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

Done by: *AMMAR KHODJA Rayane* **Supervised by:** *Hedi Tabia*

Introduction

Key Words: Pose Estimation, Action Recognition, VGG-16, Multitask Deep Learning, State Of The Art, Computer Vision, Pooling, ReLU, Soft-argmax, Elastic Net Loss, SGD, RMSprop.

Key Points:

1. Multitask deep learning architecture for joint 2D and 3D pose estimation and action recognition.
2. Presenting the Network architecture which is based on VGG-16, and into the parallelism between Pose estimation and Action recognition.
3. Project is based on research papers trying to work for the State Of The Art results.

Multitask Framework

- 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.

My Work: I critically analyze existing methods and clarify complex points. In this project I focused in Research paper **"2D/3D Pose Estimation and Action Recognition using Multitask Deep Learning"** [1] and made changes from **Invention V4** into **VGG-16.**

Experiments over the Project

The General Multitask Framework

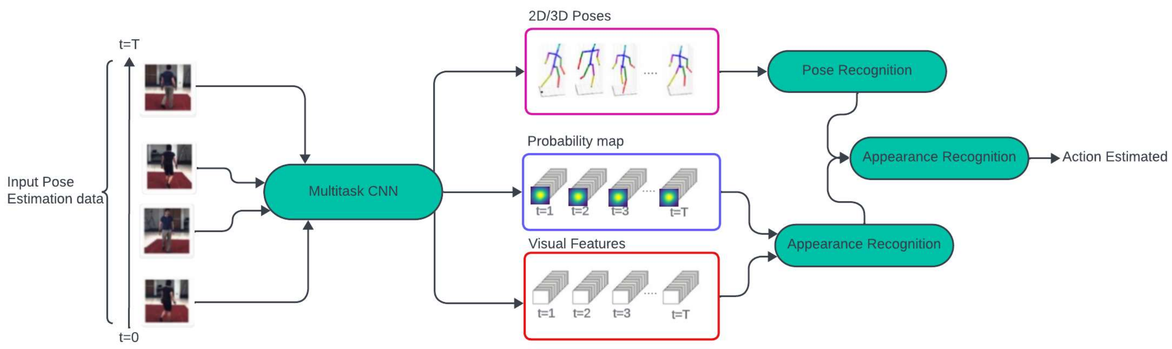


Figure.1 Multitask framework for joint 2D and 3D pose estimation and action recognition; which depict the flow of information through the network and how the model integrates information from RGB images to perform both tasks simultaneously.

Network Architecture

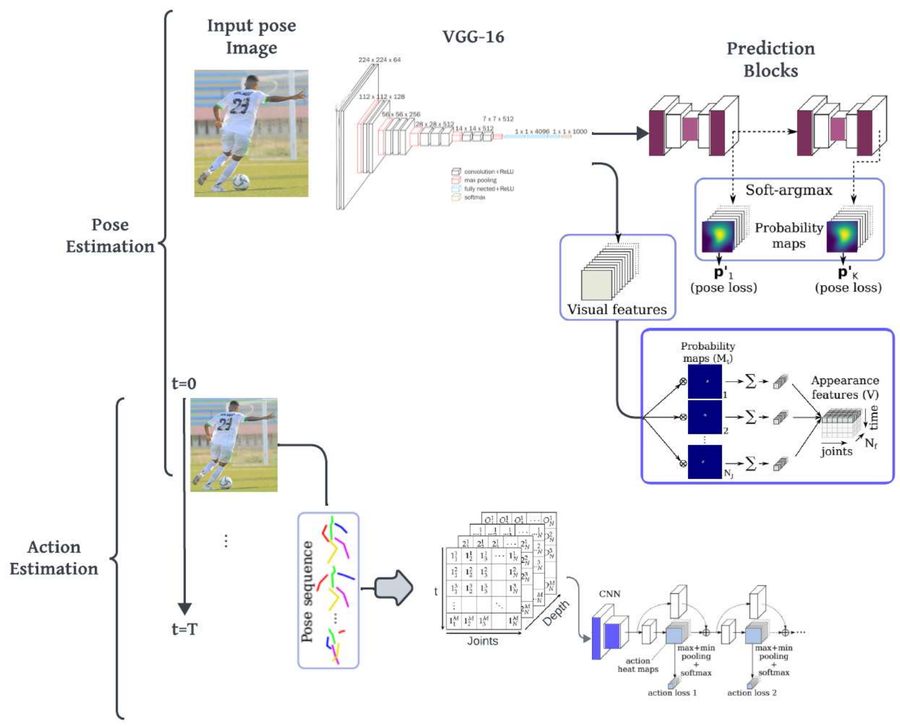


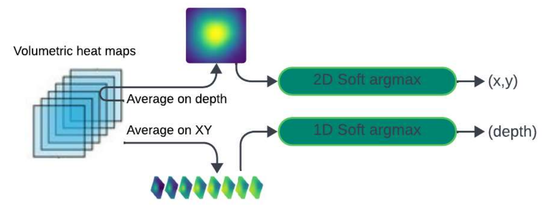
Figure.2 Unified Multitask Deep Architecture for Joint 2D and 3D Pose Estimation with Action Recognition.

Soft argmax Fuction

- 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

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

- 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.



Training

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

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

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

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

Data Sets

The project utilizes two main datasets for evaluation:

- The **MPII Human Pose** Dataset for Pose Estimation (25000 images, with 15,000 training samples, 3,000 validation samples, and 7,000 testing samples).
- The **Penn Action** dataset for Action Recognition (comprises 2,326 videos capturing 15 different actions).

Advantages

- Multitask learning facilitates the model in acquiring a shared representation, capturing common features across tasks and enhancing generalization.
- Jointly learning multiple tasks promotes better generalization, particularly for related tasks, leading to enhanced individual task performance, especially with limited data.
- Multitask learning acts as regularization, preventing overfitting by encouraging the model to learn representations beneficial for multiple tasks.

Results

(a) Bench press action with accurracy 54,6%

(b) Sit ups action with accurracy 33,41%

Figure.1 Some results of the Penn Action dataset with their accuracy in the top left, here we can see some difficulties in recognition and most of it is related to some anomalies in the knees, most of the best results for action recognition is focused on the parts around the neck and bellow the chins.

Conclusion

- The framework employs attention-based pooling on body parts, leveraging shared weights to address four tasks: 2D pose estimation, 3D pose estimation, 2D action recognition, and 3D action recognition with a single model. Extensive experiments demonstrate the proposed approach's ability to match or surpass dedicated approaches, yielding state-of-the-art results across various datasets.

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,