Bayesian Gait-based Gender Identification (BGGI) Network on Individuals Wearing Loosely Fitted Clothing

Aman Kumar[1,2], Anisha Jain[1,2] and Amarjot Singh[3]

[1]Skylark Labs LLP., India
[2]National Institute of Technology, Warangal, India
[3]Skylark Labs LLC., San Francisco, USA
{amank, anishaj, amarjot}@skylarklabs.ai

Abstract

Suspicious individuals often attempt to hide their identity to avoid detection by safety and security systems. Wearing clothes of the opposite gender is one of the several techniques used by these individuals. Several promising attempts have been made to recognize gender using gait recognition. However, these systems only focused on recognizing gender for individuals who wore tightly fitting attire which made it easier to detect the body joints further making it possible to differentiate both genders. In this work, we attempt to solve a challenging real-world problem faced by security agencies in which the individuals mask their identity by wearing loosely fitted clothes (LFC) of the opposite gender. LFC makes it difficult to locate the body joints in effect making the gender classification, in this situation, a complicated problem. We propose a Bayesian Gait-based Gender Identification (BGGI) technique that is used for gender recognition in LFC conditions, in dense real-world videos. This research releases the loosely fitted clothes individuals (LFCI) dataset used for training the deep network. This may encourage researchers interested in using deep learning for this task. The pose estimation and gender recognition achieve great performance with state-of-the-art techniques.

1. Introduction

Video surveillance has become an essential safety tool in today's society. In the last decade, the number of installed video surveillance cameras has reached the point where the vast majority can no longer be manually monitored by security personnel. This results in an increasing demand for automatic and intelligent video content analysis systems. Such systems can enable more efficient monitoring by only presenting footage of interest to the security personnel. This is achieved by characterizing individuals by their attributes such as motion, age and clothing.

From the safety and security perspective, gender identification through gait is an attractive modality because it may be performed at a distance surreptitiously, without the need of the face which may not be visible for subjects far away from the camera. Gender than can be applied for efficient search and retrieval of persons from video footage.

Researchers have attempted to model human motion using gait which is further used for gender classification. A popular technique used for gait is the Gait Energy Image (GEI) [5], a binary image produced by averaging the motion in one gait cycle of a subject. Numerous attempts have been made to improve the GEI method by optimizing the joint intensity and the space metric to improve the robustness and reduce the hindering effects of objects carried by the subject on the gender recognition performance [9], [10].

Deep networks have been recently used for more accurate gender classification. Chéron et al. [3] sampled patches of for RGB images of certain human joints to model human motion which was further used to perform promising gender classification. In another approach, Zhang et al. [21] extracted features from a sequence of individual images where were aggregated at the fully connected layers to learn complex high-level features used to perform the task of gender classification.

However, these systems perform identification only when the subjects are recorded from a close proximity. This limits the applicability of these systems in real-world scenarios as the image in the database needs to be matched to a video or video frames which may contain numerous faces that can appear at different positions, orientations, and scales. These videos can also be affected by noise and illumination variations which further complicates this problem.

Figure 1. The illustration shows the frames of males and females at different variations from the proposed Loosely Fitted Clothing (LFC) dataset.

The data for the above mentioned works has mainly been acquired in a Western environment with subjects wearing Western clothing. In some cases, non-Western clothing can obscure the subject's joints and its movements suppressing visual features pertaining to the gender of the individual. This makes gait recognition and gender classification even more challenging. The face of the person may be covered hiding any facial features adding to the complexity of the challenge.

This paper introduces the Bayesian Gait-based Gender Identification (BGGI) Network on individuals wearing Loosely Fitted Clothing. The video recorded by the CCTV cameras is first decomposed into frames. The network then extracts the humans and estimates their poses using the ScatterNet Hybrid Part Affinity Fields (SH-PAF) Network constructed by replacing the first convolutional, relu and pooling layers of the Part Affinity Fields (PAFs) network [1] with the hand-crafted ScatterNet [15] as shown in Fig. 3. The poses are then used to identify suspicious (Fig. 1) individuals using the 3D ResNext [6] with Bayesian uncertainty estimates. The features obtained from the deeper layer of the BGGI network are also used to perform one-to-one person re-identification.

The novelties of the proposed BGGI network are detailed below:

- ***Rapid learning with ScatterNet and Structural Priors***: The proposed SH-PAF network is constructed by with the hand-crafted ScatterNet (front-end) that extracts translation, rotation, and scale invariant low-level edge features from the input images (similar to the replaced layer). These features can be used by the PAF (back-end) network to learn more complex features from the start of learning as the edges are already present, resulting in accelerated training. The ScatterNet invariant features are particularly useful for this application as the human can appear at different locations, orientations, and scales. The training of the PAF network is further accelerated by initializing the filter weights with structural priors learned (unsupervised) using the PCANet [2] framework (Fig. 3). The initialization with priors also ***reduces the need for sizeable labeled training datasets*** for effective training which is especially advantageous for this task or other applications [14, 7] as it can be expensive and time-consuming to generate keypoint annotations.
- ***Bayesian Uncertainty***: The proposed network uses dropout at test time to make several predictions. The mean and standard deviation of these predictions is calculated which can aid the user in deciding if a certain prediction can be trusted.
- ***Loosely Fitted Clothing (LFC) Dataset***: The paper presents the Loosely Fitted Clothing dataset of 2400 videos of 25 individuals wearing loosely fitted clothes that cover their body joints. The LTC dataset contains humans recorded at different variations of scale, position, illumination, blurriness etc. This dataset may en-

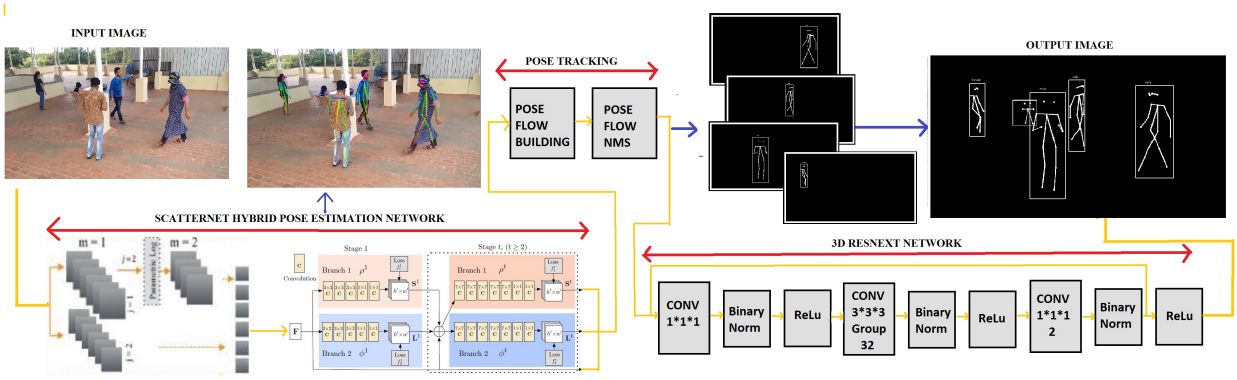

Figure 2. The illustration presents the proposed Bayesian Gait-based Gender Identification (BCGI) network for loosely fitted clothes. The input image is first fed to the scatternet that extracts the low-level translation invariant features which are then used by the pruned part affinity fields network to estimate pose of the humans. The estimate pose is binarized and fed into a 3D ResNext network which uses 16 frames to estimate the gender of the humans.

courage researchers interested in using deep learning for aerial surveillance applications.

The pose estimation, individuals identification, and re-identification performance of the system is compared with the state-of-the-art techniques.

The paper is divided into the following sections. Section 2 presents the introduced LFC dataset while Section 3 introduces the proposed BGGI system. Section 4 details the experimental results and concludes this research.

2. Loosely Fitted Clothing Dataset

This research proposes an annotated Loosely Fitted Clothing dataset which is used for pose estimation and gender classification. The dataset consists of 2160 videos of humans in loosely fitted clothes walking along a straight line. The subjects in the dataset wear attire with the following attributes (i) scarf with loose dress (ii) hoodie with skirt (iii) scarf with skirt (iv) hoodie with loose pants (v) scarf with loose pants. The subjects in the dataset include 8 females and 10 males. Each subject in the dataset is recorded under the following observation angles: $0°, 45°, 90°, 135°, 180°, 225°, 270°, 315°$. These activities are performed by the 18 subjects between the ages of 18-22 years. These videos are recorded from an height of 2.5-3 meters. Each video is 920x540 pixels and around 120 frames.

The gender classification task from these videos is an extremely challenging problem as these videos can be affected not only by the elementary issues of illumination changes, shadows, poor resolution, and blurring, but also, more importantly, by the nature of the clothing. The joints and limbs of the individuals are hidden from sight making it difficult to observe the movement of the body. In addition to these variations, the humans can appear at different locations, orientations, and scales. The proposed dataset includes videos with the above-detailed variations as these can significantly alter the appearance of the humans and affect the performance of the surveillance systems. The 3D Convolutional Neural Network, when trained on the Loosely Fitted Clothing Dataset with these variations, can learn to recognize human gender despite these variations.

3. Bayesian Gait-based Gender Identification (BGGI) Network

This section introduces the Bayesian Gait-based Gender Identification (BGGI) Network which first uses the proposed ScatterNet Hybrid Part Affinity Fields (SH-PAF) Network to estimate pose for the humans, whose output is further fed to the 3D ResNext, which captures the motion of an individual to predict the gender. The system uses cloud computation to achieve the identification of the gender of the person of interest in real-time. Each part of proposed network is explained in the following sub-sections.

3.1. ScatterNet Hybrid Part Affinity Fields

This section details the proposed ScatterNet Hybrid Part Affinity Fields (SH-PAF) Network, inspired from Singh et al.'s work in [16, 17, 14, 18], composed by combining the hand-crafted (front-end) two-layer parametric log ScatterNet [15] with the pruned Parts Affinity Fields (PAFs) [1] network (back-end) as shown in Fig. 3. The ScatterNet accelerates the learning of the SH-PAF network by extracting invariant edge-based features which allow the network to learn complex features from the start of the learning [16]. The regression network also uses structural priors to expedite the training as well as reduce the dependence on the annotated datasets. The ScatterNet (front-end) and pruned

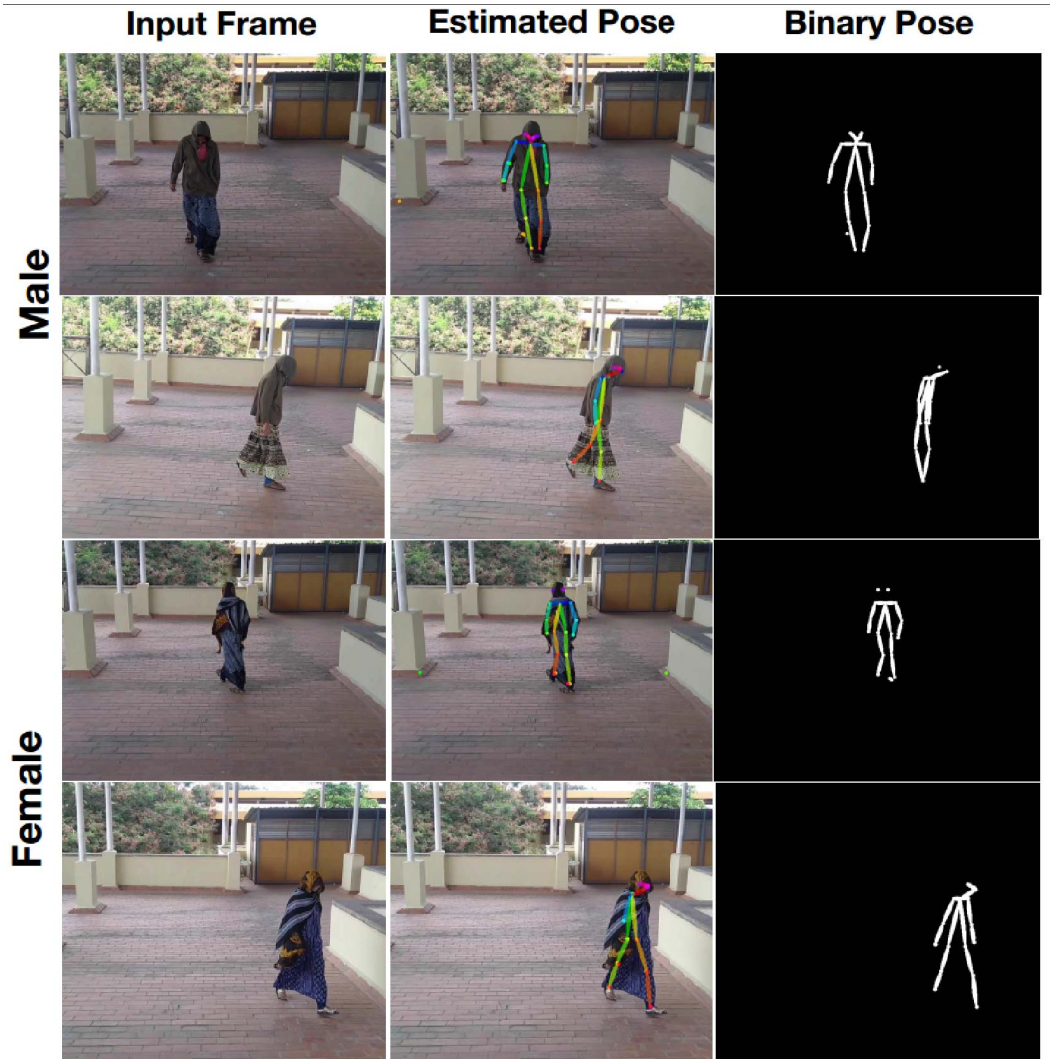

Figure 3. The illustration shows the pose estimated on two male and female examples for the proposed LFC dataset. The estimated is binarized pose and given as input to the 3D ResNext network for gender classification.

Parts Affinity Fields (PAF) are presented below.

ScatterNet (front-end): The parametric log based ScatterNet [15] is a two-layer hand-crafted network which extracts translation, rotation, and scale invariant feature representations from multi-resolution images obtained at 1.5 times and twice the size of the input image. Below we present the formulation of the parametric ScatterNet for a single input image which may then be applied to each of the multi-resolution images.

The invariant features are obtained at the first layer by filtering the input image or signal x with dual-tree complex wavelets (better than cosine transforms [8]) $\psi_{j,r}$ at different scales (j) and six pre-defined orientations (r) fixed to $15°, 45°, 75°, 105°, 135°$ and $165°$. To build a more translation invariant representation, a point-wise L_2 non-linearity (complex modulus) is applied to the real and imaginary part of the filtered signal:

$$U[\lambda_{m=1}] = |x \star \psi_{\lambda_1}| = \sqrt{|x \star \psi^a_{\lambda_1}|^2 + |x \star \psi^b_{\lambda_1}|^2} \quad (1)$$

The parametric log transformation layer is then applied to all the oriented representations extracted at the first scale $j = 1$ with a parameter $k_{j=1}$, to reduce the effect of outliers by introducing relative symmetry of pdf [15], as shown below:

$$U1[j] = \log(U[j] + k_j), \quad U[j] = |x \star \psi_j|, \quad (2)$$

Next, a local average is computed on the envelope $|U1[\lambda_{m=1}]|$ that aggregates the coefficients to build the desired translation-invariant representation:

$$S_1[\lambda_{m=1}] = |U1[\lambda_{m=1}]| \star \phi_{2^J} \quad (3)$$

The high frequency components lost due to smoothing are retrieved by cascaded wavelet filtering performed at the second layer. Translation invarinace is introduced in these features by applying the L2 non-linearity with averaing as explained above for the first layer [15].

The scattering coefficients at L0, L1, and L2 are:

$$S = \left(x \star \phi_{2^J}, S_1[\lambda_{m=1}], S_2[\lambda_{m=1}, \lambda_{m=2}] \star \phi_{2^J}\right) \quad (4)$$

The rotation and scale invariance are next obtained by filtering jointly across the position (u), rotation (θ) and scale(j) variables as detailed in [13].

The features extracted from each multi-resolution at L0, L1, and L2 are concatenated and given as input to the pruned Part Affinity Fields (PAFs) network, to learn high-level features for human pose estimation. The ScatterNet features help the proposed SH-PAF network to converge faster as the convolutional layers of the PAF network can learn more complex patterns from the start of learning as it is not necessary to wait for the first layer to learn invariant edges as the ScatterNet already extracts them.

Pose Estimation with Structural Priors (back-end): The invariant ScatterNet features are used by the pruned Parts Affinity Fields (PAFs) network [1] (initial layers replaced with ScatterNet) to learn pose estimation using the introduced AVI dataset. The AVI dataset contains aerial images with 18 annotated key-points with 36 coordinates (section 2) on the human body which are used by the network to learn the human poses.

The pruned PAF network is composed of a feedforward network which is divided into two branches which simultaneously predicts a set of confidence maps S of body part locations and a set of vector fields L of part affinities for each limb which preserves both the position and orientation information of the limb. The predictions of each of the branch are iteratively refined over successive stages following Wei et al [19]. Finally, the confidence maps and part affinity fields are parsed by greedy inference to output the body keypoints for all people in the image.

At a stage t, the loss functions for each of the branches is given by:

$$f_s^t = \sum_{j=1}^{J} \sum_{p} W(p).||S_j^t(p) - S_j^*(p)||_2^2 \quad (5)$$

$$f_L^t = \sum_{c=1}^{C} \sum_{p} W(p).||L_c^t(p) - L_c^*(p)||_2^2 \quad (6)$$

where S_j^t is the set of confidence maps and L_c^t is the set of part affinity fields at stage t. S_j^* is the confidence map ground truth, L_c^* is the part affinity field ground truth, J is the number of body parts and C is the number of vector fields (one per limb). W is a binary mask with $W(p) = 0$ when an annotation is missing at a location p which is an issue in some datasets that do not completely label all the people.

The overall objective to be minimized is given by:

$$f = \sum_{t=1}^{T} (f_S^t + f_L^t) \quad (7)$$

The detected body parts are then associated with a person using the part affinity fields of each limb.

Structural Priors: In order to accelerate the training, each convolutional layer of the joints identification network of the SHDL network is initialized with structural priors. The structural priors are obtained for each layer using the PCANet [2] framework. By minimizing the following reconstruction error, it learns a family of orthonormal filters:

$$\min_{V \epsilon\ R^{z_1 z_2 \times K}} \left\|X - VV^T X\right\|_F^2,\ s.t.\ VV^T = I_K \quad (8)$$

Where X are patches sampled from N training features, I_K is an identity matrix of size $K \times K$. The solution of Eq. 8in its simplified form represents K leading principal eigenvectors of XX^T obtained using Eigen decomposition.

The structural priors for the first layer of joins identification network are learned on the ScatterNet features, the following layers structural priors are learned on the previous layers outputs and so on. This is applied to both of the branches present in the network. The structural priors for the joints identification networks layers learn filters that respond to a hierarchy of features which is similar to the features learned by CNNs. These learned structural priors are used to initialize each of the convolutional layer resulting in accelerated training. Since the determination of structural prioris is fast, the training process is much faster than that of CNNs with random weight initializations. However, the PCA framework may learn undesired checkerboard filters. In order to detect the checkerboard filters from the learned filter sets, we use the method defined in [4] and are then avoided as filter priors.

3.2. Gender Classification using the Bayesian 3D-ResNext

A 3D ResNext [20] is trained on 16 subsequent frames to perform gender classification. In the proposed system, we use Monte Carlo dropout at prediction time to measure the 3D ResNext models uncertainty. We make 50 predictions at test time with dropout enabled. The variance of these predictions can be used to measure how certain the model is about the prediction.

4. Results

This section presents the training details and the performance of the Bayesian Gait-based Gender Identification (BGGI) Network for gender classification on the proposed

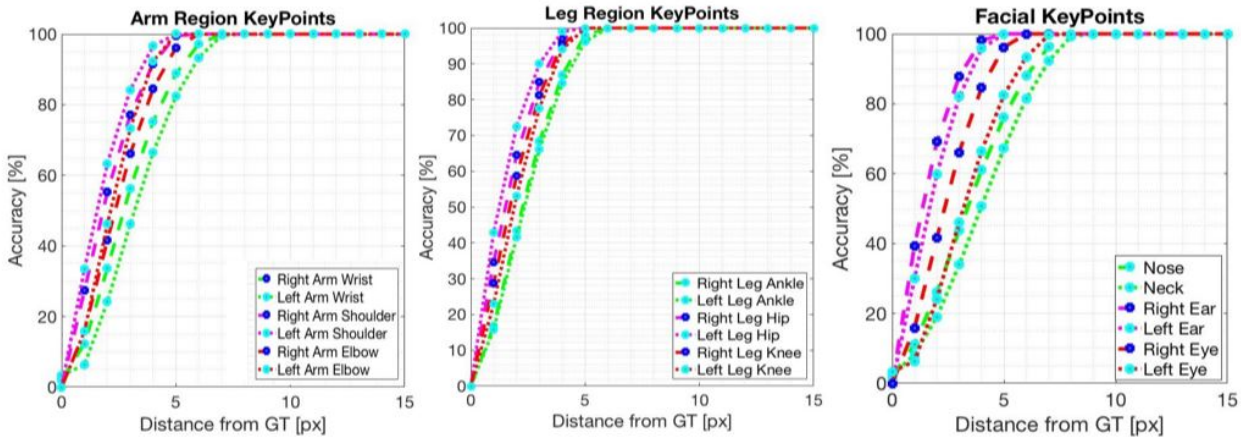

Figure 4. Illustration shows the pose estimation performance via the detection of key-points for the (a) arms region, which constitutes the wrist, shoulder and elbow, (b) legs region, which includes ankle, knee, and hip, and, (c) facial regions with the head and neck.

Loosely Fitted Clothing (LFC) dataset. The BGGI network first uses the SH-PAF network for human pose estimation, next the estimated poses are binarized and given (16 frames) as input to the temporal 3D ResNext network for gender classification. The next sections detail the training details of the SH-PAF network along with the performance of each part of the BGGI network. The classification performance is also compared with the state-of-the-art techniques.

4.1. SH-PAF Parameters and Training

The SH-PAF network is constructed by combining the scatternet with the pruned PAF network.

ScatterNet: The scatternet extracts invariant low-level features using DTCWT filters at 2 scales, and 6 fixed orientations at layers L0, L1, and L2.

PAF Network with Structural Priors: The back end of SH-PAFnetwork is a pruned PAF. The pruned PAF is trained on the scatternet features that are extracted from the 2570 humans. Out of 2570 humans, 60% are used for the training set and 20% for validation and 20% for the testing. The splitting of the dataset of extracted features is completely random. The network parameters are as follows: The base learning rate is 10^{-4}, which we decrease to 10^{-5} after 15 iterations, the dropout is 0.5, the batch size is 32, and the total number of iterations (epochs) is 70. To accelerate the training, the convolutional layers are initialized with structural priors.

4.2. Key Points Prediction Performance

The pose estimation performance of the SH-PAF network is evaluated within a set distance of d pixels from the ground truth key-point, as shown in Fig. 4 via the accuracy vs.distance graphs, for different regions of the body.

The key-points detection analysis for the arms, legs, and facial, region is presented below:

Arms: This region comprises of six points : wrist key-points (P5 and P8), shoulder key-points (P3 and P6), and elbow key-points(P4 and P7). As the figure shows, the SHDL network can detect the wrist region points with an accuracy of 60% for pixel distance, d=5, while the accuracy for elbow and shoulder region is higher, at about 85% and 95% each for the same value of d.

Legs: This region consists of six points : hip key-points, knee key-points and ankle key-points. As the figure shows, the SHDL network can detect the hip key-points with 100% accuracy for pixel distance of d=5. The accuracy for knee key-points varies between 85% and 90%. But for the ankle key-points the accuracy falls to around 85%.

Facial : This region consists of only two points head and neck. The network detects neck key-point with an accuracy of around 95% while the accuracy falls down to 77% for the detection of head key-point for the same pixel distance d=5.

4.3. SH-PAF Performance and comparison

The human pose estimation performance of the SH-PAF network on the LFC dataset is presented in Table 1. The performance is compared with 3 networks namely CoordinateNet (CN), CoordinateNet extended (CNE) and SpatialNet. The key-point detection accuracy results for the dataset are is shown in the table. The SH-PAF network outperforms the other networks by a significant margin and performs as well as SpatialNet.

4.4. Gender Classification

The estimated body jointed are connected together to form a human skeleton structure as shown in Fig. 3. A set of

Figure 5. The illustration presents the gender classifications results on the three images with multiple humans from the LFC dataset.

Table 1. Comparison of KeyPoint detection accuracies(of various architectures namely Coordinate Net (CN) [11], Coordinate extended (CNE) [11], Spatial net [12] and SH-PAF network on LFC dataset

Dataset	other architecture			
	SHDL	CN	CNE	Spatial Net
LFC	80.3	76.1	74.8	78.2

16 frames containing these skeleton structures are given as input to the 3D ResNext which performs the binary (male vs female) classification. The performance of the system for different attributes of the LFC dataset and orientations at which the human appears to the camera is shown in Table 2 below:

Table 2. Table shows the accuracy distribution (in %) of the proposed model over different attributes and angles (in degrees)

Clothing	0	45	90	135	180	225	270
Scarf with loose dress	100	98	100	97	98	97	96
Hoodie with skirt	98	100	100	95	98	100	86
Scarf with skirt	100	98	98	93	98	95	93
Hoodie with loose pants	98	96	100	100	100	98	98
Scarf with loose pants	100	100	100	98	98	98	96

The classification performance for the humans at different distances from the camera is also shown in Table 3.

As the distance of the humans increases, the accuracy of the proposed system decreases as often the pose estimated for humans which are large distances is not accurate.

The classification performance of the system is also shown for multiple humans in Table 4 as shown below.

The dataset also contains samples with more than 5 humans as well.

The classification performance is also compared with the state-of-the-art technique which were developed to perform gender classification using gait as shown in Table. 5. The proposed BGGI was able to outperform the state-of-the-art methods by more than 4% on the LFC dataset.

Table 3. The table presents the classification accuracies(%) with the increase in distance (m) for individuals in the LFC dataset.

	Height (m)			
	5	10	15	20
BGGI	93.1	91.6	88.3	85.8

Table 4. The table presents the classification accuracies(%) with the number of individuals in an image from the LFC dataset.

	Humans (No.)			
	2	3	4	5
BGGI	96.9	90.3	87.9	81.1

Table 5. Table shows the suspicious activity classification accuracy (%) compared against the two state-of-the-art method.

	Comparison		
	BGGI	Chéron [3]	Zhang [21]
Acc.	87.8	81.2	83.8

4.5. Runtime Performance

The BGGI framework consisted of two parts:: (i) human pose estimation using the SH-PAF network, and (iii) classification of the estimated human pose as male or female using the Bayesian 3D ResNext. Its runtime performance is computed on the cloud. The framework was trained and evaluated using the cuDNN framework and NVIDIA Tesla GPUs. For an image frame, the system detected and determined genders of individuals at 5 fps per second to 16 fps for upto ten people. The processing time varies in accordance with the number of individuals in the image frame.

5. Conclusion

This paper proposes a Bayesian Gait-based Gender Identification framework that can determine the gender of an individual from videos using their walking pattern or gait.

The framework first uses the proposed SH-PAF network consisting of Joints Identification network using Part Affinity Fields to detect humans and estimate their pose. The estimated poses are used by the 3D ResNext to follow the gait and predict the gender of individual. The proposed SH-PAF network uses ScatterNet features with structural priors initialization to achieve accelerated training using relatively fewer labelled examples. The use of fewer labelled examples is beneficial for this application since it is expensive to collect annotated examples. The paper also introduced the Loosely Fitted Clothing (LFC) Dataset which can benefit other researchers aiming to use deep learning methods for gender classification on videos and images where the faces and body shapes of the individuals is difficult to perceive. The proposed framework outperforms the state-of-art techniques on the LFC dataset. We believe the framework would be instrumental in detecting disguised individuals in public areas or large gatherings.

References

[1] Z. Cao, T. Simon, S. Wei, and Y. Sheikh. Realtime multi-person 2d pose estimation using part affinity fields. *CoRR*, abs/1611.08050, 2016.

[2] T.-H. Chan, K. Jia, S. Gao, J. Lu, Z. Zeng, and Y. Ma. Pcanet: A simple deep learning baseline for image classification? *IEEE Transactions on Image Processing*, 2015.

[3] G. Chéron, I. Laptev, and C. Schmid. P-CNN: pose-based CNN features for action recognition. *CoRR*, abs/1506.03607, 2015.

[4] A. Geiger, F. Moosmann, Ö. Car, and B. Schuster. Automatic camera and range sensor calibration using a single shot. In *Robotics and Automation (ICRA), 2012 IEEE International Conference on*, pages 3936–3943, 2012.

[5] J. Han and B. Bhanu. Individual recognition using gait energy image. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 28(2):316–322, Feb 2006.

[6] K. Hara, H. Kataoka, and Y. Satoh. Can spatiotemporal 3d cnns retrace the history of 2d cnns and imagenet? *CoRR*, abs/1711.09577, 2017.

[7] S. Jain, S. Gupta, and A. Singh. A novel method to improve model fitting for stock market prediction. *International Journal of Research in Business and Technology*, 3(1):78–83.

[8] V. Jeengar, S. Omkar, A. Singh, M. K. Yadav, and S. Keshri. A review comparison of wavelet and cosine image transforms. *International Journal of Image, Graphics and Signal Processing*, 4(11):16, 2012.

[9] Y. Liu, J. Zhang, C. Wang, and L. Wang. Multiple hog templates for gait recognition. In *Proceedings of the 21st International Conference on Pattern Recognition (ICPR2012)*, pages 2930–2933, Nov 2012.

[10] Y. Makihara, A. Suzuki, D. Muramatsu, X. Li, and Y. Yagi. Joint intensity and spatial metric learning for robust gait recognition. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 6786–6796, July 2017.

[11] T. Pfister. Advancing human pose and gesture recognition. In *University of Oxford*, 2015.

[12] T. Pfister, J. Charles, and A. Zisserman. Flowing convnets for human pose estimation in videos. In *IEEE International Conference on Computer Vision*, 2015.

[13] L. Sifre and S. Mallat. Rotation, scaling and deformation invariant scattering for texture discrimination. In *Computer Vision and Pattern Recognition (CVPR), 2013 IEEE Conference on*, pages 1233–1240, 2013.

[14] A. Singh, D. Hazarika, and A. Bhattacharya. Texture and structure incorporated scatternet hybrid deep learning network (ts-shdl) for brain matter segmentation. *International Conference on Computer Vision Workshop*, 2017.

[15] A. Singh and N. Kingsbury. Dual-tree wavelet scattering network with parametric log transformation for object classification. In *International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2017.

[16] A. Singh and N. Kingsbury. Efficient convolutional network learning using parametric log based dual-tree wavelet scatternet. *IEEE International Conference on Computer Vision Workshop*, 2017.

[17] A. Singh and N. Kingsbury. Scatternet hybrid deep learning (shdl) network for object classification. *International Workshop on Machine Learning for Signal Processing*, 2017.

[18] A. Singh and N. Kingsbury. Generative scatternet hybrid deep learning (g-shdl) network with structural priors for semantic image segmentation. *IEEE International Conference on Acoustics, Speech and Signal Processing*, 2018.

[19] S.-E. Wei, V. Ramakrishna, T. Kanade, and Y. Sheikh. Convolutional pose machines. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4724–4732, 2016.

[20] S. Xie, R. B. Girshick, P. Dollár, Z. Tu, and K. He. Aggregated residual transformations for deep neural networks. *CoRR*, abs/1611.05431, 2016.

[21] Y. Zhang, Y. Huang, L. Wang, and S. Yu. A comprehensive study on gait biometrics using a joint cnn-based method. *Pattern Recognition*, 93:228 – 236, 2019.