Research Project of Master Thesis

Joint 2D and 3D Pose Estimation and Action Recognition.

Author:

- **AMMAR KHODJA Rayane**

Supervised by:

- **Professor HEDI Tabia**

Master 2 MMVAI January 2024.

Paris Saclay University.

Laboratoire de Recherche Paris Saclay University, IBISC Evry Val d'Essonne.

Abstract

The paper presentted by Diogo C. Luvizon1 , David Picard and Hedi Tabia develops a multitask deep learning architecture for joint 2D/3D pose estimation and action recognition from RGB images. The proposed method achieves state-of-the-art results on both tasks and can be trained with data from different categories simultaneously. The architecture consists of a shared feature extractor followed by two task-specific branches for pose estimation and action recognition. The pose estimation branch uses a continuous regression function to predict joint locations in 2D and 3D, while the action recognition branch uses an attention-based pooling method to extract visual features from body parts. The authors demonstrate the effectiveness of their approach on several benchmark datasets and show that it outperforms separate learning for pose estimation and action recognition. Thus , In this project we will try to estimate the same results by training using a VGG-16 [VGG 16] and check if we can achieve the same presented results in the paper.

1 Introduction

The paper `2D/3D Pose Estimation and Action Recognition using Multitask Deep Learning` by `Diogo C. Luvizon1 , David Picard` and `Hedi Tabia` highlights the importance of human action recognition and pose estimation in computer vision, emphasizing their relevance in applications such as video surveillance and human-computer interfaces. The authors note that while these tasks are typically treated as distinct problems in the literature, they propose a unique end-to-end trainable multitask framework to jointly handle 2D and 3D human pose estimation and action recognition. They argue that their approach offers several advantages over separate learning methods and aim to demonstrate the effectiveness of their proposed architecture through state-of-the-art results on multiple datasets, as shown in Figure 1

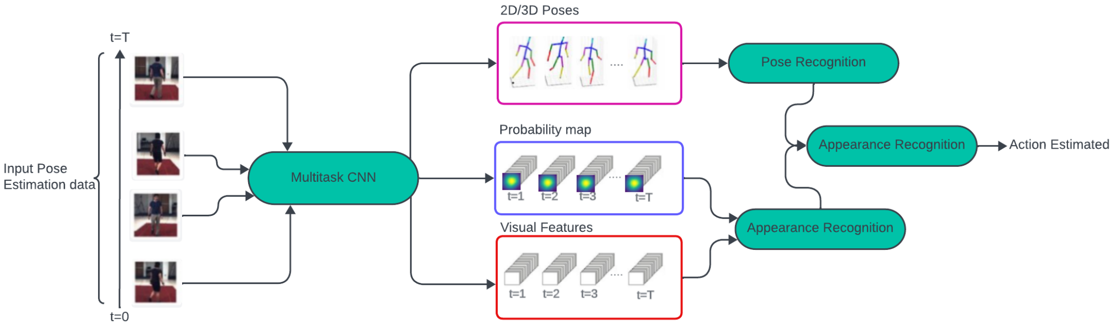


Figure 1: Proposed multitask framework for joint 2D and 3D pose estimation and action recognition; which depict the flow of information through the network and how the model integrates information from RGB images to perform both tasks simultaneously.

Important Note

- This whole project is based on the research paper of `2D/3D Pose Estimation and Action Recognition using Multitask Deep Learning` by `Diogo C. Luvizon1 , David Picard` and `Hedi Tabia` [**Tabia**]. Hence, most of the ideas are driven from that research paper, however, I changed the Network architecture from Inception-V4 [**Inception-v4**] into VGG presented by [**VGG 16**].

2 Background from the research paper

Authors of the research paper mention that the their work aims to address the lack of joint methods for 2D and 3D pose estimation and action recognition in the literature. The proposed architecture is described as an end-to-end trainable framework that can handle both tasks efficiently. It is also highlighted that the method can be trained with data from different categories simultaneously, allowing for seamless generalization of 3D predictions from 2D annotated data. The authors express their intention to demonstrate the effectiveness of their approach through extensive experiments and comparisons with existing methods.

2.1 Human pose Estimation

Human pose estimation is the task of estimating the 2D or 3D positions of the joints of a human body from an image or a video. It is a fundamental problem in computer vision with many applications, such as human-computer interaction, surveillance, and sports analysis. Accurate human pose estimation is essential for many computer vision tasks, including action recognition, human tracking, and gesture recognition.

However, human pose estimation is a challenging task due to several factors. One of the main challenges is the high degree of articulation and variability in human poses. The human body can take on a wide range of poses, and the appearance of the body parts can vary significantly depending on factors such as clothing, lighting, and camera viewpoint. Another challenge is the occlusion of body parts, which can occur when the body parts are partially or completely hidden from view. Occlusion can make it difficult to accurately estimate the position of the occluded body parts.

Over the years, many methods have been proposed for human pose estimation, with varying degrees of success. In recent years, deep learning methods have shown promising results in both 2D and 3D pose estimation. In 2D pose estimation, convolutional neural networks (CNNs) have been used to learn features from images and estimate the positions of the body joints. In 3D pose estimation, methods based on depth sensors or multi-view cameras have been proposed, but these methods are often limited by the availability of data and hardware.

Some of the most successful methods for 2D pose estimation include the OpenPose [**OpenPose**] and Mask R-CNN [**Mask RCNN**] frameworks, which use CNNs to estimate the positions of the body joints. For 3D pose estimation, methods based on volumetric representations [**VTP**] and multi-view geometry [**MVGT**] have shown promising results. However, these methods often require large amounts of data and computational resources, which can be a limiting factor in practical applications.

2.2 Action Recognition

Action recognition is the task of identifying the action being performed by a person or object in a video sequence. It is an important problem in computer vision with many applications, such as surveillance, sports analysis, and human-computer interaction. Accurate action recognition is essential for many computer vision tasks, including activity recognition, event detection, and video summarizing.

However, action recognition is a challenging task due to several factors. One of the main challenges is the high degree of variability in human actions. The same action can be performed in many different ways, and the appearance of the body parts can vary significantly depending on factors such as clothing, lighting, and camera viewpoint. Another challenge is the temporal dimension of the problem, as actions can occur over varying lengths of time and can involve complex temporal dynamics.

Over the years, many methods have been proposed for action recognition, with varying degrees of success. In recent years, deep learning methods have shown promising results in action recognition. These methods typically involve using convolutional neural networks (CNNs) to learn features from video frames and then using recurrent neural networks (RNNs) or temporal convolutional networks (TCNs) to model the temporal dynamics of the actions.

Some of the most successful methods for action recognition include the Two-Stream CNN [**Two-Stream RNN/CNN**], which uses separate CNNs to model spatial and temporal information, and the 3D CNN , which extends the traditional CNN architecture to include temporal information. Other methods, such as the C3D [**C3D**] and I3D architectures, use 3D convolutions to model the spatiotemporal information directly.

2.3 Multitask Learning

Multitask Learning

Multitask learning is a machine learning paradigm where a model is trained to perform multiple related tasks simultaneously. In the context of computer vision, multitask learning involves training a single model to perform multiple vision-related tasks, such as object detection, semantic segmentation, pose estimation, and action recognition, among others.

The significance of multitask learning in computer vision lies in its ability to leverage shared representations and learn from related tasks to improve the overall performance of the model. By jointly learning multiple tasks, the model can benefit from the complementary information present in the different tasks, leading to improved generalization and better performance on each individual task.

Advantages of Multitask Learning

1. **Shared Representation Learning:** Multitask learning allows the model to learn a shared representation that captures common features across multiple tasks. This shared representation can help the model generalize better and learn more robust features.

2. **Improved Generalization:** By jointly learning multiple tasks, the model can learn to generalize better, especially when the tasks are related. This can lead to improved performance on each individual task, especially when data for each task is limited.

3. **Regularization Effect:** Multitask learning can act as a form of regularization, preventing overfitting by encouraging the model to learn representations that are useful for multiple tasks.

3 Multitask Framework

3.1 Introducing the Multitask Framework

The research paper proposes a multitask deep learning architecture for joint 2D and 3D pose estimation from still images and human action recognition from video sequences. The proposed framework aims to leverage shared representations and learn from related tasks to improve the overall performance of the model. By jointly learning multiple tasks, the model can benefit from the complementary information present in the different tasks, leading to improved generalization and better performance on each individual task.

The human pose regression problem (Non-linear mappings between inputs and outputs) involves an input RGB image $I \in R^{W \times H \times 3}$, where W and H are the width and height of the image, respectively. The goal is to predict the estimated pose $p^\wedge \in R^{N_J \times D}$, representing NJ body joints of dimension D. This prediction is achieved through a regression function f_r defined as follows:

$$p^\wedge = f_r(I, \theta_r) \qquad (1)$$

Here, θ_r represents a set of trainable parameters associated with the function f_r. The objective is to optimize these parameters (θ_r) to minimize the error between the estimated pose $p^\wedge$ and the ground truth pose p.

3.2 Architecture of the Network

The network architecture is divided into four parts: the multitask stem, the pose estimation model, the pose recognition model, and the appearance recognition model. Depth-wise separable convolutions, batch normalization, and ReLu activation are used. The multitask stem architecture of VGG 16 is detailed in Figure 2 . Each pose estimation prediction block is implemented as a multi-resolution CNN, and 16 heat maps are used for depth predictions. The prediction block for pose estimation involves the use of volumetric heat maps and Soft-argmax for 2D/3D pose loss . The training parameters include merging different datasets, using an alternated human pose layout, and optimizing the pose regression part using the RMSprop optimizer with an initial learning rate of 0.001. For action recognition, both pose and appearance models are trained simultaneously using a pretrained pose estimation model with weights initially frozen. The SGD optimizer with Nesterov momentum of 0.98 and initial learning rate of 0.0002 is used, and the final learning rate is divided by 10 when validation accuracy stagnates. The appearance-based recognition relies on local appearance features instead of joint coordinates, and the visual features are trained using both action sequences and still images captured ”in the wild”.

The Soft-argmax layer is used to convert heat maps to joint coordinates, and the joint visibility is computed by the Sigmoid function on the maximum value in the corresponding input heat map. The unified 2D/3D pose estimation involves expanding 2D heat maps to volumetric representations and using the Soft-argmax operation for both (x,y) and z components (Figure 3) . The proposed method integrates high-level pose information with low-level visual features in a multitask framework, allowing the sharing of the network entry flow for both pose estimation and visual features extraction .

$$\Psi(\mathbf{x}) = \left(\sum_{c=0}^{W_{\mathbf{x}}} \sum_{l=0}^{H_{\mathbf{x}}} \frac{c}{W_{\mathbf{x}}} \Phi(\mathbf{x})_{l,c}, \sum_{c=0}^{W_{\mathbf{x}}} \sum_{l=0}^{H_{\mathbf{x}}} \frac{l}{H_{\mathbf{x}}} \Phi(\mathbf{x})_{l,c} \right)^T \tag{2}$$

where $H_{i,j}$ is the heat map at position (i, j), and k and l are the indices of the heat map.

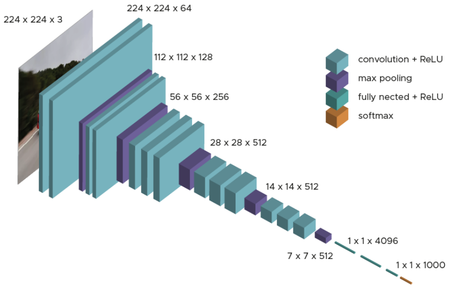


Figure 2: VGG-16 Architecture

The proposed approach for human pose estimation is a regression method, and the network architecture is based on VGG-16 (shown above Figure 2). The entry flow is used for basic features extraction, and prediction blocks are used to refine estimations. The Soft-argmax layer is used to indirectly learn joint probability maps, and the predictions are refined at each prediction block . The method allows for the training of the network with mixed 2D and 3D data, enabling the generalization of 3D predictions from 2D annotated data (Figure 3). The approach also benefits from images ”in the wild” for both 2D and 3D predictions, which has been proven to be an efficient way to learn visual features .

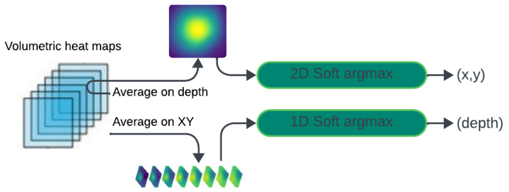


Figure 3: Unified 2D/3D pose estimation by using volumetric heat maps.

3.2.1 Pose Re-injection Property

As systematically noted in recent works [**Heatmap Regression**], predictions re-injection is a very efficient way to improve precision on estimated poses. Differently from all previous methods based on direct heat map regression, our approach can benefit from prediction re-injection at different resolutions, since our pose regression method is invariant to the feature map resolution. Specifically, in each PB at different pyramid and different level, we compute a new set of features $\mathcal{X}_t^{p,l}$ based on features from previous blocks and on the current prediction, as follows:

$$\mathcal{X}_t^{p,l} = \mathbf{W}_r^{p,l} * \mathbf{h}_t^{p,l} + \mathbf{W}_s^{p,l} * \mathbf{d}_t^{p,l} + \mathcal{Z}_t'^{p,l} + \mathcal{Z}_t^{p,l}$$

where $\mathbf{W}_r^{p,l}$ and $\mathbf{W}_s^{p,l}$ are weight matrices related to the reinjection of 2D pose and depth information, respectively. With this approach, further PB at different pyramids and levels are able to refine predictions, considering different sets of features at different resolutions.

3.3 Advantages of the Proposed Framework over Existing Methods

The proposed framework offers several advantages over existing methods. First, it allows for joint 2D and 3D pose estimation and action recognition, leveraging shared representations and complementary information across multiple tasks. Second, the Soft-argmax layer enables end-to-end training of the model, improving the accuracy of the pose estimation. Third, the incorporation of joint visibility information improves the accuracy of the pose estimation, especially in cases where joints are occluded. Finally, the proposed framework achieves state-of-the-art results on several benchmark datasets, demonstrating its effectiveness in joint pose estimation and action recognition tasks.

4 Datasets and Implementation Details

4.1 Datasets

The project utilizes two main datasets for evaluation: the MPII Human Pose Dataset and the Penn Action dataset.

MPII Human Pose Dataset

The MPII Human Pose Dataset is used for single person pose estimation and consists of approximately 25,000 images, with 15,000 training samples, 3,000 validation samples, and 7,000 testing samples. The images are sourced from YouTube videos covering 410 different human activities, and the poses are manually annotated with up to 16 body joints. This dataset provides a rich source of diverse human poses and activities, making it suitable for training and evaluating pose estimation models.

Penn Action Dataset

The Penn Action dataset comprises 2,326 videos capturing 15 different actions, such as "baseball pitch," "bench press," and "strum guitar." One of the challenges of this dataset is the presence of missing body parts in many actions and significant variations in image scales across different samples. This dataset provides a real-world representation of human actions, making it suitable for evaluating action recognition models.

4.2 Implementation Details

Pose Estimation Task

- For training the pose estimation task, an elastic net loss function is used, focusing on the n-th joint positions.

$$L_{\mathbf{p}} = \frac{1}{N_J} \sum_{n=1}^{N_J} \left(\|\hat{\mathbf{p}}_n - \mathbf{p}_n\|_1 + \|\hat{\mathbf{p}}_n - \mathbf{p}_n\|_2^2 \right), \tag{3}$$

  where $\hat{\mathbf{p}}_n$ and $\mathbf{p}_n$ are respectively the estimated and the ground truth positions of the n^{th} joint.

- Bounding boxes are cropped around the target person using either ground truth annotations or the person's location if available.

- Ground truth visibility flags are set based on whether a body joint falls inside the cropped bounding box during training. If outside, visibility is set to zero; otherwise, it is set to one.

- The ground truth visibility information is used to supervise the predicted joint visibility vector (v) using binary cross-entropy loss.

- Evaluation results are presented for both single-crop and multi-crop scenarios. Single-crop involves using one centered image for prediction, while multi-crop involves cropping multiple images with small displacements and horizontal flips, with the final pose being the average prediction.

Action Recognition Task

- For the action recognition task, the network is trained using categorical cross-entropy loss.

$$CE = -\sum_{i}^{C} t_i \log\left(f(s)_i\right) \tag{4}$$

- During training, fixed-size clips with T frames are randomly selected from video samples.

- During testing, results are reported for single-clip or multi-clip evaluation. Single-clip involves cropping a single clip in the middle of the video, while multi-clip involves cropping multiple clips temporally spaced $T/2$ frames apart. The final scores for multi-clip are computed by averaging results across all clips from one video.

- To estimate the bounding box during testing, an initial pose prediction is performed using full images from the first, middle, and last frames of a clip. The maximum box enclosing all initially predicted poses is then selected.

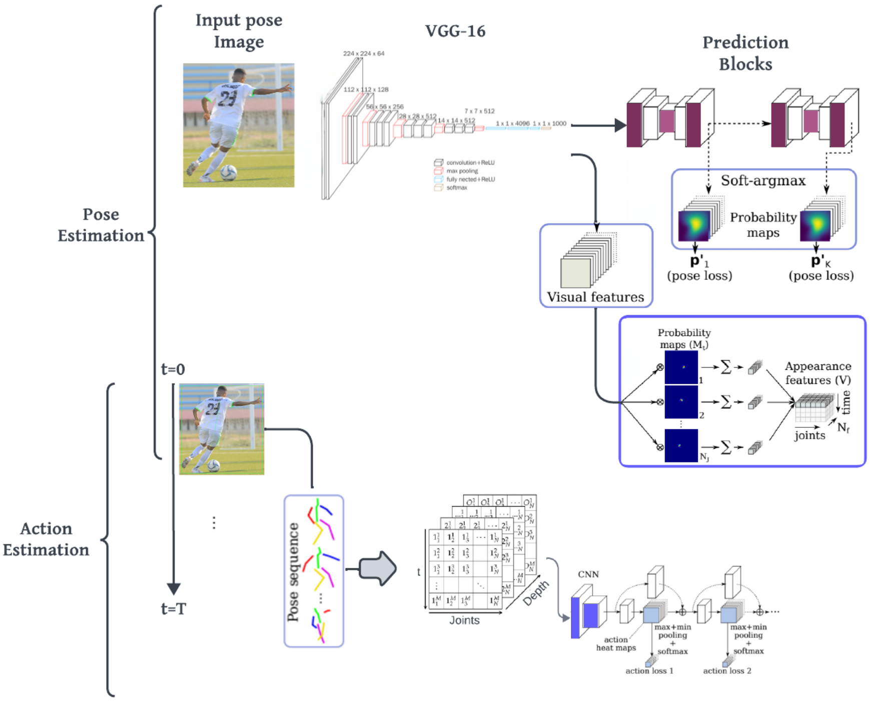


Figure 4: Unified Multitask Deep Architecture for Joint 2D and 3D Pose Estimation with Action Recognition

5 Results

5.1 Performance on Datasets

PennAction Dataset (Action Recognition)

- Achieved an accuracy of 44.10
- Highlights the need for further optimization to enhance accurate action recognition.

MPII Dataset (Pose Detection)

- Keypoint accuracy varied across body parts, with the highest in the neck (7.4) and the lowest in the left knee (1.0).
- Indicates a better performance in detecting upper body parts compared to lower body parts, with overall room for improvement in accuracy.

Our recently implemented architecture for sequential image data processing marks a significant shift from the previous model, introducing key characteristics:

[label=–]**Increased Initial Feature Channels:**

1. - The model begins with 1024 channels, a substantial increase from the previous architecture.
   - This higher channel count signifies a more complex network design, aiming to extract a richer set of features from each input frame.

2. **Complex Network Design:**

   - The augmented initial feature channels reflect a more intricate network design.
   - This complexity is intended to enhance the model's ability to capture nuanced features from sequential image data.

3. **Utilization of TimeDistributed Layers:**

   - The architecture continues to employ TimeDistributed layers, reinforcing its suitability for processing video and sequential image data.

4. **Progressive Increase in Feature Maps:**

   - A notable aspect is the consistent increment in the number of feature maps/channels as the network progresses (e.g., from 1024 to 1120).
   - This design choice is expected to improve the model's capacity to capture complex and nuanced features.

5.2 Factors Influencing Performance

- The model's performance was affected by a lower number of training epochs than ideally required.

- This limitation likely restricted the model's ability to fully learn and adapt to the complexities of the datasets.

- Utilizing a Saturn Cloud instance with a V100-2XLarge GPU encountered significant computational limitations.

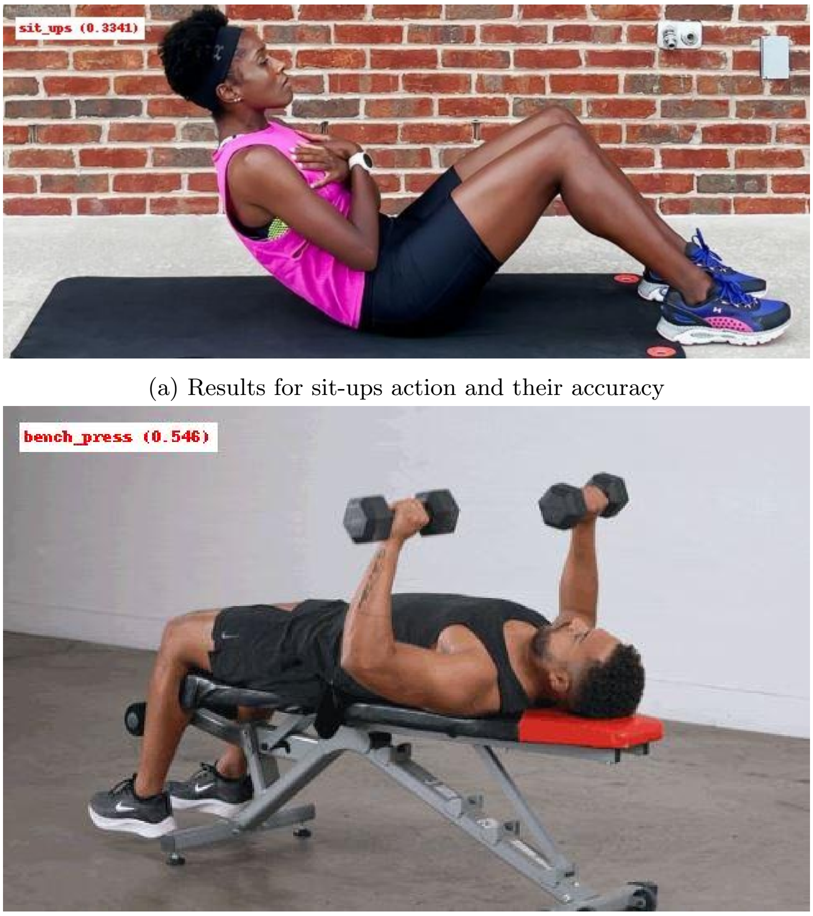
- Constraints in GPU capacity and associated costs restricted the exploration of more computationally intensive models, impacting the depth and extent of training.

- Strict project timelines resulted in a shorter experimentation and tuning phase.

- This limitation potentially led to suboptimal model performance due to reduced exploration of sophisticated models and depth of learning.

6 General Conclusion

This report presents a multitask deep architecture for joint 2D and 3D pose estimation along with action recognition. The proposed framework leverages the temporal evolution of body joint coordinates and performs attention-based pooling on human body parts to predict the action performed in the video. By sharing weights and features efficiently, the model addresses four different tasks - 2D pose estimation, 3D pose estimation, 2D action recognition, and 3D action recognition - using a single model. Extensive experiments have demonstrated that the proposed approach is capable of equaling or even outperforming dedicated approaches on all these tasks, as evidenced by the state-of-the-art results achieved across multiple datasets.



(a) Results for sit-ups action and their accuracy

(b) Results for bench-press action and their accuracy

Figure 5: Final Results for two different actions in Penn Action dataset

References

[1] 2D/3D Pose Estimation and Action Recognition using Multitask Deep Learning Diogo C. Luvizon1 , David Picard1,2 , Hedi Tabia1 1ETIS UMR 8051, Paris Seine University, ENSEA, CNRS, F-95000, Cergy, France 2Sorbonne Universite, CNRS, Laboratoire d'Informatique de Paris 6, LIP6, F-75005 Paris, France (link : Hedi Tabia's GitHub link for research paper)

[2] OpenPose: Realtime Multi-Person 2D Pose Estimation using Part Affinity Fields Zhe Cao, Student Member, IEEE, Gines Hidalgo, Student Member, IEEE, Tomas Simon, Shih-En Wei, and Yaser Sheikh (link : Openpose link)

[3] Multi-task Deep Learning for Real-Time 3D Human Pose Estimation and Action Recognition Diogo C. Luvizon Hedi Tabia David Picard (link : Last research paper with Mr.Tabia link)

[4] Mask R-CNN Kaiming He Georgia Gkioxari Piotr Dollar Ross Girshick ´ Facebook AI Research (FAIR) (link : Mask R-CNN link)

[5] VTP: Volumetric Transformer for Multi-view Multi-person 3D Pose Estimation Yuxing Chen, Renshu Gu , Ouhan Huang, Gangyong Jia (link : VTP link)

[6] W. Du, Y. Wang, and Y. Qiao. Rpan: An end-to-end recurrent pose-attention network for action recognition in videos. In The IEEE International Conference on Computer Vision (ICCV), Oct 2017

[7] Multiple View Geometry Transformers for 3D Human Pose Estimation Ziwei Liao1,* Jialiang Zhu2, Chunyu Wang3, Han Hu3 Steven L. Waslander1 1 University of Toronto 2 Southeast University 3 Microsoft Research Asia (link : MVGT link)

[8] C. Szegedy, S. Ioffe, and V. Vanhoucke. Inception-v4, inception-resnet and the impact of residual connections on learning. CoRR, abs/1602.07261, 2016. (link : Inception-v4 link)

[9]

[10] VERY DEEP CONVOLUTIONAL NETWORKS FOR LARGE-SCALE IMAGE RECOGNITION Karen Simonyan Andrew Zisserman Visual Geometry Group, Department of Engineering Science, University of Oxford (link : VGG 16 link)

[11] Two-Stream RNN/CNN for Action Recognition in 3D Videos Rui Zhao1 , Haider Ali2 , and Patrick van der Smagt3 Two-Stream RNN/CNN link)

[12] C3D: Learning Spatiotemporal Features with 3D Convolutional Networks (Video Classification Action Recognition) C3D link)

[13] J.Martinez, R. Hossain, J. Romero, and J. J. Little. A simple yet effective baseline for 3d human pose estimation. In ICCV, 2017

[14] D. Mehta, S. Sridhar, O. Sotnychenko, H. Rhodin, M. Shafiei, H.-P. Seidel, W. Xu, D. Casas, and C. Theobalt. Vnect: Real-time 3d human pose estimation with a single rgb camera. In ACM Transactions on Graphics, volume 36, 2017

[15] A. Bulat and G. Tzimiropoulos. Human pose estimation via Convolutional Part Heatmap Regression. In European Conference on Computer Vision (ECCV), pages 717–732, 2016. Heatmap Regression link)