**Human Pose Tracking and Mask Generation in Video Sequences**

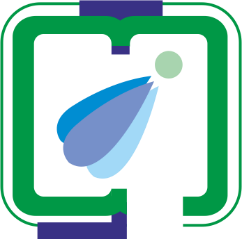
***By***

Gursimar Singh

(Roll No. 2015101)

**Supervisors:**

|  |  |
| --- | --- |
| Internal  Dr. Atul Gupta  Associate Professor  PDPM IIITDM Jabalpur | External  Mrs. Sheena KP  Senior Manager  Canon ISDC, Bengaluru |



Electronics & Communication Engineering, B.Tech 2015

INDIAN INSTITUTE OF INFORMATION TECHNOLOGY, DESIGN AND MANUFACTURING, JABALPUR

29h May, 2018 - 10th July, 2018

# **Introduction**

Human extraction and segmentation in natural environments is one of the most discussed problems of modern times in computer vision. Many pedestrian detectors gained both popularity and accuracy after state-of-the-art object localization deep learning algorithms were introduced. But a bounding box is not enough for many applications and thus pixel-wise segmentation is required. During the tenure of this report, I worked on this problem and explored and experimented various conventional and recent image processing and computer vision algorithms. The task of human mask generation can be eased if we have information about the pose of the human to be segmented. In the process to generate the mask, first task is to extract the person of interest from video sequences. Second task is to identify essential key points of target human and then constructing a skeleton mask which could be used as a seed for mask generation.

# **Progress**

During the period of this report I have worked on various image processing and deep learning techniques for mainly accomplishing two tasks:

1. **Tracking person of interest and its essential key points for pose in video sequences**
2. **Human full body segmentation**

**1.1 Tracking person of interest**

The first task is to create a ROI for the person be segmented. A lot of work has been done in pedestrian tracking but they fail on persons in complex poses, therefore they could not be regarded as pedestrians. A lot person detection models are available like HOG+SVM, Mask R-CNN etc. Mask R-CNN is state of the art object detection deep learning model.

* + 1. **Approach**

A feature matching approach was used to track the person. The convention features like SIFT, SURF, ORB etc. did not provide dense correspondences. Therefore, a multi-scale convolutional based feature matching approach was used. The extra background features are removed using velocity mean based segmentation with the help of deep flow.

**1.2 Keypoint Tracking**

The essential key points of human body define the pose of the human. A lot of work is being done on pose detection of humans using deep learning algorithms. The keypoints detected using the deep learning have to be refined as it fails in complex poses, low resolution images etc. Thus, assuming the keypoints of the first frame are correct, the keypoints are tracked in the subsequent frames using deep matching. The reference frame is reinitialized when the number of keypoints become less than a certain number or is there is a drastic change in the location of the keypoints.

**2.** **Human mask generation**

Pixel-wise segmentation has gained popularity after the introduction of U-Net style deep learning architectures. But the recent works concentrate on a number of semantic classes and does not incorporate pose or keypoint information for mask generation, thus are not very accurate. I till now have worked on various image processing techniques in order perform this task and currently exploring deep learning techniques or a way design an algorithm combining both of them. Two major algorithms have been tested using the ensemble of the following techniques:

* **Weighted Graph Cut algorithm**
* **Simple Linear Iterative Clustering (SLIC) superpixel segmentation**
* **MLE using Color based Gaussian Mixture Models**
* **Color based region growing**

The essence of the algorithms is same and they are focused on removing as much background region as possible. A mask is to be developed to give it input to the graph cut algorithm. A skeleton mask generated by interconnecting the keypoints is available and an area of sure background is known by expanding the skeleton mask to the extent when no body part is being labelled as background.   
**Approach 1:**

An example of the SLIC algorithm is shown in Fig 2.1. The intersection of the superixels clusters and the skeleton mask generate a grown mask for the image Thus, the generated mask is given as seed to graph cut algorithm.



Fig 2.1: Super pixel clustering

**Approach 2:**

The problem with the above algorithm is that some of the limbs or body parts are also labelled as background thus the output of the above approach is fed into the region growing algorithm. Gaussian Mixture Model is them made out of the sure background pixels and then the probability of all the candidate foreground pixels is computed to lie in the background. GMM clustering can be done two ways, i.e. pixel-wise and superixel-wise. In superpixel GMM clustering mean, median and standard deviation is computed for RGB values of each superpixel. These three features are used as one descriptor in the GMM model.

# **Result and Discussions:**

During the duration of this report, I worked on the following tasks namely human tracking, body key point tracking and human mask generation. They are cases when they fail and there is a scope of improvement.

**Person Tracking:**

The result of the person tracking approach were satisfactory. The tracking is good when the frames are continuous but fails when there are missing frames.

**Further improvements**

* Find way to detect false detection and reinitialize bounding rectangle either by manual intervention or some algorithm.
* Make use of variance of the motion information to improve segmentation of foreground and background

**Human Key Point labelling:**

The results are satisfactory, but there is an offset error which needs to be compensated.

During the duration of this report I worked on finding a solution to, combine all the mentioned techniques in some order to get fine results for mask generation. The generated masks with current approach and flow are very accurate in cases with homogeneous backgrounds. In cases with non-homogenous backgrounds, results are not very fine at the edges and also lose some foreground in some cases.

**Further improvements**

* Find an efficient order in which these techniques can be used
* Trying including more features in GMM model for segmentation

**Human Mask Generation**

The generated masks with current approach and flow are very accurate in cases with homogeneous backgrounds. In cases with non-homogenous backgrounds, results are not very fine at the edges and also lose some foreground in some cases. There is also a problem of the foreground being labelled as background.

**Further improvements**

* Explore some deep learning models for the purpose like the Joint Pose and Parsing network
* Efficient Graph cut implementation like Cooperative cut which has a discounting function that helps to preserve thin limbs.

# **Conclusion**

The human mask generation techniques used to till now depend a lot on the variation of the background thus is not very reliable in many cases. During the term of this report I learnt a lot about various segmentation techniques and their working. I also learnt about the motion tracking and features extraction algorithms. I also analyzed their results in various cases and scenarios.