GAIT Analysis: 3D Pose Estimation and Prediction in Defence Applications using Pattern Recognition

B Kiran Kumar Ashish[1], Manoj[2], Ashwik Sagi[3], V Pooja Nandana[4,] Kalyan Reddy[5]

Anurag group of Institutions, Hyderabad, India

{15h61a04j3, 16h61a1295, 16h61a0461, 16h61a04b6, 16h61a0491}@cvsr.ac.in

Abstract

The latest advancements in Computer Vision have enabled major solutions for numerous applications in Defence research and commercial aspects. The existing solutions for human action recognition have been into deployments but the predicting of next course of pose is not accurate in the existing solutions. The existing methodology can't be defined on any edge device which can be deployed in a real-time scenario. This paper analyses the pose of the object using Computer Vision using 3D plots for better analytical reports which can further deployed on any edge device. This paper describes combination of LSTM and top-down approach. This uses a normal HD camera and no special sensor or stereo camera is required to capture the image. This paper proposes a unique method to analyse a large corpus of crowd videos by segmenting each point in the pose and plot it in a 3D plot. The analysis was performed particularly on stone pelting videos and massive crowd videos. Pose estimation for people in the video achieved 95% accuracy (mAP >0.5).

Keywords: Computer Vision, 3D plots, LSTM, top-down approach, pose.

1. Introduction

Human action recognition [1, 4] through Pattern Recognition have received an important attention in the modern Deep Learning era [1, 4, 11, 12]. Pattern Recognition in image and video analytics gathers the pose from the targets and connects the patterns that lead to a final action summary. Pose estimation and action recognition are usually handled as distinct problems in sports [1, 10, 16] and Defence applications. Even though pose is of extreme relevance for action recognition, to the best of our knowledge, there is no method in the literature that solves both problems in a joint way to the benefit of action recognition and predicting the next sequence of action that can take place based on previous and current pose [1]. The existing methodologies couldn't able to predict or estimate the pose in a huge crowd.

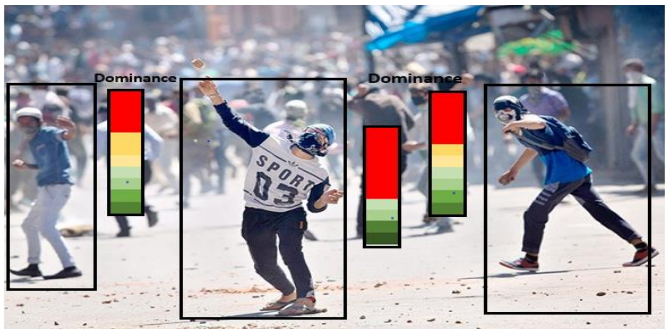

Fig 1: Aiming at automatically to detect the dominance level of the stone pelters through simple camera.

Fig 2: Dominance levels of normal people in a crowd

Fig (1) and (2) describes our proposed methodology of showing the dominance levels of various persons in a crowd. The red levels indicate the alert signal and the green defines the normal pose.

Computer Vision has been playing an important and crucial role in Defence analytics. The images collected from various sources made their memory much larger than before which can be very useful for computer vision algorithms for analytics. Applications of Computer Vision in Defence analytics are pose estimation and prediction [1], face recognition and tracking, video summarization [1, 4, 12], object detection like bombs, grenades, landmines and likewise and much more. Today there are numerous videos available on thefts, criminals, stone pelters and so on. These are generally short videos and may vary with circumstances. These maybe CCTV footage or TV news agency streaming videos. All of these have multiple viewpoints. These can be efficiently utilised to train the model with multiple viewpoints for better efficiency of the model. Even today, most of the surveillance videos are manually monitored 24x7 by the experts and most of them are aftermath incidents which is expensive in nature. Since, it becomes hectic tasks for any human to consistently monitor huge crowds at once and detect abnormalities between them, a unique method is proposed which detects and predicts the abnormalities which works even in crowded places using RMPE architecture [12]. Deep Learning techniques have enabled wide range of tasks such as action recognition [10, 11], object detection and tracking, semantic segmentation, image segmentation into an increased performance ratio with precise accuracies. The most specific study of human motion by using the observer's brain and eye which is used to enlarge the measure of body movements and the muscle activities called GAIT analysis is much useful in Defence analytics. The movement of body pose will be unique for each person and differs from others. This data can be used to identify *"the one in a million crowds".*

CCTV footage and news feed often pose greater challenges like blur footage, angle view, resolution or zooming. This will be a major challenge in automated analysis. The human movements are very quick and lasts for seconds. Sometimes, camera awareness is also pre-cautioned by them and try not to reveal themselves where in most cases face recognition fails. Since, pose estimation works where face recognition fails. The estimated time for pose estimation for a normal video footage is approximately *5 seconds*. Tracking poses in different view for each CCTV or video footage taken from different camera angles poses much more difficulties. The crucial points for getting the pose are their joints. The camera angle plays a crucial role in detecting the joints [1, 10, 11]. Therefore, topmost view was discarded and have trained narrow and wide-angle view footage [1].

2. Related Work

2.1 Computer Vision in Defence Applications: Computer Vision and Pose detection has its significant role in Defence related fields. This work deals with analysis of stone pelters Kashmir, India and Israel-Palestine boarder. The video feed was mostly taken from You Tube and several News channels videos in both India and Israel. The data was trained by annotating the action and object by object detection and smart image segmentation. The stone pelters pose a major problem to the security personnel. Identifying them in a crowd is a bit complex task as they often cover their faces with masks, where face recognition

model fails. Anurag Ghosh et al. [1] proposed each pose segment for each stroke played in badminton sports. Each stroke and its position are analysed and best position for stroke was then analysed. Similar approach by calculating the trajectories from their pose is analysed and tracked further based on his pose tracking. Most violent pose structures are analysed for future advancements. As not much research has been done on Defence use cases because presence of greater complexity remains as main challenge. Now, as the data is available of specific test cases, a novel method is proposed to detect their action and track them based on their pose. Goerg Poier et al [9] proposed hand pose [9, 3] estimations using Autoencoders, but the challenging task is to detect the hand pose in a huge crowd. It tries to achieve that hand pose [3] separately by just focussing on the hand part by isolating it with the vicinity using GrabCut method [3]. The matrix which models the transformation of objects in 3D plot [10, 11] is $([x\_, y\_, z\_]F[x, y, z]^T = 0)$, which is used for 3D visualization for better accuracy and training in all possible angles as shown in Fig (3).

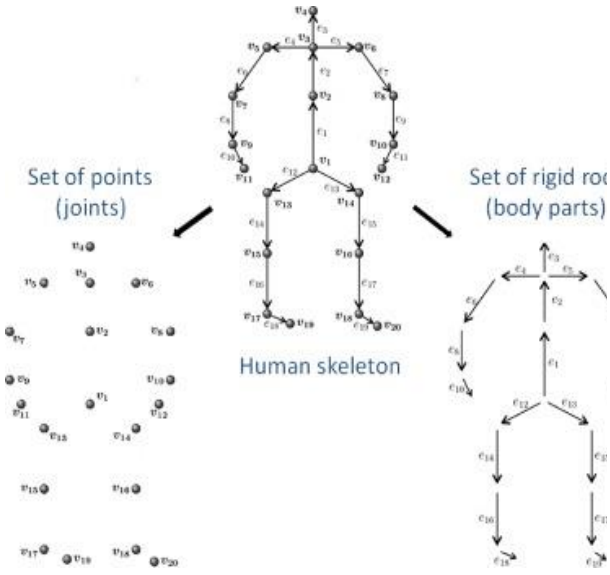

Fig 3: Shows how 3D visualization works mathematically for plotting locomotion

2.2 Action Recognition through Pattern Recognition: A hybrid method is used for training a specific data for single and multi-pose estimation. The analytics can be applied on a single person itself for better analytical report on him/her. Smart segmentation using GrabCut method [3] isolates the surroundings and masks only required object using OpenCV. This output can be applied for pose estimation and a detailed report on his action performance such as body angle, movement, alertness, walking style, joint stress and likewise psychological factors can be estimated through pose. Suppose a child is lost in a huge crowd such as *Kumbh* or *Hajj*. This child will be in a lost state and will not have control on his body and its movement. He will be in fear and very much scared such that he will be searching for his beloved one's in the surroundings like a lightening jolt which is very much different from a normal search. These pose patterns can be combined together and thus results in a classification of lost person or normal person in a crowd. These pose patterns can be used to easily identify to trace lost persons in any crowd and then can perform face recognition for confirmation. Initial step of face recognition in a crowd would fail the model since not all faces turn up to the cameras and not all cameras are at all angles

This work combines these patterns and analyse in a 3D plot to analyse in all ways that gives us precise accuracy while testing in a crowded area.

2.3 Smart Image Segmentation: Smart Image Segmentation is taken for training the dataset for specific test cases. Image segmentation is performed on single stone pelters and lost children in a crowded area are trained with respect to all angles possible. Carsten Rother et al [3] proposed the GrabCut algorithm [3] which does the

background and foreground subtraction and boarder matter. Smoothing regularizer is used for accurate boarder cut.

$$\mathbf{E} = \sum_{n \in T_U} \boldsymbol{D}^{\sim}(\boldsymbol{\alpha_n}) + \sum_{t=1}^{T} \boldsymbol{V}^{\sim}(\Delta_t, \boldsymbol{\sigma_t}, \Delta_{t+1}, \boldsymbol{\sigma_{t+1}})$$

This energy function is minimized over t for boarder matting, where V˜ is smoothing regularizer given by:

$$\boldsymbol{V}^{\sim}(\Delta, \boldsymbol{\sigma}, \Delta', \boldsymbol{\sigma}') = \boldsymbol{\lambda_1}(\boldsymbol{\Delta} - \boldsymbol{\Delta}')^2 + \boldsymbol{\lambda_2}(\boldsymbol{\sigma} - \boldsymbol{\sigma}')^2$$

2.4 Detecting and Tracking: For specific pose for specific persons image segmentation was applied and their poses were trained for a class 'A' action in different scenarios. We detect the abnormal actions such as missing in crowd, stone pelting and track them based on their pose. This work is to detect the abnormalities in a huge crowd and track them based on their pose through pattern recognition and computer vision techniques.

The major contributions of this paper are:

1. An end-to-end inference for pose estimation and predict the pose for next 4-5 frames based on current and past frames.
2. Object detection is done using Transfer Learning and then pose is estimated within that bounding box.
3. Smart segmentation using GrabCut algorithm [3] is used for image segmentation where only one specific object is obtained by background subtraction and apply pose analysis for it.
4. Analysed various metrics on several use cases of several persons which gives analytical report to the Defence services and hence, qualitative analysis on war-prone zones and crowded areas.
5. Collection of large variety of dataset and isolated from background for several images for getting high-level understandable pose for specific use cases which will be discussed later.

3. Proposed Methodology

In this section, architectural flow is explained in detail. Single and multi-pose plots in 3D view is proposed [13] for defence analytical reports. The key part in our architecture is the dataset and pre-processing. The dataset varies from specific poses from specific targets. Image segmentation is used for getting a specified person and estimate his pose and then train for those class 'A' pose category because the specific pose data is rare to obtain. Since, we are dealing with stone pelters and missing children in a huge crowd, the data clips would be small. Hence, we are specially dealing with collecting the data in multi-viewpoint and labelling them inti class 'A'.

3.1 Dataset Collection: We worked on collection of news footage and CCTV footage on stone pelters from Kashmir, India and Israel-Palestine clashes. The footage is short and very much noisy and has multi-angle views which is very unstructured. Therefore, we are refining the data and discarding which is very unstructured which gives no output. We are using pre-trained COCO dataset for better accuracy. Concatenating pre-trained data model with our custom dataset model saves time as well as gives better accuracy.

3.2 Data Pre-Processing: As data pre-processing is the basic brick of the model, it is a key player in the framework. The video footage is saved into frames using OpenCV. We extracted *10 frames* per second such that there would be no duplication issue and same pose wouldn't be repeated which is unnecessary. Since, the human movement is not stable, 10 frames are extracted from video such that each pose would last for at least 10 seconds

and we got 100 frames from a 10 seconds video which is visualized in 3D plot from different viewpoints and trained it from all possible angles as similar to data augmentation.

For focussing on one specific target in a crowd, image segmentation was applied so that we just get the target image and rest is subtracted. We can apply pose for it and visualise the exact pose in multiple angles. This is much useful when we have a very rare data available and is much crucial for visualizing it.

3.3 Algorithms

Pipeline: Our Symmetric STN consists of STN and SDTN which are attached before and after the SPPE. The STN receives human proposals and the SDTN generates pose proposals. The Parallel SPPE acts as an extra regularizer during the training phase. Finally, the parametric Pose NMS (p-Pose NMS) is carried out to eliminate redundant pose estimations [10, 12]. Unlike traditional training, the SSTN+SPPE module was trained with images generated by PGPG [12].

Symmetric STN and Parallel SPPE: Human proposals provided by human detectors are not well-suited to SPPE. This is because SPPE is specifically trained on single person images and is very sensitive to localisation errors. It has been shown that small translation or cropping of human proposals can significantly affect performance of SPPE [10, 12]. Our symmetric STN + parallel SPPE was introduced to enhance SPPE when given imperfect human proposals [12]. The module of our SSTN and parallel SPPE is shown in Figure. STN and SDTN The Spatial Transformer Network (STN) has demonstrated excellent performance in selecting region of interests automatically. In this paper, we used the STN to extract high quality dominant human proposals. Mathematically, the STN performs a 2D affine transformation which can be expressed as

$$\begin{pmatrix} x_i^s \\ y_i^s \end{pmatrix} = [\boldsymbol{\theta_1} \quad \boldsymbol{\theta_2} \quad \boldsymbol{\theta_3}] \begin{pmatrix} x_i^t \\ y_i^t \\ \mathbf{1} \end{pmatrix}$$

Where θ1, θ2 and θ3 are vectors in

$\mathbf{R^2}$**.** $\mathbf{\{x^s_i, y^s_i\}}$ **and** $\mathbf{\{x^t_i, y^t_i\}}$ are the coordinates before and after transformation, respectively. After SPPE, the resulting pose is mapped into the original human proposal image. Naturally, a spatial de-transformer network (SDTN) is required to remap the estimated human pose back to the original image coordinate [12]. The SDTN computes the γ for de-transformation and generates grids based on γ:

$$\begin{pmatrix} x_i^t \\ y_i^t \end{pmatrix} = [\boldsymbol{\gamma_1}\ \boldsymbol{\gamma_2}\ \boldsymbol{\gamma_3}] \begin{pmatrix} x_i^s \\ y_i^s \\ \mathbf{1} \end{pmatrix}$$

Since SDTN is an inverse procedure of STN, we can obtain the following:

$$[\boldsymbol{\gamma_1} \quad \boldsymbol{\gamma_2}] = [\boldsymbol{\theta_1} \quad \boldsymbol{\theta_2}]^{-1}$$

$$\boldsymbol{\gamma_3} = -1 \times [\boldsymbol{\gamma_1} \quad \boldsymbol{\gamma_2}]\, \boldsymbol{\theta_3}$$

Parallel SPPE: To further help STN extract good human dominant regions, parallel SPPE branch was added in the training phrase [10, 12]. This branch shares the same STN with the original SPPE, but the spatial de-transformer (SDTN) is omitted. The human pose label of this branch is specified to be centred. All the layers of this parallel SPPE were freeze during the training phase. The weights of this branch are fixed, and its purpose is to back-propagate centre-located pose errors to the STN module. If the extracted pose of the STN is not centre-located, the parallel branch will back-propagate large errors. In this way, we can help the STN focus on the correct area and extract high quality

human-dominant regions. In the testing phase, the parallel SPPE is discarded. The effectiveness of our parallel SPPE will be verified in our experiments [12].

Pose NMS: Human detectors inevitably generate redundant detections, which in turn produce redundant pose estimations. Therefore, pose non-maximum suppression (NMS) is required to eliminate the redundancies [12].

Pose Distance: Now, we present the distance function *dpose*(P_i , P_j). We assume that the box for P_i is B_i. Then we define a soft matching function [1, 10, 12, 14].

$$\mathbf{K_{Sim}(P_i , P_j \mid \sigma_1)} = \sum_n \boldsymbol{tanh}\frac{c_i^n}{\sigma_1} \boldsymbol{.\ tanh}\frac{c_j^n}{\sigma_1}, \text{ if } k_j^n \text{ is within } B(k_i^n)$$

$$\mathbf{0}, \text{ otherwise}$$

Where $B(k_i^n)$ is a box centre at k_i^n, and each dimension of $B(k_i^n)$ is 1/10 of the original box B_i. The tanh operation filters out poses with low-confidence scores. When two corresponding joints both have high confidence scores, the output will be close to 1. This distance softly counts the number of joints matching between poses.

The spatial distance [12] between parts is also considered, which can be written as

$$\boldsymbol{H_{sim}}\ \mathbf{(P_i, P_j \mid \sigma_2)} = \sum_n \boldsymbol{exp}\left[-\frac{(k_i^n - k_j^n)^2}{\sigma_2}\right]$$

By combining both the equations, the final distance function can be written as

$$\mathbf{d\ (P_i, P_j \mid \Lambda)} = \boldsymbol{K_{sim}}\ \mathbf{(P_i, P_j \mid \sigma_1)} + \boldsymbol{\lambda . H_{sim}}\ \mathbf{(P_i, P_j \mid \sigma_2)}$$

Where λ is a weight balancing the two distances and $\Lambda = \{\sigma 1, \sigma 2, \lambda\}$. Note that the previous pose NMS set pose distance parameters and thresholds manually. In contrast, our parameters can be determined in a data-driven manner [12].

3.4 Multi-Person Pose Estimation

Multi-Person pose estimation is more difficult than the single person case as the location and the number of people in an image are unknown. Typically, we used top-down method by incorporating a person detector initially using transfer learning, followed by estimating the parts and then calculating the pose for each person. Our work is much easier to implement than the other approaches as we are adding a person detector is much simpler than adding associating or grouping algorithms. The model is much dependent on the accuracy of the person detector, as pose estimation is performed on the region where the person is located. Hence, errors in localization and duplicate bounding box predictions can cause the pose extraction algorithm to perform sub-optimally. To resolve this issue, we are using Symmetric Spatial Transformer Network (SSTN) which extracts a high-quality single person region from an inaccurate bounding box. A Single Person Pose Estimator (SPPE) [11, 12] is used in this extracted region to estimate the human pose skeleton for that person. A Spatial De-Transformer Network (SDTN) [10] is used to remap the estimated human pose back to the original image coordinate system. Finally, a parametric pose Non-Maximum Suppression (NMS) technique is used to handle the issue of redundant pose deductions.

Furthermore, we introduce a Pose Guided Proposals Generator to augment training samples that can better help train the SPPE and SSTN [10, 11, 12] networks. The salient feature of RMPE is that this technique can be extended to any combination of a person detection algorithm and an SPPE.

4. Experimental Results:

We have initially performed detection for stone pelting and normal civilian stuck there based on pose as Zhu et al [11]

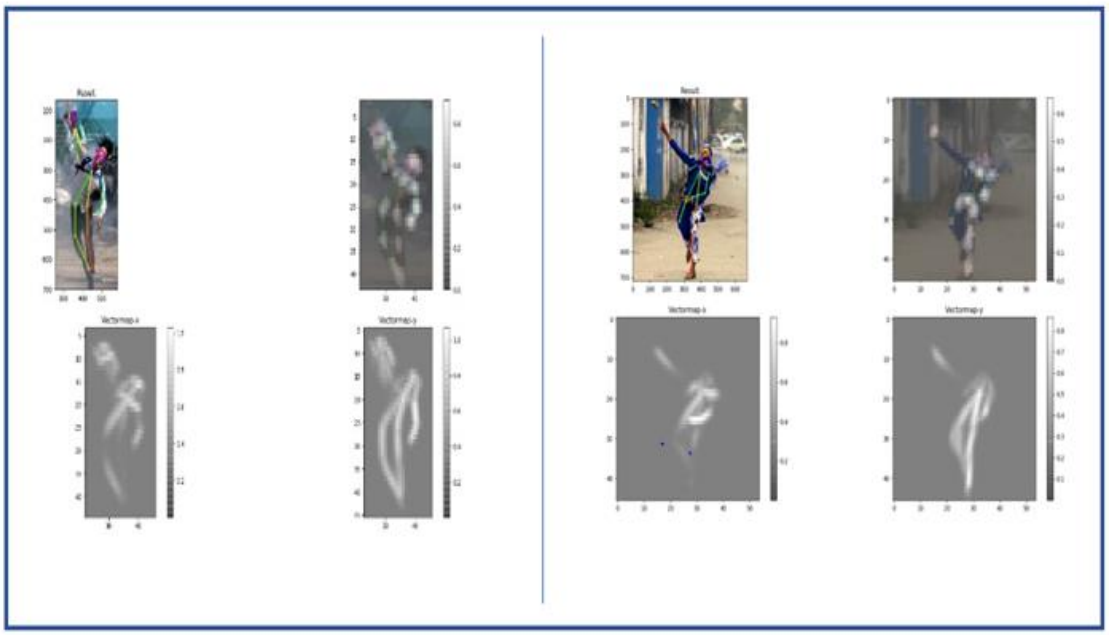

Fig 4: Target is focussed, and pose is estimated along with vector heatmap

performed for badminton [1] in which they classified as offensive or defensive.

We have taken the dataset which is available in the You Tube and several news videos feed and documentaries on stone pelting in Kashmir and Israel. The challenging task for us it to differentiate normal people and stone pelting from a mob. Therefore, we were just focussing on the subject target by isolating him/her from the background and plot the pose. This pose is taken account into class 'Lost' and for training data and then the model is trained based on the labelled pose. Same methodology was used for identifying lost child in a crowd or street. We isolated the child and plotted the pose of the child to label it and trained the model, versus the normal children. The code is written in Python and open source library TensorFlow which is used to plot the pose. We used YOLO object detection to detect the persons in a frame and GrabCut method [3] which uses OpenCV used for image segmentation to isolate the object from the background.

The vector heatmap is plotted to check out the dominance intensity of the subject target which will differentiate into respective classes such as 'Normal Pose' or 'Stone Pelting' or 'Lost Child' as shown in Fig (4).

These pose patterns can be connected and then fed to the model which then performs predictive analytics on the current poses as shown in Fig 5. These patterns can accurately predict the estimated pose for next 10 frames such that detection + prediction analytical report can precisely detect whether the target is suspicious or not. Getting a particular subject target is very challenging task hence we have used GrabCut method [3] for smart segmentation where we just get the target object and can isolate the rest in the foreground and background. We can apply 3D plots on each one of them and perform a detail analytical report on their pose/locomotion geometry. Fig (5) describes the Prediction of leg motion based on present leg motion.

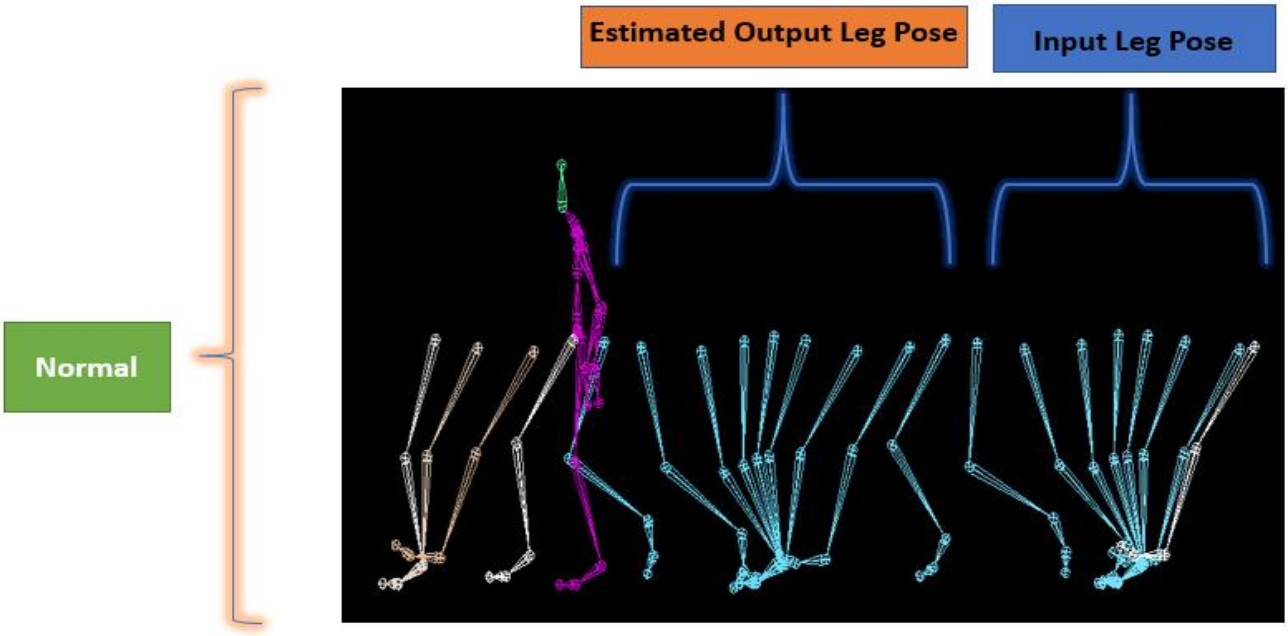

Fig 5: Locomotion of predictive movement of leg based on past and current pose data

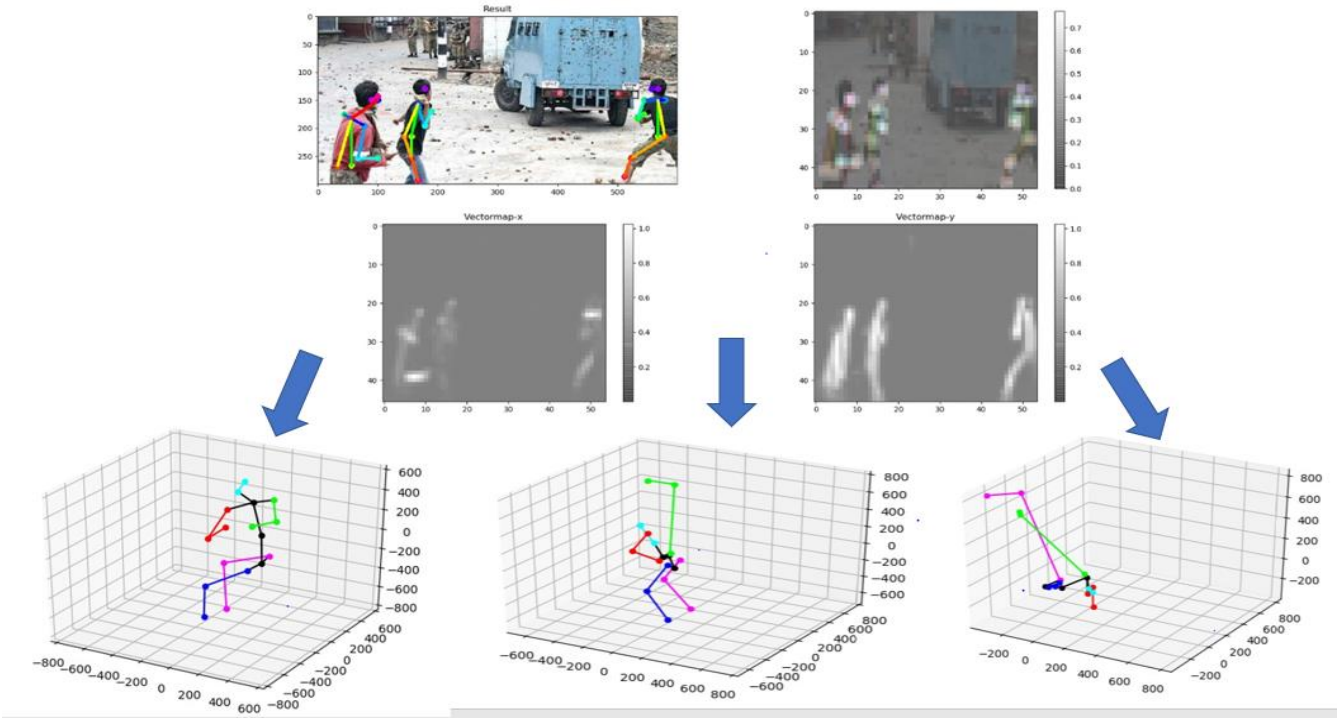

Fig 6: Snaps of Detection of Stone Pelters and the 3D plot view along with vector heatmap

Fig (6) describes the 3D plot and vector heatmap of stone pelters which detects the dominance level and further predicts the pose if class is abnormal segment.

Top-down method is used after person detection using YOLO object detection. Since, we are isolating the targets from the background and pose estimation is based on the local region where the person is located, Single Person Pose Estimator (SPPE) is used in this extracted region to estimate the human pose skeleton for that person. This is based on the centre coordinates, where the joints are mapped from the centre and a Spatial De-Transformer Network (SDTN) is used to remap the estimated human pose back to the original image coordinate system. Finally, a parametric pose Non-Maximum Suppression (NMS) is used for redundant pose deductions. Since, we are using YOLO object detection for person, no other objects cab be detected and where pose of other objects need not be an issue. This is done by changing the coco dataset labels to "Don't_Show 'object' " in the *labels.txt* file for all other objects except for person. 3D plots are visualized after applying affine 2D transform and then converting it into 3D plots by computing $([x\_, y\_, z\_]F[x, y, z]\tau = 0)$ matrix as mentioned in section 2.1.

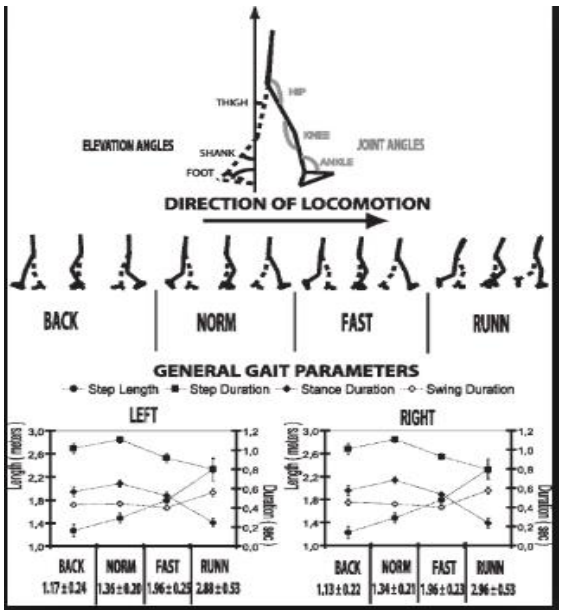

Fig 7: Geometry Locomotion on GAIT Cycle

Fig (7) shows the geometry locomotion on GAIT analysis on the human posture in different operational parameters taken from Bio-Medical research report on locomotion. This shows the difference of pose gestures of different parameters which can differentiate even in a huge crowd. The centre coordinates are fixed to the target where the pose distance is calculated and sets to fix the joints and then by combining them together for getting a pose gesture. Applying transformation on this 2D vectors to 3D vector, gives us to view the pose from 360-degree angle.

Table 1 given below, describes the accuracy comparison of different methodologies used in GAIT analysis. The method used by us, RMPE proves to give highest accuracy when compared to other techniques.

Fig (8) describes the pose estimation of each target object (person in our case) where we have used image segmentation to isolate the targets and plot pose individually which can be viewed in a 3D plot.

The other techniques were focussing on a single segment, but our method focusses on all critical segment aspects such as hand, body and movement. We extracted each frame and isolated each target for getting accurate pose pattern. Hence, with the help of existing methods such as RMPE, we outperformed the remaining methods with a higher accuracy result.

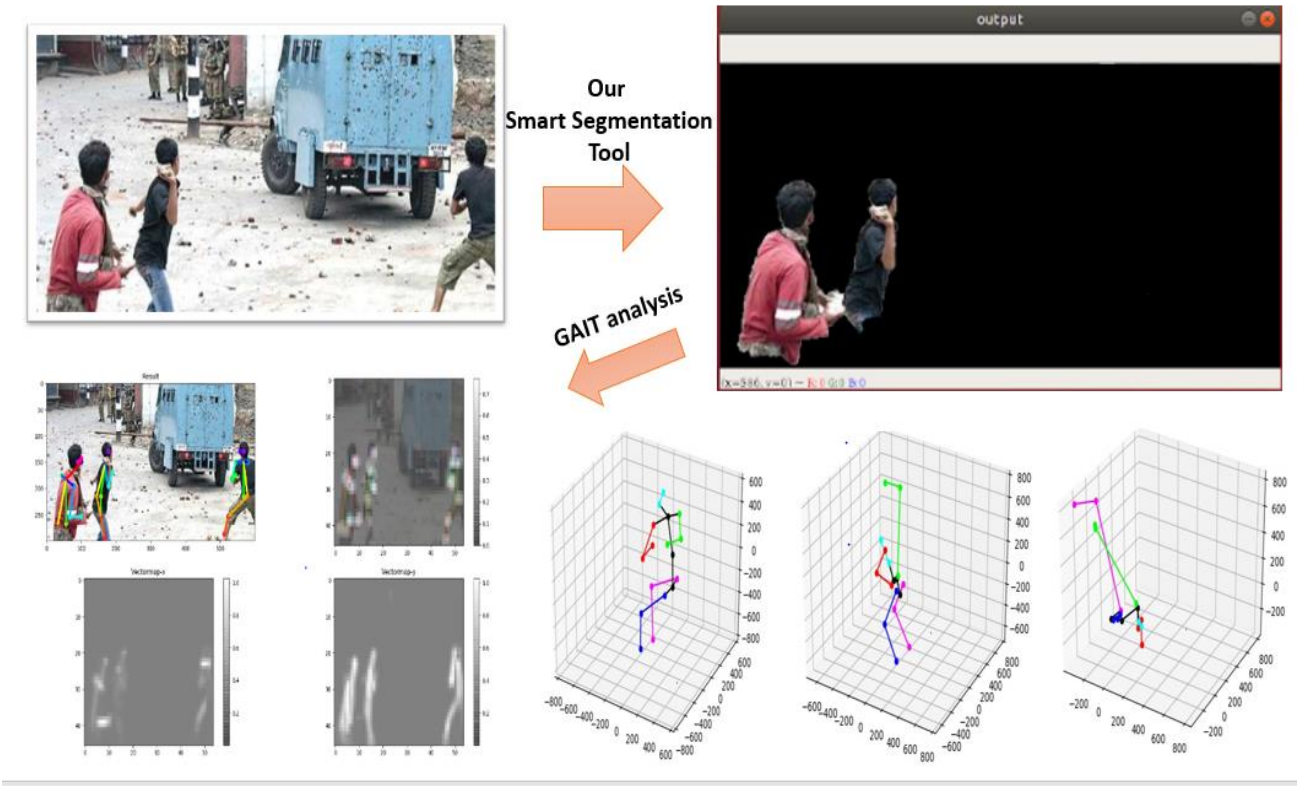

Fig 8: Final Output of Pose Estimation (Proposed Method: along with image segmentation shown in the second block)

Methods	Head	Shoulder	Elbow	Wrist	Hip	Knee	Ankle	Total
RMPE, Full	**90.7**	**89.7**	**84.1**	**75.4**	**80.4**	**75.5**	**67.3**	**80.8**
w/o SSTN+Parallel SPPE	89.0	86.9	82.8	73.5	77.1	73.3	65.0	78.2
w/o Parallel SPPE only	89.9	88.0	83.4	74.7	77.8	74.0	65.8	79.1
w/o PGPG	82.8	81.0	77.5	68.2	74.6	66.8	60.1	73.0
Random Jittering	89.3	87.8	82.3	70.4	78.4	73.3	63.8	77.9
w/o PoseNMS	85.1	83.6	79.2	69.8	76.4	72.2	63.6	75.7
PoseNMS	88.9	87.8	83.0	73.8	78.7	74.6	66.3	79.1
PoseNMS	90.0	88.6	83.7	74.6	79.7	75.1	67.0	79.9
Straight forward two-steps	81.9	80.4	74.1	68.5	69.0	66.1	62.2	71.7
Oracle human detection	94.3	93.4	87.7	80.2	84.3	78.9	70.6	84.2

Table 1: Accuracy of all postures of body on detecting through pose model [14]

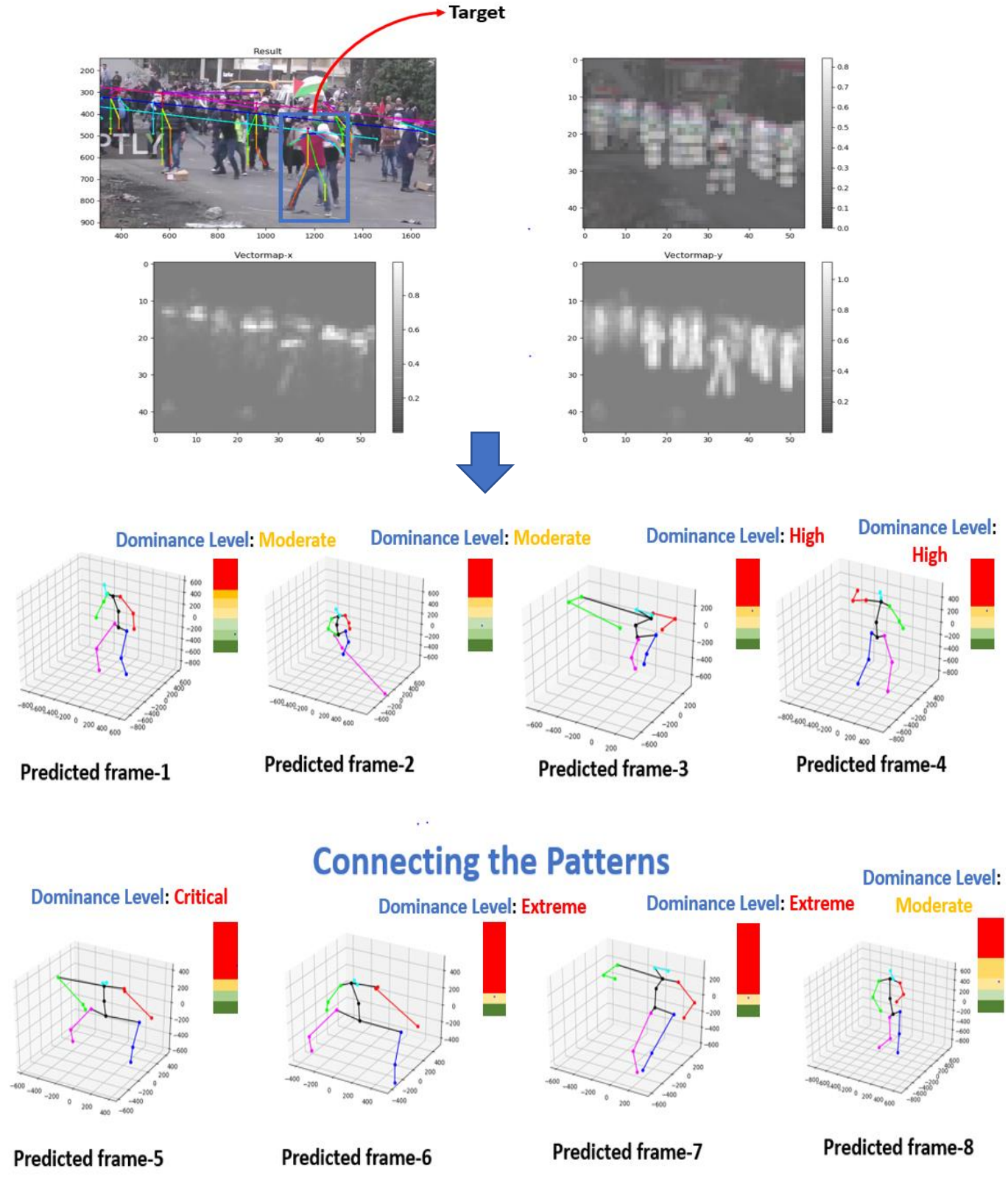

Fig 9: Connecting the Patterns of predicted pose from given input image

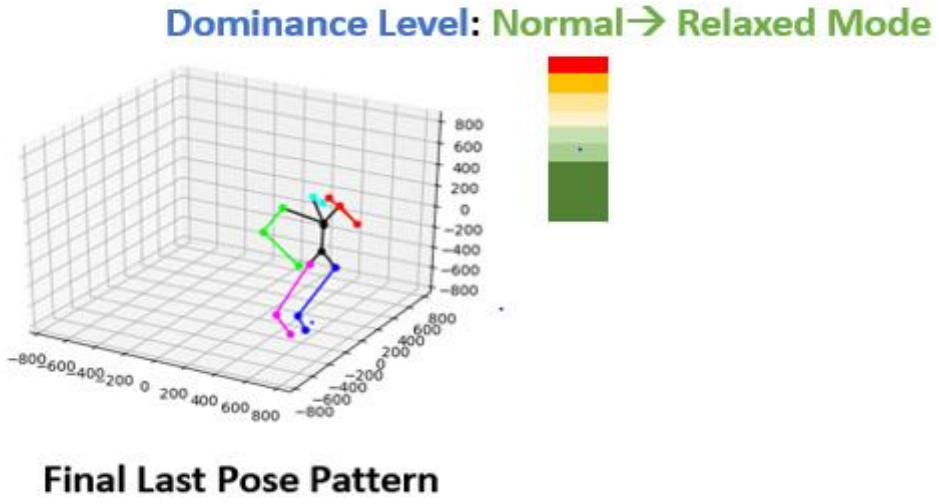

Fig 10: Final pose pattern of predicted pose of given input action image

Fig (9) and Fig (10) describes the final pose prediction from given input in-action image extracted from local Israel news channel video of Palestine stone pelting on the Israel Army. We have predicted the upcoming 9 frames of a particular single target which is shown in the boundary box in Fig (9). The predictions are plotted in 3D view for better analysis. Fig (10) describes the final pose of the predicted action when he comes to relaxed state or normal states. This pattern recognition of identifying the stone pelting action is predicted by series of patterns. The dominance levels of each predicted pose are shown according to which the analytical report will summarize the target situation at that scenario such *as {stone pelter, normal person, lost child, normal child}*. These patterns are used for detecting the dominance level of the target in the crowd and differentiating them with the rest.

5. Conclusion and Future Scope:

The current methods can find single pose accurately but for multi-pose it failed in huge crowds. Although the existing methods worked for multi-pose but when implemented in a huge crowd, it failed to plot pose for all the present people. Our proposed method worked even in the huge crowds, but the pose for farther view is rather negligible pose as their coordinates are merely nearer to zero. Their 3D plot is as simple as a single point. Our method has outperformed on all 2D and 3D datasets. For stone pelting, their body posture and especially hand pose [3] were much focussed on since the dominance levels of the hands is more critical than the body pose for both stone pelting and lost child in a crowd.

Pose estimation and prediction can be used in bio-medical applications such as cancer flow, virus growth and likewise. This can be explored to wide range of fields such as cyber security where identifying hack patterns. Our 3D view can be used for many applications where lesser dataset is available and this multi-angle views can be used as training dataset. We visualized the 3D plots and 3D viewpoints of a single pose in different angles for better psychological analysis.

References

[1] Anurag Ghosh, Suriya Singh, and C.V.Jawahar. Towards Structured Analysis of Broadcast Badminton Videos. 23 December 2017.

[2] B. Ghanem, M. Kreidieh, M. Farra, and T. Zhang. Context-aware learning for automatic sports highlight recognition. In Proc. ICPR, 2012.

[3] Carsten Rother, Vladimir Kolmogorov, Andrew Blake. "–Cut": interactive foreground extraction using iterated graph cuts. ACM Transaction on Graphics (TOG). Volume 23 Issue 3, August 2004. Pages 309-314.

[4] Chhaya Methani, and Anoop M. Namboodiri. Pose Invariant Palmprint Recognition. *International Conference on Biometrics*, 2009.

[5] S. Chen, Z. Feng, Q. Lu, B. Mahasseni, T. Fiez, A. Fern, and S. Todorovic. Play type recognition in real-world football video. In Proc. WACV, 2014.

[6] W.-T. Chu and S. Situmeang. Badminton Video Analy-sis based on Spatiotemporal and Stroke Features. In Proc. ICMR, 2017.

[7] A. Fathi and J. M. Rehg. Modeling actions through state changes. In Proc. CVPR, 2013.

[8] A. Fathi, X. Ren, and J. M. Rehg. Learning to recognize objects in egocentric activities. In Proc. CVPR, 2011.

[9] B. Ghanem, M. Kreidieh, M. Farra, and T. Zhang. Context-aware learning for automatic sports highlight recognition. In Proc. ICPR, 2012.

[10] Georg Poier, David Schinagl, and Horst Bischof. Learning Pose Specific Representation by Predicting Different Views. CVPR, 2018.

[11] Georgios Pavlakos, **Luyang Zhu**, Xiaowei Zhou, Kostas Daniilidis. Learning to Estimate 3D Human Pose and Shape from a Single Color Image. IEEE/CVF Conference on *Computer Vision and Pattern Recognition* (CVPR), 2018

[12] R. Girdhar, G. Gkioxari, L. Torresani, M. Paluri and D. Tran Detect-and-Track: Efficient Pose Estimation in Videos IEEE/CVF Conference on *Computer Vision and Pattern Recognition* (CVPR), 2018

[13] A. Hanjalic. Generic approach to highlights extraction from a sport video. In Proc. ICIP, 2003.

[14] Hao-Shu Fang, Shuqin Xie, Yu-Wing Tai, and Cewu Lu. RMPE: Regional Multi-Person Pose Estimation. 2016

[15] Mengyuan Liu, and Junsong Yuan. Recognizing Human Activities as the Evolution of Pose Estimation Maps. CVPR, 2018.

[16] K. Simonyan and A. Zisserman. Two-stream convolutional networks for action recognition in videos. In Proc. NIPS, 2014.

[17] K. Simonyan and A. Zisserman. Very deep convolutional networks for large-scale image recognition. In Proc. ICLR, 2014.

[18] Sirnam Swetha, Vineeth N Balasubramaniam, and C.V.Jawahar. Sequence-to-Sequence Learning for Human Pose Correction Videos. ACPR, 2017.

[19] V. Ren, N. Mosca, M. Nitti, T. DOrazio, C. Guaragnella, D. Campagnoli, A. Prati, and E. Stella. A technology plat-form for automatic high-level tennis game analysis. CVIU, 2017.

[20] S. Singh, C. Arora, and C. Jawahar. First person action recognition using deep learned descriptors. In Proc. CVPR, 2016.

[21] K. Soomro, A. Roshan Zamir, and M. Shah. UCF101: a dataset of 101 human actions classes from videos in the wild. 2012.

[22] S. Stein and S. J. McKenna. Combining embedded ac-celerometers with computer vision for recognizing food preparation activities. In Proc. UbiComp, 2013.

[23] M. Sukhwani and C. Jawahar. Tennisvid2text: Fine-grained descriptions for domain specific videos. In Proc. BMVC, 2015.

[24]M. Sukhwani and C. Jawahar. Frame level annotations for tennis videos. In Proc. ICPR, 2016.

[25]D. Tran, L. Bourdev, R. Fergus, L. Torresani, and M. Paluri. Learning spatiotemporal features with 3d convolutional net-works. In Proc. ICCV, 2015.

[26]K.-C. Wang and R. Zemel. Classifying nba offensive plays using neural networks. In MITSSAC, 2016.

[27]L. Wang, Y. Qiao, and X. Tang. Action recognition with trajectory-pooled deep-convolutional descriptors. In Proc. CVPR, 2015.