.

Figure 1.5 Gollum in *The Lord of the Rings*.

HCI is the study of interaction between people (users) and computers which is based on computer vision and aims to improve the interaction with real-time systems and marker-less systems. Vision-based human-computer interaction systems are the way of the future because the current advances in computing technology is

pushing the application of HCI. For example, the touch screen phone is everywhere in our life. Paper [43] detected the head and simultaneously estimate the pose to make an automatic HCI system. Figure 1.6 is one picture from Carnegie Mellon University's Robotics Institute, which have enabled a computer to understand body pose and movements of multiple people from video in a real time. Usually, HCI is based on gesture and realizes the robust control of mouse and keyboard events. Researcher Pei [44] design a real-time HCI system based on hand gesture.

Figure 1.6 Computer reads body language.

Activity is a collection of actions and or interaction that compound to describe a high level event, e.g. 'tidy room' and 'eat an apple'. Each interaction and action can be thought of a sub-activity event in computer vision. Human actions [45] can be used to improve pose estimation with high-level information about activities to incorporate higher-order part dependencies. Pose-base methods are mostly influenced by viewpoint of the person, [46] build one model with holistic and pose based features to make activity recognition. From Figure 1.7 we can see that pose estimation is the basic knowledge to

understand gesture, action, interaction and activity.

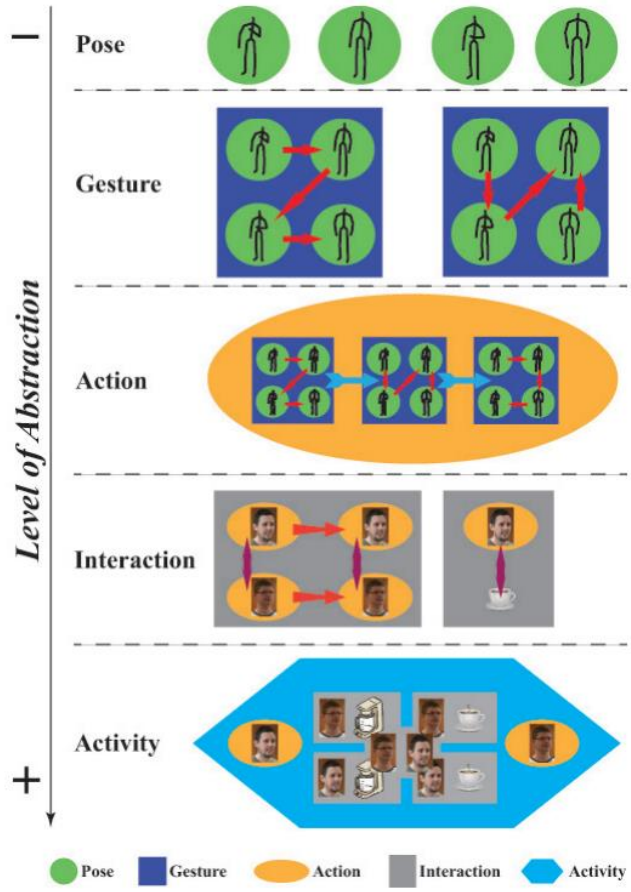

Figure 1.7 Levels of abstraction within human action recognition [47].

It is important to understand the human body language so that human pose is still a popular topic in videos with people. There is a strong evidence that representation based on human pose are highly effective for a variety of tasks in video. Researchers bring video-based body pose and articulated tracking together working on visual human analysis. From single person to multi-person, it becomes more and more complexity with HPE. Umar. etc. [48] proposed a model that jointly multi-person pose estimation and tracking in a single

formulation. They also built a new challenging "Multi-Person PoseTrack" dataset which provides detailed and dense annotations for multiple persons in every video. As shown in Figure 1.8, each color corresponds to a unique person identity and we can see the estimated poses for all persons.

Figure 1.8 Estimated poses for all persons in a video [48].

What's more, pose estimation can help to do gait analysis and rehabilitation in medicine. Pose estimation can be also applied to sport medicine, like injury prevention, performance analyses and performance enhancement. Especially, a pose estimation is also used to correct athlete's posture. As the development of AI, pose estimation also opens the door to design ergonomic products.

1.3 Challenges in Human Pose Estimation

Despite many years of research with significant process made recently, Estimating human pose from images is still a very challenge task. HPE is an important computer vision problem as well as plays critical role in a variety of real-world applications. An effective human pose estimation system requires us to map the input images with large variations into multiple body keypoints which must satisfy a set of geometric constrains and interdependence.

There are five main obstacles to be solved: cluttered background, motion blur, occlusion and self-occlusion, varying illumination and foreshortening.

Background clutter. The presence of background clutter is a common problem in object detection. It is likely that the background

of an image contains very unbalanced and complex objects. These background objects tend to divert from foreground objects that can lead to false positive results. High complexity can make HPE a really tough task for human eye. As shown in the top row of Figure 1.9 (a), these two images contain many people in the background.

Motion blur. Fast moving parts can invalidate smoothness constraints, which always happens in videos because of relatively long exposure time on cameras. Even a high per second frame extraction is used, there is still a high degree to correspond dislocation parts. Figure 1.9 (b) shows the limb appearances of the athletes and a boy are in motion-blurred and no clearly solid region can be seen.

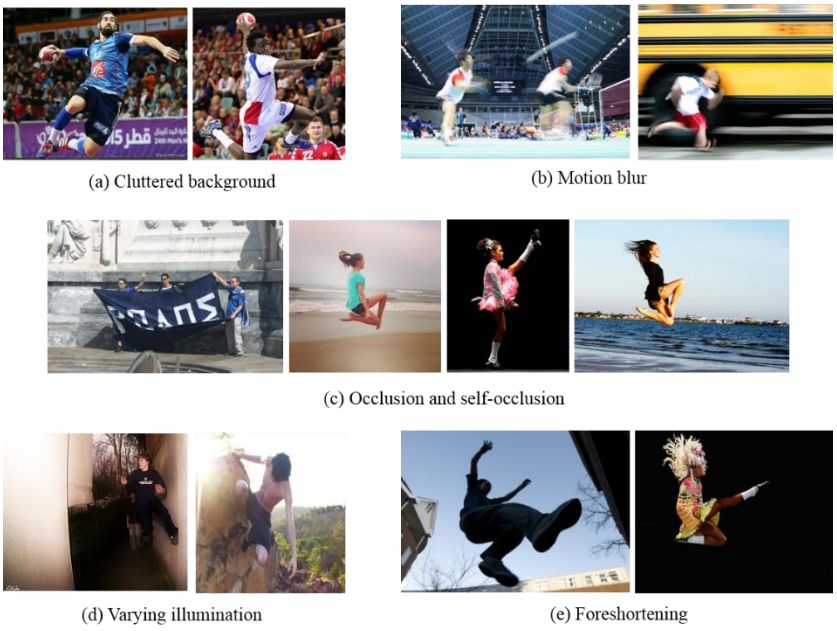

Figure 1.9 Various real-life challenges impeding in HPE and make difficult to see human body part clearly.

Occlusion and self-occlusion. There is one of the most common problems is occlusion and self-occlusion of objects in natural senses.

Generally, a person gets included by other people in groups or an element of the surroundings. The localizations of parts are not visible with such full occlusion. Above mentioned occlusions have occurred from not the human-in-consideration. Self-occlusion happens due to different viewpoints of objects. Such phenomenon is observed in Figure 1.9 (c), the first image shows some parts of human are shaded by banner and rest three images show self-occlusion.

Varying illumination. Illumination can be manipulated to make the person-of-interest with different lighting conditions. The overall brightness of the same scene may be different due to the existence of shadows. What's more, cameras tend to lead under-exposure or over-exposure problems with automatically adjust exposure for the light of environment. The bottom row of Figure 1.9 (d) present unclear visibility of complete human because of varying illumination conditions highlighting the background and the foreground.

Foreshortening. Arm and legs foreshortening can happen because human bodies can take on a large variety of possible pose. When human body part plane has a higher angle or not in parallel, the pose can be extremely complex. The child part elusive with respect to the parent part is made by this difficulty. Figure 1.9 (e) shows feet and arms foreshortening when a person is jumping.

1.4 Outlines

In this thesis, the main goal is to propose a novel HPE system for single still image. In particular, this system offers improvement in its computational complexity and accuracy, compared to previous researches. Specifically, the main contributions of this thesis are briefly summarized in the following.

Our major contributions of this thesis can be divided into three

parts.

(1) We control the global information and local information by using gates for weighted control of feature maps through multi-stages.

(2) We replace a simple convolution layer with a fire module introduced in squeeze net to reduce the big number of parameters. And skip connections are deployed in the module to maintain the global and local context information.

(3) We take pyramid dilated convolution to learn multi-scale representation of each image. Our parallel dilated convolution layer with different dilation factors can capture multi-scale information from feature maps.

To demonstrate the details of our work, the remainder of this thesis is organized as follows. We start by introducing existing approaches on HPE in Chapter 2. First, the traditional pose estimation methods are presented. Then, a focus is taken toward more popular methods built upon deep learning. Finally, the very recent classic approaches based on CNNs for HPE are discussed. Our networks are proposed in Chapter 3. Experiments and evaluations are demonstrated in Chapter 4. Finally, we summarize the system of HPE built by us and discuss its superiority over other models. We also look into future works in the area of HPE in Chapter 5.

Chapter 2 Existing Approaches of Pose Estimation

In this chapter, an overview of the available literatures in the field of HPE is provided.

Looking back at past work on HPE, a variety of models with efficient inference have been proposed. Generally, there are two basic methods to estimate pose: traditional approaches and deep learning for pose estimation. These methods are discussed in the following sections.

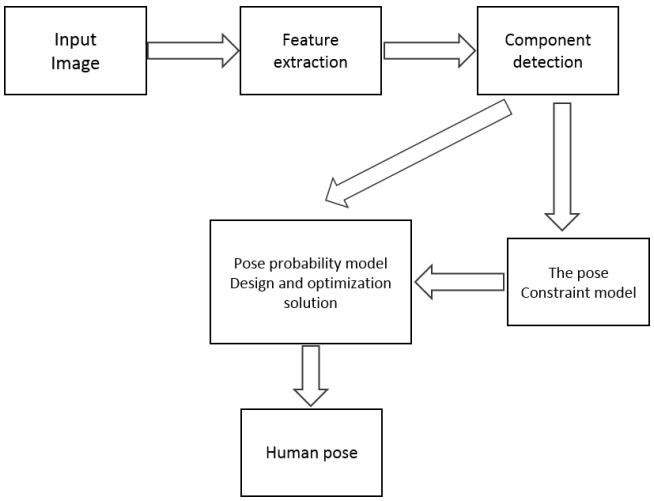

Figure 2.1 General pose estimation process.

Including the performing estimation methods based on deformable part models, the best ones are typically based on body part detectors which commonly consist of multiple staged of processing as shown in Figure 2.1. Processing in a typical pipeline, the first stage consists of extracting sets of low-level characteristics such as SIFT [54], HOG [53], or other filters describing orientation statistics in local image patches. And in the next stage, in order to reduce the size of the

representation and develop local shift scale invariance as well, there features are pooled over local spatial regions and across multiple scales sometimes. In the final stage, the aggregate features are mapped to a vector, and then the vector is input to a standard classifier such as a support vector machine (SVM) or the next stage of processing(e.g. assembling the parts into a whole). When remaining invariant to the various nuisance factors (lighting, viewpoint, scale, etc.), much work is devoted to engineering the system producing a vector representation that is sensitive to class (e.g. head, hands, torso).

2.1 Traditional Approaches with Handcrafter-feature

The traditional approaches for HPE are based on handcrafter feature. In early time, Marr and H.K.Nishihara [49] took the information about head, torso, arm and leg to shape with a 3D model in 1978. They also made a 3-D model's axes specified in terms of pairwise relations with location, rotation and scale as shown in Figure 2.2.

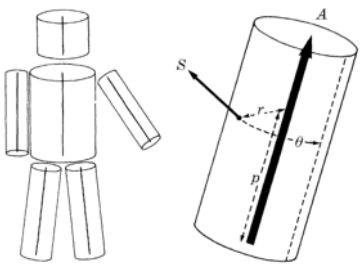

Figure 2.2 The organization of person shape information and pairwise relations [49].

However, Pictorial Structure (PS) is most classic approach in traditional approaches. In 1973, Fischler and Elshlager [50] devised the concept of pictorial structures for objects that are consist of

different prominent sub-parts. There is a human face as an object shown in Figure 2.3 (left) and its sub-parts hair, eye, nose etc. interconnected by springs to maintain a feasible configuration. The right side shows different parts of one human's body are connected with springs.

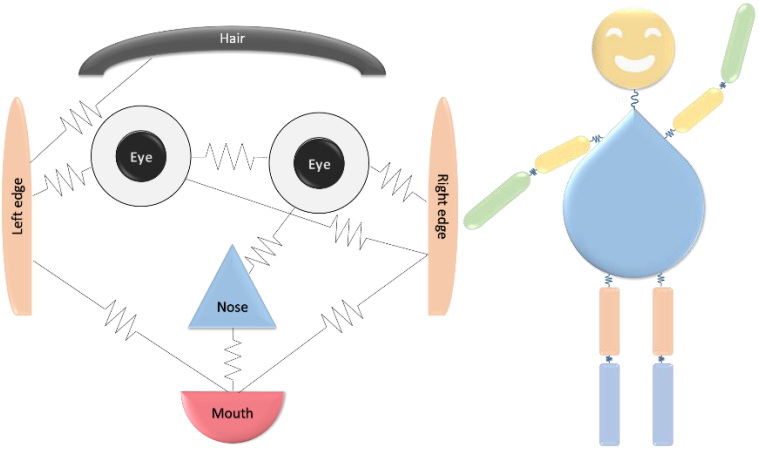

Figure 2.3 Pictorial structure for face (left) and human (right) with springs.

2.1.1 Pictorial Structure based HPE

In pictorial structure, model is represent by a graph $G=(V,E)$ where $V=\{v_1,v_2,\ldots,v_n\}$ are the parts and $(v_i,v_j)\in E$ indicates a connection between parts. The optimal location for object is given by $L^*=(l_1^*,\ldots,l_n^*)$. The function (1) is shown in detail.

$$L^*=\arg\min_L\left(\sum_{i=1}^{n}m_i(l_i)+\sum_{(v_i,v_j)\in E}d_{ij}(l_i,l_j)\right) \qquad (2.1)$$

$m_i(l_i)$ is the cost of the placing part i at location l_i and $d_{ij}(l_i,l_j)$ is a deformation cost. Here n parts and h locations gives h^n

configurations. If graph is a tree, it can use dynamic programming. As shown in Figure 2.4, it is one tree model for HPE and it illustrates how the output distribution are produced by decision tree from input depth map and pixel coordinate. Here they [21] used a random forest with cascade approach working on HPE.

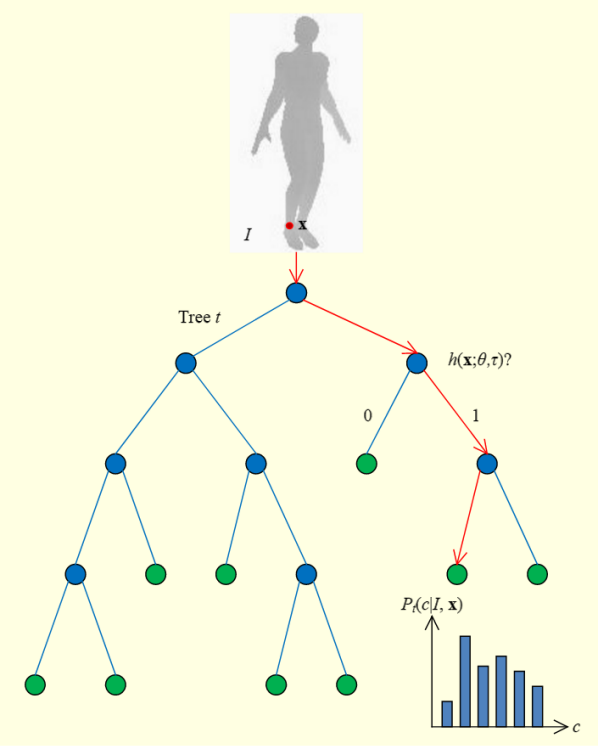

Figure 2.4 Tree model for human pose [21].

However, it is still quite difficult to capture the whole range of appearances with pictorial structures. There is a big problem that even projections of a simple cylinder into 2D yields many different appearances.

2.1.2 Mixtures of Parts based HPE

In paper [51], they did many miniature part model that using multiple parts connected with springs to appear a single limb, such as Figure 2.5. The lower leg of the panda could be modeled with two parts and the torso was modeled with four parts. Small parts were

connected with springs that incorporate flexibility in their connections, shown in Figure 2.6.

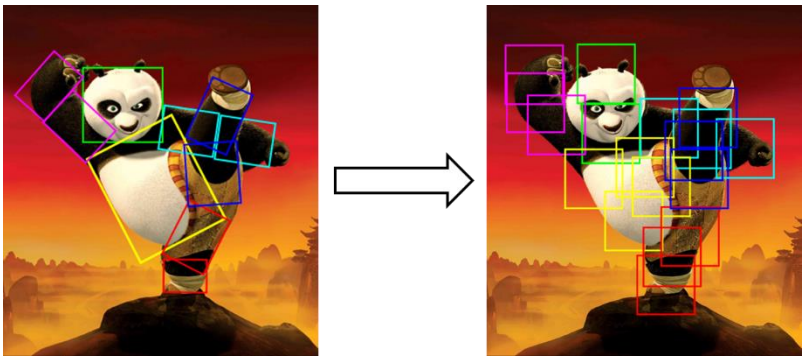

Figure 2.5 Example of mini part model.

For a graph $G=(V,E)$ given a test image I , this mixtures of parts model full score equation written as:

$$S(I,L,M)=\sum_{i\in V}\alpha_i^{m_i}\cdot\phi(I,l_i)+\sum_{ij\in E}\beta_{ij}^{m_i m_j}\cdot\psi(l_i,l_j)+S(M) \qquad (2.2)$$

Parameters m_i is mixture of part i and $\alpha_i^{m_i}$ is unary template for part i with mixture m_i . $\beta_{ij}^{m_i m_j}$ is pairwise springs between part i with mixture m_i and part j with mixture m_j . It encodes both the rest position and rigidity of the spring. $\phi(I,l_i)$ is the local image features at location l_i . $\psi(l_i,l_j)$ is spatial features between l_i and l_j . Here $S(M)$ is a co-occurrence bias. This equation can be seen as three terms. The first term scores a local match for placing a part template at a particular image location. The second term consists of a deformation model that evaluates the relations of pairs of parts. The third term scores the combination of itself.

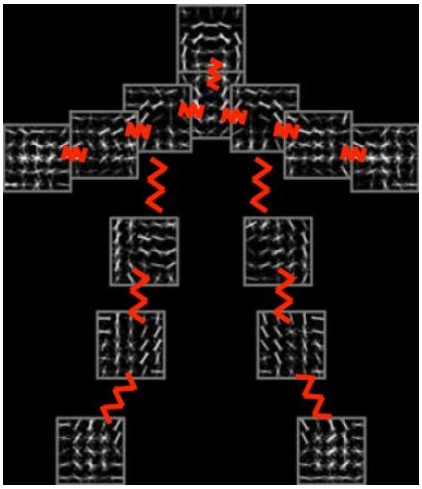

Figure 2.6 Visualization of a flexible mixture parts model connected with springs in one person.

This model shared computation among a large number of oriented and foreshortened limbs with a small number of local templates. This method made inference faster and learn easier due to fewer parameters.

2.1.3 Multimodal Decomposable Models based HPE

Sapp and Taskar [52] captured multi-modality at a higher granular level of half-bodies. They proposed a multimodal, decomposable model for HPE and this model was one pictorial structure model which could capture the wide range of appearance present in monocular images. They called it as MODEC pose model, which was a multimodal decomposable model via clustering human body joint configurations in a normalized image-coordinate space. Every model was corresponds to a discriminative structure linear model as shown in Figure 2.7 (Left). In this figure (right) was the MODEC pose model with some model average images which were representative of different modes. They used square Euclidean distance measure to obtain modes. This model aimed to train parameters that helped in

identifying the closest matching mode and locating the body keypoints with k-means clustering.

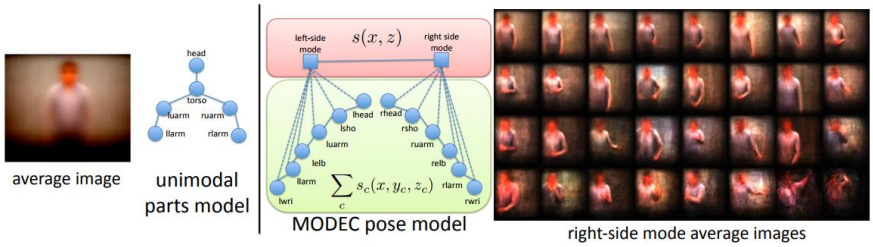

Figure 2.7 Left: One linear model. Right: The MODEC pose model proposed by [52].

This method is different from a static pictorial structures model such that a mode represents a part of the body and each of them handles the configuration parameters locally. However, the size of tree in the model is dynamic.

2.2 DeepNet Architecture based Approaches

The research has headed from traditional methods [49-54] into convolutional neural networks (ConvNets, CNNs) [1, 11-19, 22-24, 26-28] in recent years. Because of this, significant improvements in accuracy have been achieved. Efficient ConvNet architectures command largely computation and much memory in training phase. The range of network parameters quantity is from few millions to hundreds of millions. In basic, Powerful GPUs are mandatory for training which can take few days or weeks for training a network.

In 2014, significant improvements have been accomplished by convolutional Neural Networks (ConvNets), that is, deeppose do with the developments of deep learning and map a novel algorithm ground on a Deep Neural Network (DNN). Of late, Convolutional Pose Machine [40] pooled the spatial correlations inference among body parts within the ConvNets. Most advanced performance is gained by

the stacked hourglass network [29] and its variant [5] using reduplicated pooling down and up sampling process to learn the spatial distribution. As the developing of Generative Adversarial Networks (GANs), this technology is also used to HPE and has achieved outperformance [22-24].

2.2.1 Convolutional Neural Networks and Architecture

Convolutional Neural Networks (CNNs) stand out for their ubiquity of use in Artificial Neural Networks (ANNs) which have been applied to some tools such as automatic translation, search engines or video classification. Here we will make an illustrated explanation of CNNs which includes many layers.

A feature is a pattern obtained from the input images and will be learned by the distinct layers. The edge contours of the objects serve as feature maps in classification and object detection tasks. A filter (or kernel) applied in a sliding window fashion to extract features from the input refers to an operator which can transform the information encoded in the pixels. Kernel size is smaller than the input feature map size, and they will do a convolving which is calculating the dot product in the kernel matrix. Then the dot product values are summed together to be a convolved value. Kernels can be stacked to create high-dimensional representations of the input feature maps. At each operation, a matrix multiply of the kernel and current region of input is calculated and the processing is demonstrated in Figure 2.8. The mathematical function of convolution is implemented in each layer defined as:

$$S(i, j) = \sum_{m} \sum_{n} I(m, n) K(i + m, j + n) \tag{2.3}$$

Where K is the convolution kernel and I is the image. Kernel size is specified by $(m \times n)$ and the parameters of the kernels are shared by

all nodes in a layer.

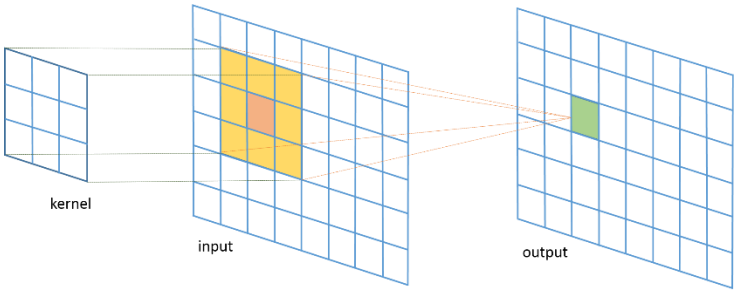

Figure 2.8 Convolution operation using a 2D kernel.

The receptive field in the CNN refers to the part of the image that is visible to one kernel at a time. Each neuron in a hidden layer is only connected to a small number of units in the previous layer. For instance, an output neuron in the hidden layer will only be connected to a small region of the input image. We call this region as receptive field. See Figure 2.9 for a graphical representation of receptive field.

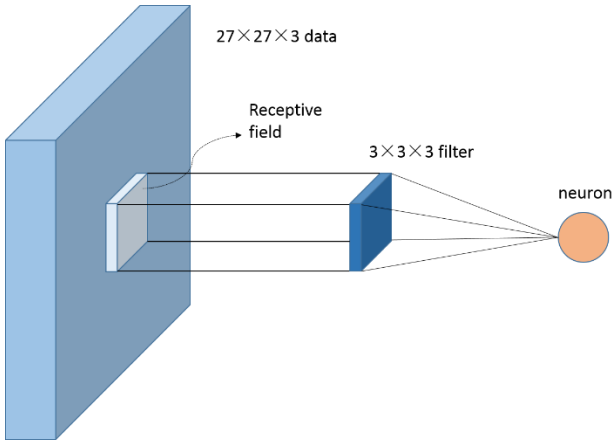

Figure 2.9 An example of receptive field.

Convolution neural networks extract multiple levels of representation from the input images. For example, in Figure 2.10, the first layer recognizes edges, the middle layers will get facial features like a nose or an eye. Until the final layer extracts full faces. It shows that deep learning can learn well with image processing and higher layers in CNNs not only learn higher level features, but also catch more specific to particular object categories.

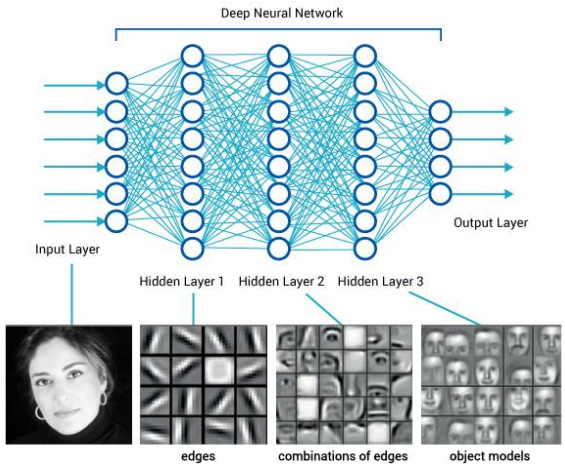

Figure 2.10 Layers and their abstraction in deep learning.

Image recognition as measured by ImageNet classification performance has improved dramatically with the development of deep learning. It was apparent that deep learning would take over computer vision and other methods would not be able to catch up in 2012, corresponding AleNet's [36] rising as demonstrated in Figure 2.11. However, human performance zone is between 5% and 10%.

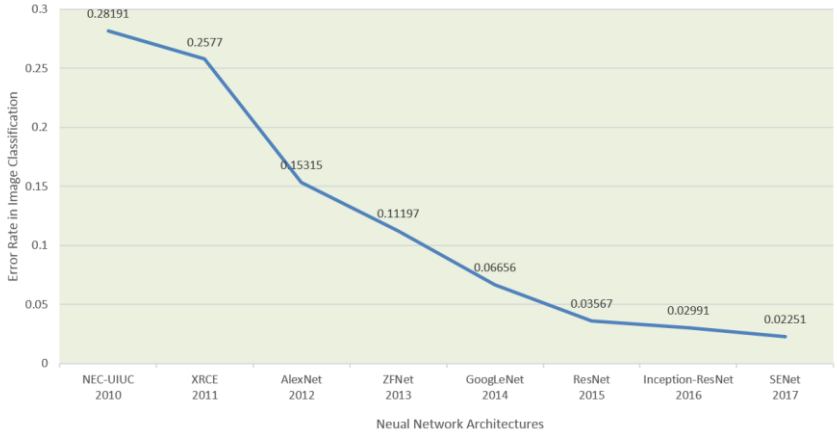

Figure 2.11 The ILSVRC saw an exponential decline in top 5 error rate for CNNs for Image Classification.

ConvNets [2-10, 29-31, 33, 36-37] are very classic frameworks.

The neural network from simple to complexity, becomes more and more powerful. AlexNet (2012) [36] was published by Krizhevsky in 2012. It was made up of five simple convolutional layers, max-pooling layers and three fully connected layers. ReLU was taken as active function to reduce the computational complexity. Dropout technique was used to cut out certain neurons during training selectively to avoid over-fitting in the model.

VGGNet (2014) [37] is a very but simple neural network which has 16 layers. In VGGNet, each convolutional layer has a 3x3 kernel unlike the other networks that have different sized kernels for each layer. VGGNet believes that if you use a large convolution kernel will cause a lot of time to waste, so use the small convolution kernel to reduce the parameters, saving computing overhead.

Inspired by [9], the Bottleneck layer of GoogleNet (2014) reduces the number of features, thereby reducing the complexity of the operation of each layer, so it can speed up the reasoning time. The number of features is reduced by about four times before passing the data to next layer. The architecture deserves frame because of saving a large number of computing costs. The reason for the success is that the input characteristics are relevant so that redundancy can be removed by appropriately with 1x1 convolution. Then after the convolution has a smaller number of features, they can be extended again and act as input on the next layer. GoogleNet works well on the ImageNet dataset. Later there is a series Inception, such as Inception v2 [6], Inception v3 [7] and Inception v4 [8].

ResNet (2015) [5] was published in December 2015, using skipping layers. Skipping two layers can be seen as a small classifier in the network. It's a very fatal architecture, because this architecture achieves more than 1000 layers through this method. Before it, we use only a few layers. The layer first uses 1x1 convolution and then

outputs 1/4 of the original feature, then use a 3x3 convolution kernel followed by a 1x1 convolution kernel again. But the output is the original input size, so it looks like a Bottleneck. The Bottleneck makes a lot of reduction in the amount of calculation, but it remains a wealth of high-dimensional feature information. The residual connections are shown in Figure 2.12.

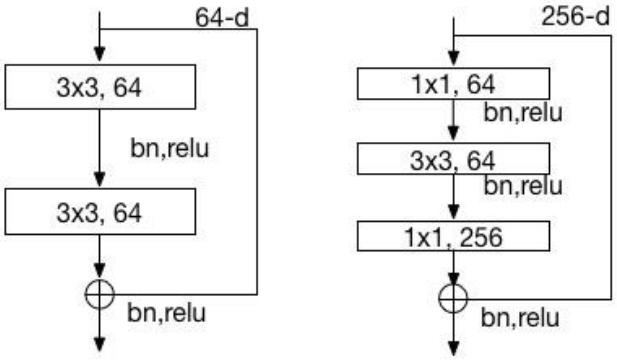

Figure 2.12 ResNet block [5].

As we all know, in the last two years, researchers usually take two directions to improve the performance of convolution neural. One is to make network deeper, such as ResNet, to solve the network when the depth of the gradient disappear. Another one is to make network wider, such as GoogleNet Inception. However DenseNet (2017) [4] starts from the feature. The ultimate use of feature achieves better results with fewer parameters. DenseNet has access to multiple previous layers and employs multiple channels to process the raw input and learns to combine them. In DenseNet, it encourages heavy feature reuse although the number of connection grows quadratically with depth as shown in Figure 2.13. There is no need to learn redundant feature maps again because all layers have direct access to every feature map.

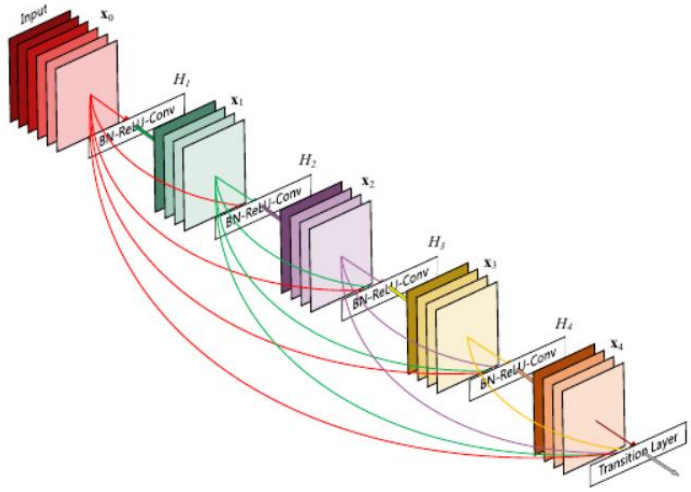

Figure 2.13 DenseNet architecture [4].

2.2.2 Convolutional Pose Machines

Convolutional Pose Machines (CPM) [28] design a sequential architecture that directly operate on belief maps from last stages. Without the need for explicit graphical model, the part locations can get refined estimates increasingly. Multi-stages are also used in the model to produce multi-loss avoiding vanishing gradients during training. This method can be taken as intermediate supervision.

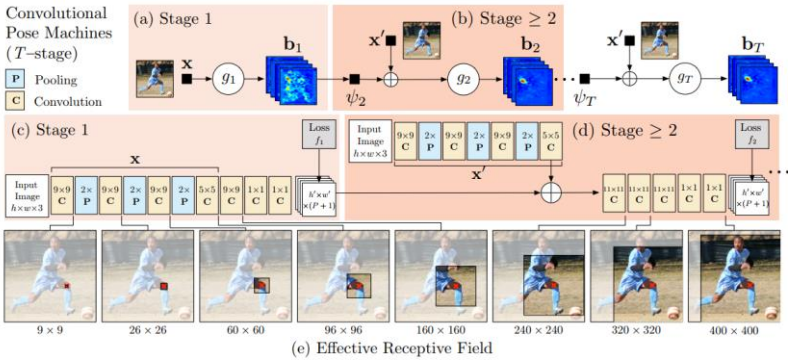

Figure 2.14 Architecture and receptive of CPMs [28].

Figure 2.14 shows the architecture and receptive of CPMs, which

has T stages. The computed belief maps provide an increasingly refined estimate for each part location in every stage.

CPMs can achieve end-to-end joint training in all stages. This network corresponds receptive field for the subsequent stages. They use SGD to jointly train all the T stages in the network.

2.2.3 Stacked Hourglass based PoseNet

Another state-of-the-art convolutional network architecture is stacked hourglass [27] poseNet, which repeats bottom-up and top-down processing used in conjunction with intermediate supervision to improve the performance of HPE. In Figure 2.15, this module allows for repeated bottom-up, top-down inference across scales.

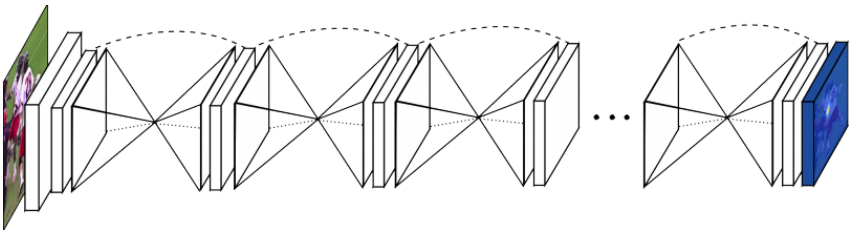

Figure 2.15 Stacked hourglass modules [27].

This network can capture and consolidate information across all scales of the image. Like many convolutional approaches that produce pixel-wise outputs, hourglass poseNet pools down to a very low resolution first, then up samples and combines features across multiple resolutions. Here they expand on a single hourglass by consecutively placing multiple hourglass modules together and make an end-to-end training.

2.2.4 GANs based PoseNet

Recently, Generative Adversarial Networks (GANs) model with deep learning techniques have been particularly popular due to their principle ability to generate sharp images through adversarial loss.

Neural networks have made a great process and GANs have been applied to human pose estimation [22-24]. Adding adversarial training strategy to HPE has brought some benefits to improve the performance of pose estimator.

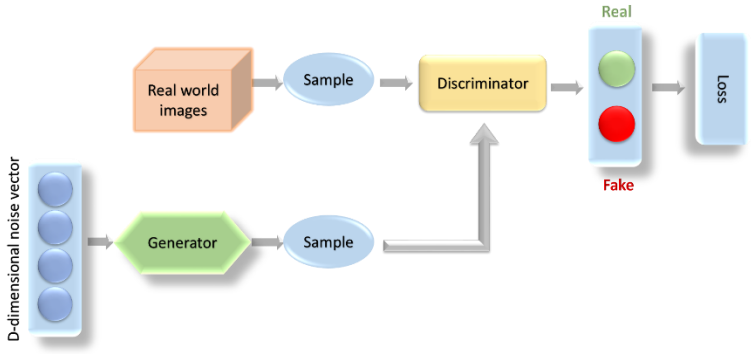

Figure 2.16 Processing of GANs.

The main idea behind discriminative and generative methods is to have two competing neural network models as shown in above Figure 2.16. The generator will take noise as input and generate samples and the discriminator will receive samples from both the training data and generator data. The discriminator should be able to distinguish training between generator data. It is like they play a continuous game that discriminator is learning to distinguish generated data from real data, and the generator is learning to produce more and more realistic data.

As mentioned in Figure 2.17, the model [22] consists of three parts: the pose generator network G, the pose discriminator network P and the confidence discriminator C. C and P are two discriminator networks to correct some low confidence and incorrect location pose estimation here. They use a bottom-up and top-down network as the generative network. This method achieves considerably better results on two popular datasets: extended Leeds Sports Poses (LSP) and MPII Human Pose.

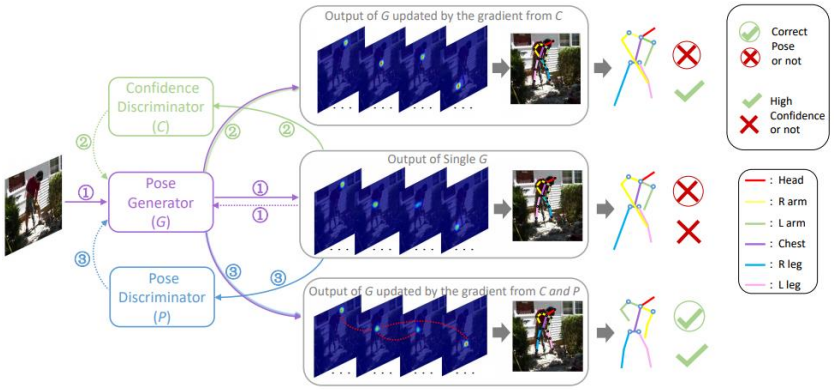

Figure 2.17 Overview of the proposed GANs for HPE [22].

2.3 Bottom-up and Top-down Methods

No matter traditional methods or deep learning methods, HPE can be done into distinct manners. The first approach is bottom-up based on two steps: parts detection anod parts association. The second approach is top-down based on another two steps: person detection and pose estimation.

2.3.1 Bottom-up Pose Estimation

In bottom-up pose estimation approaches, human body parts are detected by pieces of image evidence, then assembled into a human occurrence as shown in Figures 2.18 and 2.19. Generally, the pieces describe different features which predict human pose or localize body parts. To represent a direct human model, a tree-structured model with a number of sticks and joints is used and sticks connect these joints. Pictorial structure (PS) is the most successful technique in bottom-up methods due to the fact that human instances are highly articulated. In the past decades, more and more researchers have focused on this approach [11-16]. There are two disadvantages with traditional models like PS. Firstly, their limited representation power can't handle large variations in pose and appearance, because these tree-structured models are based on traditional hand-crafted features. Secondly, the coupling between pairs of parts with geometric priors is not strong and data-dependent. In order to overcome these difficulties,

some researchers [20] try to design a full relational model with graph of limbs and approximation inference by local greedy search. But this model is very time consuming with hand-crafted features. Thus, it is still limited in handling complex natural images.

Figure 2.18 Parts detection.

Figure 2.19 Parts association.

Recently, with the developing of deep learning, graphical models are combined with deep learning method becoming more powerful representation [13-20, 22-28]. The bottom-up approach has less limit in its application and more robust to rapid movements because it has the advantage that there are no specific model prior needed and no manual initialization required.

2.3.2 Top-down Pose Estimation

A top-down approach uses an object detector to find the person localization in the still image first and then makes a pose estimation from the detected region, as represented in Figures 2.20 and 2.21. It

was used in a multi-person pose estimation [17]. Usually, a top-down method takes a priori human model as the model representing the observed object. Based on this model, human model is continuously updated. For the detector part, nowadays, most researchers like to choose CNNs, which have shown their capacities in classification and detection problems. It is very important that a human model has the ability to handle occlusions by various kinematic constraints. However, when a top-down model is only based on single-view pose estimation, it has the high probability to select wrong pose and suffer from accumulation of errors. These errors can make the pose recovery more difficult.

Figure 2.20 Person detection.

Figure 2.21 Pose estimation.

Chapter 3 A Multi-stage Convolution Machine with Scaling and Dilation Approaches

PS model has a limitation in expressing complex HPE. And the hand-crafted features are sensitive to noise and variations. On the other hand, generally, CNNs are robust to variations and can extract good representations without using hand-crafted low level features. However, their performances heavily depend on their internal factors like their architectures and hyperparameters and their external factors like the size and variety of training data sets. In this thesis, our major contributions towards developing a novel neural network for HPE are collected to improve its performance.

By combining different kinds of state-of-the-art deep learning technologies, we design a new PoseNet with dependent multi-context representation to boot HPE. It is a variant of previous approaches [14, 15, 28]. To start our network, we will introduce some concepts of SqueezeNet, Dilation Convolutions and SENet.

3.1 Related Basic Methods

3.1.1 SqueezeNet

SqueezeNet [29] is a small CNN architecture which achieves AlexNet-level accuracy on ImageNet with 50x fewer parameters. The authors outline three strategies to reduce the parameters drastically.

Firstly, by replacing a group of 3x3 filters with 1x1 filters, this strategy reduces the quantity of parameters 9x. As to a larger 3x3 convolution filter takes spatial information of pixels close to each other. On the other side, 1x1 convolutional filters zero in on a single

pixel and capture relationships amongst its channels as opposite to surrounding pixels. Secondly, to remain 3x3 filters, they reduce the number of inputs. This strategy reduce the number of parameters with fewer filters and feed 1x1 layers into another parallel layers consist of 1x1 and 3x3 convolution filters. In this paper, they call last 1x1 layer squeeze and next layer expand as shown in Figure 3.1. And the authors of this paper call this specific architecture the "fire module". Thirdly, they use downsample late in the network to receive large activation maps. In our network we use "fire module" block for reducing parameters.

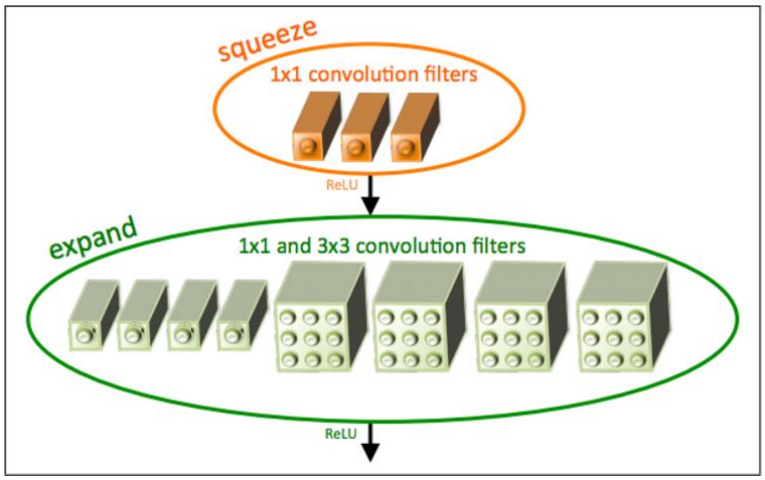

Figure 3.1 Squeeze fire module [29].

3.1.2 Dilated Convolutions

Dilation here just refers to the fact that a certain number input values is skipped when applying the filter of a convolutional layer. [2] build a network out of multiple layers of sparse convolutions, where the dilation factor l increase exponentially at every layer. The effective receptive field of units exponentially with layer depth despite the number of parameters grows linearly. The Figure 3.2 illustrates the process. (a) is a 1-dilated convolution which has a

receptive field of 3x3. In the second (b) 2-dilated convolution, each element has a receptive field of 7x7. A receptive field of 15x15 is produced by (c) which is a 4-dilated convolution. Dilated convolution is applied to keep the output resolutions high and it avoids many parameters.

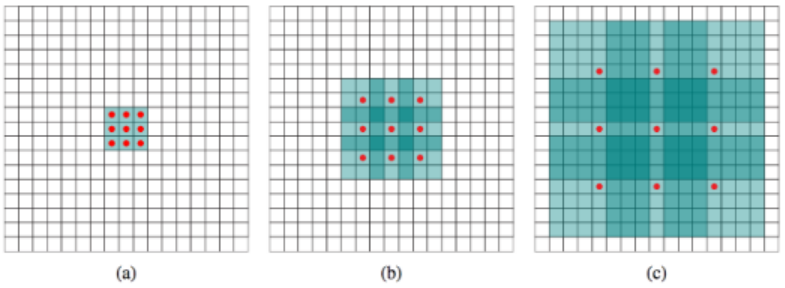

Figure 3.2 Dilated filter [2].

3.1.3 Squeeze-and-Excitation Networks

Squeeze-and-Excitation Networks (SENets) [33] introduces a building block for CNNs which improves channel interdependencies at almost no computational cost. The basic idea is that let activation maps learn a weight vector and give each channel different weightings. They wined the ImageNet competition this year and helped to improve the result by 25% from last year in image classification task.

Figure 3.3 shows a squeeze-and–excitation block which can ensure the network to increase its sensitivity to informative features. Subsequent transformation will exploit these features and less useful features will be suppressed. They use a global average pooling to squeeze global spatial information in to channel descriptors. This step is called Squeeze. Another step is Excitation to adapt recalibration. They use one simple gating mechanism with a sigmoid activation and the final output is obtained by rescaling.

SENet block have made huge performance boost in image classification. They can be also easily added to existing architectures such as Inception module, ResNet module.

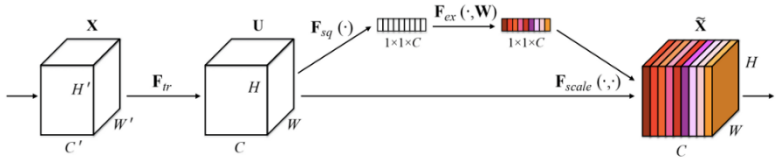

Figure 3.3 SENet block [33].

3.2 Proposed Architecture

Some state-of-the-art human pose estimation methods have been proposed by complex ConvNet architectures [27, 28] and they perform considerably well in human pose datasets. Especially, multi-stage convolutional networks can capture information in large receptive field and can be trained while avoiding problems of difficult optimization by using an intermediate supervision for each stage. And they refine heat-maps of body joints via multi-stages.

In order to improve accuracy of joint predictions, the proposed architecture is based on the famous Convolutional Pose Machine (CPM) architecture [28]. However, we only choose 4 stages to train our network, Because CPM showed that the performance increases monotonically till 5 stages and too many layers can't improve the performance significantly. Every stage produces 15 joint confidence maps (heat-maps) for each still image and then the heat-maps are sent into next stage.

Figure 3.4 shows an overview of our proposed network architecture which consists of one pre-stage and other four stages,

where F is fire module block, G is Gate technology and D is pyramid dilation convolution. Rectified Linear Units (ReLU) are used for faster training and dropout is employed during training to prevent from overfitting. We also take Stochastic Gradient Descent (SGD) to make a back propagation. We apply one big gate scaling to every stage to control the importance of the feature map weights. At the same time, there is one small gate in the end of each stage to catch the global and local information of the images.

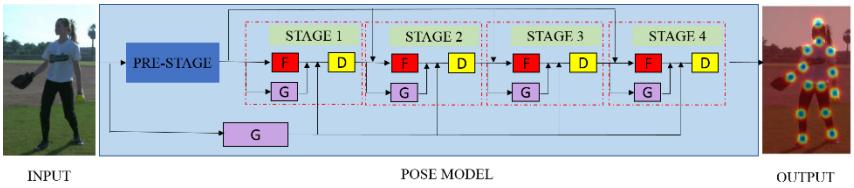

Figure 3.4 Overview of our proposed architecture.

As shown in Figure 3.4, feature maps obtained from an input image are downsampled through the pre-stage to do a down sample first and the output will be 8 times smaller than the input size because of doing pooling 3 times. In the pre-stage, we also apply Squeeze [29] method which is famous for smaller parameters but same level accuracy. In our architecture, we only use some blocks but not whole Squeeze network. For down sample operation, we use a parallel pooling and convolution with stride 2 together, which can keep more information from images.

Following each stage consist of scaling and pyramid dilated convolution together to produce 15 outputs. Their inputs (except first stage) include three parts: output of previous stage, output of pre-stage and big gate. But the output of big gate will concatenate with the output of fire module and go to dilation convolution in each stage respectively. The big scaling (gate) will take a stride of 8 to keep same output size with each stage working for PRE-STAGE. But the

small scaling won't change feature map size working for fire module block in each stage.

Our neural network keeps the receptive field large enough for learning the long-range spatial relationships by deep convolutions. Also, intermediate supervision is used to produce intermediate confidence maps and refine the loss through different stages. Intermediate supervision can reduce vanishing gradients when the network becomes deeper. Figure 3.5 shows the processing of the different stages learning.

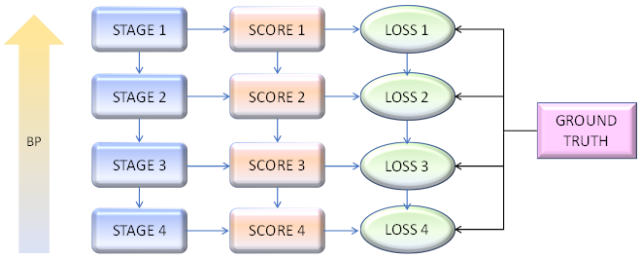

Figure 3.5 Procedure of BP in training.

3.2.1 Pre-stage and Stage 1

Several recent works regress heat-maps of body joints via convolutional neural network [27, 28]. These models have the ability to capture special properties. While our goal is to improve accuracy of joint predictions, here we leverage CPM [28] as basis for our architecture. However, we only choose 4 stages to train our network since too many layers can't improve the performance significantly. Now we will introduce our network detail inside.

All the stages of the proposed architecture have the same structure but a little different inputs. Here we will introduce pre-stage and take first stage as an example Here we will introduce pre-stage and take first stage as an example. In CPM, they only assemble convolutional

layers and pooling layers together simply. By contrast, the proposed architecture in Figure 3.6 is combined with fire module, down sample block, dilated block and Gate (Scaling). It looks our architecture is much more complex than the original one. All these technologies will be explained detail in following sessions.

As shown in Figure 3.6, the pre-stage is located in the green dashed box consists of a 3x3 convolution layer and a stack of three fire modules and down sample blocks. Every down sample block is in front of fire module. This pre-stage is applied for reducing the dimension of images.

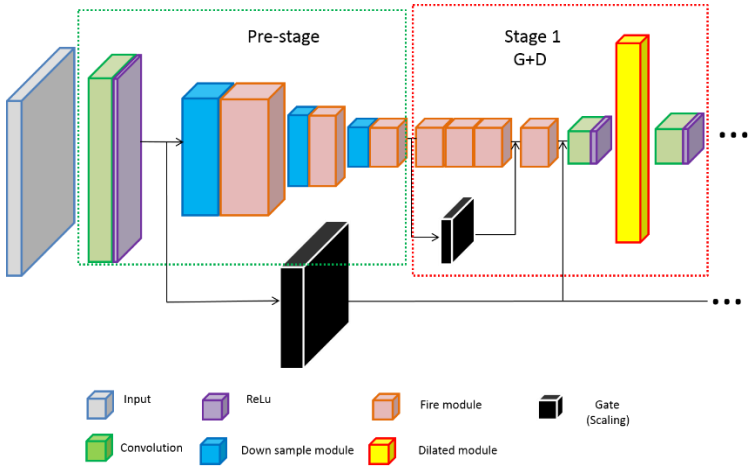

Figure 3.6 First stage of our architecture.

Stage 1 is in the red dashed box, which consists of a small gate and a pyramid dilated convolution layer. We use four fire modules instead of four simple convolution layers in the first stage. A small gate comes from pre-stage and it skips first three fire modules going to next fire module. After first three fire modules and a small gate, they will make a concatenation then go to next fire module. There is one convolution layer inserts in the fourth fire module and dilated block. This one convolution' inputs come from the fourth fire module

and the big gate. Another convolution is following dilated block and its outputs are 15 heatmaps going to next new stage.

3.2.2 Fire Module

It is popular to use 1x1 to reduce parameters in the recent advanced architectures. The original fire module from SqueezeNet [29] consists of 2 layers comprised of 1x1 and 3x3 convolution filters and it is used to replace a convolution layer. Our fire module, its variant, includes two original fire modules: one is normal and another take a bypass additionally, as shown in Figure 3.7. And in [29], the authors showed that when using a bypass made results better. That's the reason we use two original fire modules to build up our new fire module. One can keep going into a further deep process, while another can use a skip to make a concatenation which can encourage feature reuse.

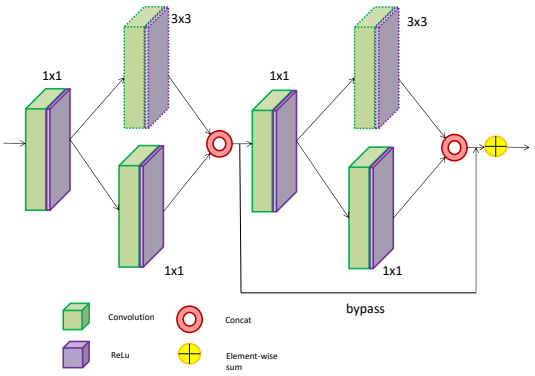

Figure 3.7 Fire module.

To understand bypass better, we can see the Figure 3.8 which shows that when you get a batch of goods, and you have to distribute them to everyone. Then you choose to transfer one by one, but the last person may not get the goods since he is too far from the goods. Now the other path is like one expressway, which can bring some

goods to give the person who doesn't get the goods. Bypass just works as an expressway in the network and make sure the neural network can keep more information during feature extraction. Usually, the deeper the layer, the more abstract concept can be covered. To get the better feature encoding for human pose estimation, shallow layers also have sufficient abstractive representation.

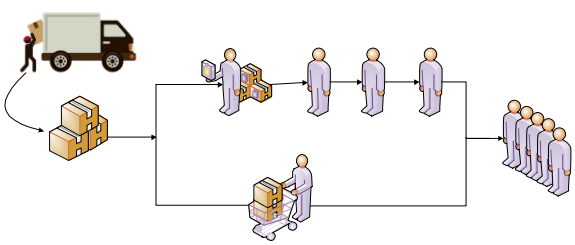

Figure 3.8 Transfer goods bypass.

3.2.3 Down Sample Module

It is common to insert a pooling layer between two convolution layers to progressively reduce the spatial size of the representation and also control overfitting. And stride can be used to control how the filter convolves around the input feature map. Figure 3.9 shows the proposed down sample module where the dual path to reduce the grid-size is heavily inspired by [31]. We choose a pooling layer in parallel with a convolution of stride 2 and concatenate these output feature maps. Before pooling, we use one dropout with a rate 0.5. The Max Pooling layer is performed with non-overlapping 2x2 windows.

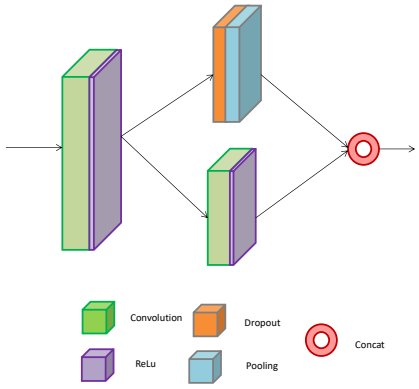

Figure 3.9 Down sample module.

3.2.4 Dilated Module

It is famous that dilated convolutions have a wide receptive field in the network and avoid overly down sampling the feature maps. Inspired by [32], we design our dilated block which employ dilated convolution in parallel to capture multi-scale context by adopting multiple dilations factors. Dilated convolutions are also known as atrous convolutions.

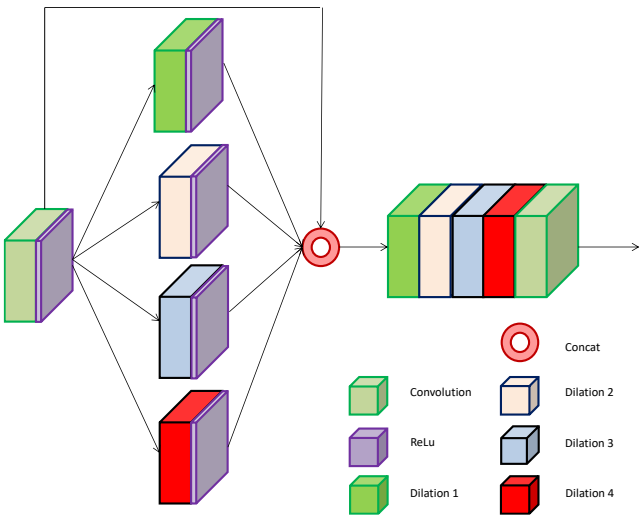

Figure 3.10 dilated Module.

Our four parallel dilated convolution with different dilations which are applied on top of the feature map. This block consists of four 3x3 convolutions with dilation 1, 2, 3 and 4 respectively. They are fused as the global prior. We also concatenate them with the original feature map. After concatenating all the branches of features, it will pass through another 1x1 convolution and generate 15 outputs.

Our dilated block has three advantages, first one is that it allows us to enlarge the field of view of filters effectively incorporating multi-scale context; second one is that with parallel dilated convolution layers, it can capture multi-scale information; the third one is that dilated convolutions will not lose resolution or coverage with exponential expansion of the receptive field. Our dilated block is able to control the resolution where feature responses are computed with ConvNet without learning extra parameters as shown in Figure 3.10.

3.2.5 Gate (Scaling)

Totally we set five gates, one big gate and four small gates. Every stage will consist of one small gate. Our gate designed is to control the weighting of each feature map. Our idea comes from [33], which is the winner of ImageNet 2017 classification task. We already introduce detail about Squeeze-and -Excitation networks (SENet) in session 3.1.3. The core idea of SENet is to learn the feature weight according to the loss in the network. The network can adaptively adjust the weighting of each feature map by adding parameters to each channel of a convolutional block. It can achieve better results by using ineffective or effective feature maps. We can get the global information by the big gate and local information by the small gates.

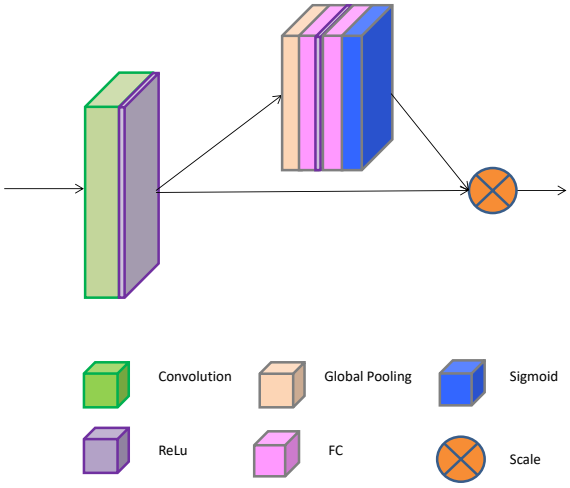

Figure 3.11 Gate (Scaling).

As shown in Figure 3.11, the gate is combined with one convolution layer and squeezed by one global pooling followed by two full connection layer and a sigmoid function. In the end they do a scaling to give each channel a weight.

Chapter 4 Experiments and Results

In this thesis, we adapt our network to HPE by applying fire module, scaling and parallel dilated convolution to extract dense features. Our implementation is built on caffe [56] framework. We evaluate the propose models on LSP dataset which includes 14 joints. The performance is measured by PCP and PCK.

4.1 Evaluation Methodology

4.1.1 PCP Evaluation

In all experiments, we use most popular criterions which are the correctly estimated body Parts (PCP) and Probability of Correct Keypoint (PCK). PCP measures the percentage of correctly localized body parts. This criteria proposed by [34] presents that one part is considered correctly localized if its predicted endpoints are within 50% part length of the corresponding ground truth segment from their annotated location. That is to say predicted endpoints are closer to their ground-truth location than a threshold of 50%, which will be labeled as correct. In all experiments, we use most popular criterions which are the correctly estimated body Parts (PCP) and Probability of Correct Keypoint (PCK). PCP measures the percentage of correctly localized body parts. This criteria proposed by [34] presents that one part is considered correctly localized if its predicted endpoints are within 50% part length of the corresponding ground truth segment from their annotated location. That is to say predicted endpoints are closer to their ground-truth location than a threshold of 50%, which will be labeled as correct.

4.1.2 PCK Evaluation

Another common evaluation is PCK, which reports the percentage of detections that locate in a normalized distance of the ground truth joint localizations. To obtain the PCK evaluation, we need ground truth key point of the body joint and the predicted keypoint localization. The Euclidian distance (l) between the ground truth keypoint and estimated keypoint. The bounding box of person to get scale (s) and the threshold (t) are also necessary. The scale is the maximum of the height (h) and the width (w) of the human bounding box respectively. Ground truth key points are given through manually annotating 14 joints of each human in each image. It is used to control the relative threshold for considering correctness. While t = 0.9, the evaluation is very tolerant and t = 0.1, the evaluation will be too strict. Hence, the less the value of the threshold, the more strict the evaluation is and we use t =0.2 in our research. If the keypoint falls within t*s pixels of the ground truth keypoint localization, the predict keypoint will be considered to be correct. Otherwise, it will be regarded as mistake estimated localization. In another word, we will get correct localized body joint if the l is less than or equal to t*s pixels.

$$s(x) = \max(h, w)$$

$$g(x) = \begin{cases} 1, & l \le t * s \\ 0, & otherwise \end{cases}$$

4.2 Data

Our architecture predicts 14 human full-body joint location. These 14 joints are head, neck, left shoulder, right shoulder, left elbow, right elbow, left wrist, right wrist, left hip, right hip, left knee, right knee, left ankle and right ankle. We train our proposed method on two

datasets: Leeds Sport Pose (LSP) and MPII Human Pose (MPII) Dataset.

Leeds Sports Pose (LSP) Dataset. This dataset [35] consists of 11,000 images for training and 1000 images for testing. These images are gathered from Flick searches with sport people.

MPII Human Pose Dataset. MPII dataset contains 28,821 annotated pose for training and 11,701 for testing. These images are collected from YouTube videos. The annotated pose have 16 body joints, some of them are not present and some are in occlusion but can be predicted by the context information.

Data augmentation. We choose to augment the data on both MPII and LSP datasets. We use random rotation degrees in $[-40^{o}, 40^{o}]$ and random rescaling in [0.7, 1.3] to make the model more robust for image changing. Input RGB images were cropped and centered on the main subject with one squared bounding box to keep people scale. Horizontal flipping is also applied to do data augmentation. After all, the training images are resized to 368x368 pixels.

4.3 Experiments

We produce two kinds of outputs with modified version of caffe from the data layer, one is the augmented image, and another is the corresponding transformed ground truth heatmaps. We train the model with an initial learning rate of 10^{-4}. Two Pascal TITAN GPUs are used to train the merged dataset of extended LSP and MPII. Both of these two datasets provide the visibility of body parts and we use them as the supervision occlusion signal during training. However, we only test our model on LSP dataset. The input resolution of the images is 368x368 with max pooling dropping down to 46. The number of output features is 15.

4.4 Software Libraries Used

Caffe. Caffe is a deep learning framework made with expression, speed, and modularity in mind. It is written in C++, with a Python interface. Caffe supports many different types of deep learning architectures, such as CNN, RCNN, and LSTM. It also supports GPU based acceleration with CuDNN of Nvidia.

OpenCV2.4. OpenCV (Open Source Computer Vision) is a library of programming functions mainly aimed at real-time computer vision. OpenCV support a wide variety of languages such as Python, Java, C++, etc., and it is available on different platforms including ios, Android, Linux, and Windows. A multitude of algorithms related to Computer Vision and Machine Learning is expanding by OpenCV developed by Intel. OpenCV is a cross-platform and free source initiative.

CUDA8. Compute Unified Device Architecture (CUDA) is a parallel computing platform created by Nvidia. The CUDA platform is accessible to software developers to use a Graphic Processing Unit (GPU) for general purpose processing. CUDA supports programming frameworks such as OpenCL and OpenACC.

4.5 Results

To validate our idea without any severe burden in time on relatively slow GPUs, we use the CPM with 3 stages as the base line and a limited iterations of 100,000. We train networks with our methods respectively and test on LSP dataset. In Tables 4.1 and 4.2, it shows that our methods mostly improve the baseline CPM. Here SQ is the technology SqueezeNet and we replace simple convolution layer with our fire module in the network. PD means we use pyramid

dilation convolution in each stage. The last Gate (scaling) is that we set five gates in the CPM.

Our method improves significantly the baseline with the same stages and training. Thus the improvement comes from the different architecture of the network. It shows that new network can do feature extract better in the human pose machine.

Table 4.1. Performance comparison of different methods on the LSP dataset (PCK@0.2)

Methods	Head	Shoulder	Elbow	Wrist	Hip	Knee	Ankle	Total
CPM [28]	94.9	65.9	58.0	51.7	61.8	54.5	53.2	62.9
SQ	95.4	70.0	62.5	56.5	66.2	57.4	54.2	66.0
PD	94.5	66.0	60.0	54.6	63.6	55.3	63.5	65.4
Gate	95.3	68.2	59.9	53.4	64.7	56.0	55.6	64.7

Table 4.2. Performance comparison of different methods on the LSP dataset (PCP)

Methods	Torso	Upper leg	Lower leg	Upper arm	Fore arm	Head	Total
CPM [28]	92.9	55.4	46.6	53.1	38.5	91.4	57.1
SQ	87.4	52.1	47.4	44.9	25.2	84.5	51.1
PD	93.2	57.5	47.1	52.9	37.5	90.6	57.4
Gate	95.0	58.4	48.3	54.5	38.7	90.8	58.6

We combined these three technologies as one model to train the network and our results are represented in Table 4.3 and Table 4.4. In order to make a comparison to other approaches on the task of HPE from a single image, we evaluate our approach on the LSP test set. It is clear that we achieve promising performance for the keypoint localization. In particular, our model performs better than the original method proposed by Wei et al. [28]. For the most challenging body parts such as ankle and wrist, our approach achieves 1.7% and 2.4% improvement compared with [28].

Table 4.3. Performance comparison with previous work on the LSP dataset (PCK@0.2)

Methods	Head	Shoulder	Elbow	Wrist	Hip	Knee	Ankle	Total
Rafi [1]	95.8	86.2	79.3	75.0	86.6	83.8	79.6	83.8
Yu [57]	87.2	88.2	82.4	76.3	91.4	85.8	78.7	84.3
Belagian [58]	95.2	89.0	81.5	77.0	83.7	87.0	82.8	85.2
Lifshitz [59]	96.8	89.0	82.7	79.1	90.9	86.0	82.5	86.7
Pishchu [60]	97.0	91.0	83.8	78.1	91.0	86.7	82.0	87.1
Insafutd [61]	97.4	**92.7**	87.5	84.4	91.5	89.9	87.2	90.1
Wei [28]	**97.8**	92.5	87.0	83.9	91.5	90.8	89.9	90.5
Bula [62]	97.2	92.1	88.1	85.2	**92.2**	91.4	88.7	90.7
Ours	97.4	**92.7**	**88.8**	**86.3**	91.9	**92.8**	**91.6**	**91.6**

Table 4.4. Performance comparison with previous work on the LSP dataset (PCP)

Methods	Torso	Upper leg	Lower leg	Upper arm	Fore arm	Head	Total
Rafi [1]	97.6	87.3	80.2	76.8	66.2	93.3	81.2
Belagian [58]	96.0	86.7	82.2	79.4	69.4	89.4	82.1
Lifshitz [59]	97.3	88.8	84.4	80.6	71.4	94.8	84.3
Pishchu [60]	97.0	88.8	82.0	82.4	71.8	**95.8**	84.3
Yu [57]	98.0	93.1	88.1	82.9	72.6	83.0	85.4
Insafutd [61]	97.0	90.6	86.9	86.1	79.5	95.4	87.8
Wei [28]	98.0	82.2	89.1	85.8	77.9	95.0	88.3
Bula [62]	97.7	92.4	89.3	86.7	79.7	95.2	88.9
Ours	**98.2**	**94.0**	**91.6**	**87.5**	**81.2**	95.6	**90.2**

Some examples of heatmaps produced by our model are shown in Figure 4.1. As we can see from left-top second to right-bottom second, the presented heatmaps correspond to head, neck, left shoulder, right shoulder, left elbow, right elbow, left wrist, right wrist, left hip, right hip, left knee, right knee, left ankle and right ankle.

Figure 4.1 Example output produced by our system. On the top left is input image and on the bottom right is the final pose estimation. From top to down, we show sample heatmaps.

As shown in Figure 4.2, we can see the most challenging parts ankle, knee and wrist in each stage are predicted as heatmaps. It shows that

accuracy improves with the stage increasing.

Figure 4.2 Prediction of challenging keypoints in each stage.

Chapter 5 Conclusions and Future Works

We have shown successfully how to improve the performance of HPE which is one of the most complex computer vision tasks. We explored many different state-of-the-art structural models and adapted their concepts and modules to propose an architecture showing better performance. The proposed HPE is a variant of the original CPM. It is based on a multi-stage structure and includes some modules such as fire module, gating module, and dilated convolution to reduce the number of parameters in the proposed network and use multi-scale information in space and feature map. The experimental results show that the proposed HPE outperforms most of the other state-of-the-art methods based on deep learning.

For the related further study, it will be required to analyze the proposed method in more detail according to their hyperparameters. And also we need to research on simplifying the proposed architecture to reduce time complexity. Furthermore, we wish we could apply this method to multi-person pose estimation and real-time pose estimation. For hand pose or face landmark location, they are also trying to find the precise location. In the future, HPE can be applied to activity recognition to understand humans and their interactions with other humans or objects.

Acknowledgements

I came to South Korea from China to start my master courses in Multimedia Lab at Chonbuk National University two years ago. During the unforgettable period, I would like to show my appreciations to these people who have supported and encouraged me. Thanks to them I could pass my joyful and fruitful time here.

Above all, I would like to express my sincere gratitude to Professor Dong Sun Park who is my supervisor. His sincere attitude on research and strong professional skill are having impressed me deeply. I respect and like him a lot. Professor Park likes talking Chinese culture with us and we enjoy a lot from it. In my hard time, I also got help from him. He always gives me a hand on the road of my scientific research and my life.

I would like to thank Professor Sang-seob Song. He is a very easygoing Professor and very patient to explain difficult questions for us. I choose his courses every semester almost. Professor Hyongsuk Kim impressed me for his professional knowledge. They are very great professors. I couldn't graduate well without their patiently and professional guidance.

I would like to thank Professor Sook Yoon, who is amiable and provides unselfish support for me. She teaches me how to improve the quality of a paper with face to face chat and gives me some ideas about my researches. I would like to thank Professor Yong Yang who introduced me to Korea and join in our lab and brother Zhihui Wang who helped me a lot.

Now my thanks will go to all my lab mates. They are Korean brother Jong Bin Park, 이재환, 장대석, 홍창표, 이동석, 인배,

규명, Korean sister 이유정, and Alvaro Fuentes, Zhang Yujia, Haseeb Nazki. We work together and support each other with a cheerful working atmosphere. I will never forget the time we travel, chat and discuss together.

I am grateful to my dear Chinese sister, Chunzao Cui, who brings me to adapt the life in Korea and solve the Korean language. She takes care about my life a lot. I also thank my other roommates: Zhenchao Wan, Huihui Pang and Yong Jiang. We share happiness and sadness in our daily time. I have to thank Ibrahim, Akanksha and Linzi Wang who help my English and paper grammar problems.

Last but not least, I must like to thanks to my family and friends in China now. They give me great support whenever I have difficulties. They encourage and comfort me.

I love you all and thank you a lot!

References

[1] Rafi, Umer, et al. "An Efficient Convolutional Network for Human Pose Estimation." BMVC. Vol. 1. 2016.

[2] Yu, Fisher, and Vladlen Koltun. "Multi-scale context aggregation by dilated convolutions." *arXiv preprint arXiv:1511.07122* (2015).

[3] Chen, Yunpeng, et al. "Dual path networks." *arXiv preprint arXiv:1707.01629* (2017).

[4] Huang, Gao, et al. "Densely connected convolutional networks." *arXiv preprint arXiv:1608.06993* (2016).

[5] He, Kaiming, et al. "Deep residual learning for image recognition." *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2016.

[6] Ioffe, Sergey, and Christian Szegedy. "Batch normalization: Accelerating deep network training by reducing internal covariate shift." *International Conference on Machine Learning*. 2015.

[7] Szegedy, Christian, et al. "Rethinking the inception architecture for computer vision." Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2016.

[8] Szegedy, Christian, et al. "Inception-v4, Inception-ResNet and the Impact of Residual Connections on Learning." *AAAI*. 2017.

[9] Lin, Min, Qiang Chen, and Shuicheng Yan. "Network in network." *arXiv preprint arXiv:1312.4400* (2013).

[10] Szegedy, Christian, et al. "Going deeper with convolutions." *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2015.

[11] Chu, Xiao, et al. "Structured feature learning for pose estimation." *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 2016.

[12] Chu, Xiao, Wanli Ouyang, and Xiaogang Wang. "Crf-cnn: Modeling structured information in human pose estimation." *Advances in Neural Information Processing Systems*. 2016

[13] Tompson, Jonathan, et al. "Efficient object localization using convolutional networks." *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 2015.

[14] Carreira, Joao, et al. "Human pose estimation with iterative error

feedback." *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 2016.

[15] Lifshitz, Ita, Ethan Fetaya, and Shimon Ullman. "Human pose estimation using deep consensus voting." *European Conference on Computer Vision*. Springer International Publishing, 2016.

[16] Tompson, Jonathan J., et al. "Joint training of a convolutional network and a graphical model for human pose estimation." *Advances in neural information processing systems*. 2014.

[17] Papandreou, George, et al. "Towards Accurate Multi-person Pose Estimation in the Wild." *arXiv preprint arXiv:1701.01779*(2017).

[18] He, Kaiming, et al. "Mask r-cnn." *arXiv preprint arXiv:1703.06870* (2017).

[19] Hwang, Jihye, Sungheon Park, and Nojun Kwak. "Athlete pose estimation by a global-local network." *Computer Vision and Pattern Recognition Workshops (CVPRW), 2017 IEEE Conference on*. IEEE, 2017.

[20] Tran, Duan, and David Forsyth. "Improved human parsing with a full relational model." *Computer Vision–ECCV 2010* (2010): 227-240.

[21] Chang, Ju Yong, and Seung Woo Nam. "Fast Random-Forest-Based Human Pose Estimation Using a Multi-scale and Cascade Approach." *ETRI Journal* 35.6 (2013): 949-959.

[22] Chen, Yu, et al. "Adversarial PoseNet: A Structure-aware ConvolutionalNetwork for Human Pose Estimation." *arXiv preprint arXiv:1705.00389* (2017).

[23] Ma, Liqian, et al. "Pose Guided Person Image Generation." *arXiv preprint arXiv:1705.09368* (2017).

[24] Chou, Chia-Jung, Jui-Ting Chien, and Hwann-Tzong Chen. "Self adversarial training for human pose estimation." *arXiv preprint arXiv:1707.02439* (2017).

[25] Isola, Phillip, et al. "Image-to-image translation with conditional adversarial networks." *arXiv preprint arXiv:1611.07004* (2016).

[26] Linna, Marko, Juho Kannala, and Esa Rahtu. "Real-time human pose estimation from video with convolutional neural networks." *arXiv preprint arXiv:1609.07420* (2016).

[27] Newell, Alejandro, Kaiyu Yang, and Jia Deng. "Stacked hourglass networks for human pose estimation." *European Conference on Computer Vision*. Springer International Publishing, 2016.

[28] Wei, Shih-En, et al. "Convolutional pose machines." *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 2016.

[29] Iandola, Forrest N., et al. "SqueezeNet: AlexNet-level accuracy with 50x fewer parameters and< 0.5 MB model size." *arXiv preprint arXiv:1602.07360* (2016).

[30] Paszke, Adam, et al. "Enet: A deep neural network architecture for real-time semantic segmentation." *arXiv preprint arXiv:1606.02147* (2016).

[31] Szegedy, Christian, et al. "Rethinking the inception architecture for computer vision." *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 2016.

[32] Paszke, Adam, et al. "Enet: A deep neural network architecture for real-time semantic segmentation." *arXiv preprint arXiv:1606.02147* (2016).

[33] Hu, Jie, Li Shen, and Gang Sun. "Squeeze-and-Excitation Networks." *arXiv preprint arXiv:1709.01507* (2017).

[34] Ferrari, Vittorio, Manuel Marin-Jimenez, and Andrew Zisserman. "Progressive search space reduction for human pose estimation." *Computer Vision and Pattern Recognition, 2008. CVPR 2008. IEEE Conference on*. IEEE, 2008.

[35] Johnson, Sam, and Mark Everingham. "Learning effective human pose estimation from inaccurate annotation." *Computer Vision and Pattern Recognition (CVPR), 2011 IEEE Conference on*. IEEE, 2011.

[36] Krizhevsky, Alex, Ilya Sutskever, and Geoffrey E. Hinton. "Imagenet classification with deep convolutional neural networks." *Advances in neural information processing systems*. 2012.

[37] Simonyan, Karen, and Andrew Zisserman. "Very deep convolutional networks for large-scale image recognition." *arXiv preprint arXiv:1409.1556* (2014).

[38] Simo-Serra, Edgar, et al. "A high performance CRF model for clothes parsing." *Asian conference on computer vision*. Springer, Cham, 2014.

[39] Tangseng, Pongsate, Zhipeng Wu, and Kota Yamaguchi. "Looking at Outfit to Parse Clothing." *arXiv preprint arXiv:1703.01386* (2017).

[40] Wu, Ziyan, Yang Li, and Richard J. Radke. "Viewpoint invariant human re-identification in camera networks using pose priors and subject-discriminative features." *IEEE transactions on pattern analysis and machine intelligence* 37.5 (2015): 1095-1108.

[41] Zhao, Liming, et al. "Deeply-Learned Part-Aligned Representations for Person Re-Identification." *arXiv preprint arXiv:1707.07256* (2017).

[42] Xu, Yuanlu, et al. "Human re-identification by matching compositional template with cluster sampling." *proceedings of the IEEE International*

Conference on Computer Vision. 2013.

[43] García-Montero, Mario, et al. "Fast Head Pose Estimation for Human-Computer Interaction." *Iberian Conference on Pattern Recognition and Image Analysis*. Springer International Publishing, 2015.

[44] Xu, Pei. "A Real-time Hand Gesture Recognition and Human-Computer Interaction System." *arXiv preprint arXiv:1704.07296* (2017).

[45] Iqbal, Umar, Martin Garbade, and Juergen Gall. "Pose for action-action for pose." *Automatic Face & Gesture Recognition (FG 2017), 2017 12th IEEE International Conference on*. IEEE, 2017.

[46] Pishchulin, Leonid, Mykhaylo Andriluka, and Bernt Schiele. "Fine-grained activity recognition with holistic and pose based features." *German Conference on Pattern Recognition*. Springer International Publishing, 2014.

[47] Edwards, Michael, Jingjing Deng, and Xianghua Xie. "From pose to activity: Surveying datasets and introducing CONVERSE." *Computer Vision and Image Understanding* 144 (2016): 73-105.

[48] Iqbal, Umar, Anton Milan, and Juergen Gall. "Pose-Track: Joint Multi-Person Pose Estimation and Tracking." *arXiv preprint arXiv:1611.07727* (2016).

[49] Marr, David, and Herbert Keith Nishihara. "Representation and recognition of the spatial organization of three-dimensional shapes." *Proceedings of the Royal Society of London B: Biological Sciences* 200.1140 (1978): 269-294.

[50] Fischler, Martin A., and Robert A. Elschlager. "The representation and matching of pictorial structures." *IEEE Transactions on computers* 100.1 (1973): 67-92.

[51] Yang, Yi, and Deva Ramanan. "Articulated human detection with flexible mixtures of parts." *IEEE Transactions on Pattern Analysis and Machine Intelligence* 35.12 (2013): 2878-2890.

[52] Sapp, Ben, and Ben Taskar. "Modec: Multimodal decomposable models for human pose estimation." *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 2013.

[53] Dalal, Navneet, and Bill Triggs. "Histograms of oriented gradients for human detection." *Computer Vision and Pattern Recognition, 2005. CVPR 2005. IEEE Computer Society Conference on*. Vol. 1. IEEE, 2005.

[54] Ho, Huy Tho, and Rama Chellappa. "Automatic head pose estimation using randomly projected dense sift descriptors." *Image Processing (ICIP), 2012*

19th IEEE International Conference on. IEEE, 2012.

[55] Toshev, Alexander, and Christian Szegedy. "Deeppose: Human pose estimation via deep neural networks." *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 2014.

[56] Jia, Yangqing, et al. "Caffe: Convolutional architecture for fast feature embedding." *Proceedings of the 22nd ACM international conference on Multimedia*. ACM, 2014.

[57] Yu, Xiang, Feng Zhou, and Manmohan Chandraker. "Deep deformation network for object landmark localization." *European Conference on Computer Vision*. Springer International Publishing, 2016.

[58] Belagiannis, Vasileios, and Andrew Zisserman. "Recurrent human pose estimation." *Automatic Face & Gesture Recognition (FG 2017), 2017 12th IEEE International Conference on*. IEEE, 2017.

[59] Lifshitz, Ita, Ethan Fetaya, and Shimon Ullman. "Human pose estimation using deep consensus voting." *European Conference on Computer Vision*. Springer International Publishing, 2016.

[60] Pishchulin, Leonid, et al. "Deepcut: Joint subset partition and labeling for multi person pose estimation." *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 2016.

[61] Insafutdinov, Eldar, et al. "Deepercut: A deeper, stronger, and faster multi-person pose estimation model." *European Conference on Computer Vision*. Springer International Publishing, 2016.

[62] Bulat, Adrian, and Georgios Tzimiropoulos. "Human pose estimation via convolutional part heatmap regression." *European Conference on Computer Vision*. Springer International Publishing, 2016.

요 약 문

사람 자세 추정은 생체역학 연구에서 물리 치료, 외골격 제어에 이르기까지 다양한 응용 분야에서 활용될 수 있으나, 복잡한 배경, 다양한 형태, 조명 변화 등이 문제를 어렵게 만들기 때문에 컴퓨터 비젼 분야에서는 도전적인 연구 주제이다. 최근, 강력한 CNN 구조가 영상으로부터 사람 자세의 중요한 정보를 뽑아내기 위해 연구되고 있다.

본 연구에서는 정지영상으로부터 사람 자세를 추정할 수 있는 종단간 학습 가능한(end-to-end trainable) CNN 구조를 제안한다. 이 CNN구조는 하나의 전처리단(pre-stage)과 다단의 콘볼루션 신경망으로 이루어지는 네 개의 단(stage)으로 구성되는 다단 구조를 가진다. 첫째, 우리는 특징맵(feature map) 가중 제어를 위해 5개의 게이트(gate, scaling)를 사용한다. 전역 정보를 제어하기 위해 1개의 큰 게이트(big gate)를 사용하고 각각의 스테이지에 작은 게이트(small gate)를 두어 지역 정보를 제어한다. 둘째, 우리는 모델의 파라미터 수를 줄이기 위하여 하나의 콘볼루션 층들을 사용하는 대신에 squeeze net의 fire 모듈을 사용하고, 모듈에서 Skip connection을 사용하여 정보를 유지하고 전역적 또는 지역적 정보 맥락을 동시에 통합하여 각 해상도에서 특징들이 보다 잘 유지되도록 할 수 있도록 한다. 셋째, 우리는 몸의 각 부위에 대해 다중 스케일 정보를 학습시키기 위해 pyramid dilated convolution을 사용한다. Dilated convolution은 콘볼루션망의 파라미터를 위한 추가적인 학습 없이 선택적으로 응답을 선택함으로써 해상도를 제어할 수 있도록 하며 또한 FOV(field of view)를 효과적으로 확대할 수 있도록 한다. 본 연구에서는 특징맵으로부터 다중 스케일 정보를 획득하기 위해 다른 dilation factor을 가지는 여러 dilated convolution layer를 병렬로 사용한다. 이러한 기술들을 결합함으로써,

우리는 넓은 범위의 공간상 관계(long-range spatial relationship)를 학습하기에 충분히 큰 수용 영역(receptive field)를 얻을 수 있다. 또한, 우리는 다단계 네트워트상에서 중간 confidence map들을 생성하고 다른 스테이지들을 통해 정제되도록 하기 위해 중간 지도 기법(intermediate supervision)을 사용한다.

본 연구는 사람 자세 추정에서 난제의 하나인 팔다리의 구조적 복잡도에 보다 정확하고 효율적으로 반응하도록 하는데 게이트를 이용한 스케일링 방식과 여러 dilation을 사용하여 특징들을 리샘플링하는 것이 효과적이라는 것을 보여주고 있다. 사람 자세 추정 연구에 주로 사용되는 데이터셋인 LSP와 MPII를 사용하여 사람 자세 추정 성능을 시험하였고, 다른 CNN 기반 방법들에 비해 개선된 성능을 보였다.

키워드: 사람 자세 추정, CNN, 다단, squeeze, 게이트, 스케일링, Pyramid dilated convolution