PERSON IDENTIFICATION USING SPATIOTEMPORAL MOTION CHARACTERISTICS

Muhammad Hassan Khan\textsuperscript{1,2}, Muhammad Shahid Farid\textsuperscript{2}, Marcin Grzegorzek\textsuperscript{1,3}

\textsuperscript{1}Research Group of Pattern Recognition, University of Siegen, Siegen, Germany
\textsuperscript{2}College of Information Technology, University of the Punjab, Pakistan
\textsuperscript{3}University of Economics in Katowice, Katowice, Poland

ABSTRACT

Biometric gait recognition has received substantial attention of researchers in the recent years due to its applications in numerous fields of computer vision, particularly in visual surveillance and monitoring systems. Most existing gait recognition algorithms solve the problem of person identification either by constructing a human body model based on various skeletal data characteristics such as joints positioning and their orientation, or use gait features, e.g., stride length, gait patterns and other shape templates. Such approaches require the extraction of the human-body’s silhouette, contour, or skeleton from the images, and therefore their performance highly depends on the silhouette segmentation accuracy. In this paper, we propose a novel gait recognition algorithm which exploits spatiotemporal motion characteristics of a person, which does not need silhouette or skeleton extraction at all. The proposed algorithm computes a set of spatiotemporal features from the video sequences and uses them to generate a codebook. Fisher vector is used to encode the motion descriptors which are classified using linear Support Vector Machine (SVM). The proposed algorithm is evaluated on three benchmark gait datasets: TUM GAID, CASIA-B, and CASIA-C. It achieved excellent results on all datasets which demonstrate the effectiveness of the proposed algorithm.

Index Terms—Gait recognition, Spatiotemporal features, Fisher vector encoding, Visual surveillance

1. INTRODUCTION

Biometrics has received significant research efforts in the recent years due to its growing applications in authentication, access control and surveillance. Studies \cite{1,2} have shown that individuals can be identified by using different distinguishing biological traits. Biometrics refers to the physiological or behavioral characteristic of the human, e.g., fingerprints, facial features, iris, DNA, voice, and gait, which have proven to be unique for each individual. Gait refers to the walking style of a person and is considered an important cue for person identification. Unlike other biometrics, gait does not require human interaction with the system which makes it the most suitable for surveillance systems. Moreover, gait biometrics can be used at low resolution in a non-invasive and hidden manner. Gait recognition, however, is challenging as many factors may affect it such as clothing, shoes, walking surface and injuries. Gait may not be as powerful as other biometric modalities such as fingerprints to identify the individuals, however its characteristic to recognize human from distance and without any interaction makes it irreplaceable in many applications such as visual surveillance.

The gait recognition approaches in literature can be divided into two broad categories: model-based and model-free approaches. The model-based techniques build the human body structure and motion models by tracking the different body parts and joint position over time using the underlying mathematical structure \cite{4}, and use them to recognize the people. These models such as \cite{5,6,7} may include stick figure, interlinked pendulum and ellipse are generally constructed based on the prior knowledge of the human body shape. Recent studies have demonstrated that such models are capable to deal with the occlusion and rotation problems. However, they are computationally inefficient and sensitive to the quality of video data, and therefore they are not considered suitable for real-world and real-time applications \cite{8}.

In this paper, we present a novel spatiotemporal gait repre-
representation using dense trajectories to characterize the distinctive motion traits of human gait. Unlike most existing gait recognition algorithms that require the extraction of the human body silhouette or other skeletal information, the proposed approach is model-free. It neither involves any kind of human body segmentation nor requires gait cycle estimation. Experiments worked out on three well-known gait databases confirm the effectiveness of the proposed algorithm.

2. PROPOSED METHOD

The proposed gait recognition algorithm works in three steps. First, dense trajectories are generated based on optical flow field and their motion information is encoded using local descriptors. Second, a codebook based on Gaussian Mixture Model (GMM) is built and the local descriptors are encoded using Fisher vector (FV). Finally, the computed features are classified using linear Support Vector Machine (SVM) to recognize the individuals.

2.1. Motion descriptor estimation

Recently, dense trajectories have demonstrated excellent results in action recognition [19,20]. Our motivation to use dense trajectories is that they encode the local motion patterns of gait and can be easily computed from video sequences. To extract dense trajectories, a set of dense points is selected from each frame and tracked in successive frames using displacement information from a dense optical flow field. Given a trajectory of length \( L \), a sequence \( S \) of displacement vector \( \Delta P_t \) is computed as given below [19]:

\[
S = (\Delta P_t, \ldots, \Delta P_{t+L-1}),
\]

where \( \Delta P_t = (P_{t+1} - P_t) = (x_{t+1} - x_t, y_{t+1} - y_t) \). \( P_t \) and \( P_{t+1} \) represent a point in frame \( t \) and \( t + 1 \) respectively. The sequence vector \( S \) is then normalized by the sum of the magnitudes of the displacement vector. That is,

\[
S' = \frac{(\Delta P_t, \ldots, \Delta P_{t+L-1})}{\sum_{j=t}^{t+L-1} \| \Delta P_j \|}
\]

The descriptor \( S' \) encodes the shape of the trajectory. Wang et al. [19] proposed the Histogram of Oriented Gradient (HOG) and Histogram of Optical Flow (HOF) features along the dense trajectories. In addition, to encode the relative motion information between pixels, the derivatives along the horizontal and the vertical components of the optical flow are also computed, known as Motion Boundary Histograms and are represented as MBH_x and MBH_y respectively. We evaluated various combinations of these descriptors on TUM GAID database [17], the results are shown in Fig. 1. These results reveal that HOG in combination with MBH outperform the rest and achieves up to 100% recognition accuracy. HOG captures the characteristics of a person’s static appearance and MBH highlights the information about the changes

2.2. Feature encoding

Inspired by the recent popularity of FV encoding in image classification, object detection and action recognition [20,21], we encode our local descriptors using FV and a codebook based on GMM. FV is derived from the Fisher kernel [21] that combines the characteristics of both discriminative and generic approaches. The basic idea is to model a feature set by gradient of its log-likelihood function with respect to model parameters. To build a codebook, we used GMM from one million randomly selected features of each descriptor. GMM is a generative model that defines the distribution over feature space and can be described:

\[
p(X \mid \theta) = \sum_{i=1}^{K} w_i N(x \mid \mu_i, \Sigma_i)
\]

where \( i = 1, 2, \ldots, K \) is the mixture (i.e., cluster) number, \( w_i, \mu_i \) and \( \Sigma_i \) are the weight, mean vector and covariance matrix of the \( i \)th cluster, respectively. Furthermore, \( \theta = \{w_i, \mu_i, \Sigma_i\} \) is the set of model parameters, and \( N(X \mid \mu_i, \Sigma_i) \) represents the \( D \)-dimensional Gaussian distribution. For a given feature set \( X = \{x_t, t = 1, \ldots, T\} \), the optimal parameters of GMM are learned using maximum likelihood estimation. The soft assignment of data \( x_t \) to cluster \( i \) can be defined as,

\[
q_t(i) = \frac{w_i N(x_t \mid \mu_i, \Sigma_i)}{\sum_{j=1}^{K} w_j N(x_t \mid \mu_j, \Sigma_j)}
\]

We assume that each model represents a specific motion pattern shared by the local descriptors in the codebook. The Expectation Maximization (EM) algorithm of GMM applies soft assignments of the feature descriptor to each mixture component. Therefore, the local descriptors will be assigned to multiple clusters in a weighted manner using the posterior component probability given by the descriptor. The feature
set $X$ can be modeled into a vector by computing the gradient vector of its log-likelihood function at the current $\theta$,  
\[
F_X = \frac{1}{T} \nabla_\theta \log p(X|\theta),
\]
where $F_X$ represents the FV and $\nabla_\theta$ is the gradient of the log-likelihood function, which describes the contribution of parameters in the generation process. Let $x_i$ be the local descriptor, $q_l(i)$ be the soft assignment of $x_i$ to cluster $i$, $\sigma_i$ is the diagonal element of $\sum_i$, $u_i$ and $v_i$ are the gradient vector with respect to $\mu_i$ and $\sigma_i$, respectively [22]:
\[
u_i = \frac{1}{\sqrt{\omega_i}T} \sum_{t=1}^{T} q_l(i) (x_t - \mu_i)^2 \sigma_i^2 - 1.
\]

Equation (6) and (7) are known as the first and the second order differences of descriptor points to cluster centers, respectively. The final gradient vector (i.e., FV encoding for the set of local descriptors $X$) is computed by concatenating the all $u$ and $v$ for all $K$ clusters. That is,
\[
f = [u_1^T, v_1^T, u_2^T, v_2^T, \ldots, u_K^T, v_K^T]^T.
\]

The total size of encoded vector is $2K D$, where $K$ is the total number of clusters and $D$ is the dimension of the descriptor. We encode our HOG, MBH$_x$ and MBH$_y$ descriptors using the above described method and fuse them using representation level fusion [20].

### 2.3. Gait classification

The encoded vectors are classified using Linear Support Vector Machine (SVM). SVM is considered a powerful tool for solving classification problems in many applications [23, 24]. Due to the high dimensionality of our features, we decided to use SVM as a classifier. In contrast to SVM, the other similarity based classifiers like K-Nearest Neighbor and probability based classifiers such as Naive Bayes do not perform well on high dimensional features [23]. SVM first maps the training samples in high dimensional space and then extracts a hyperplane between the different classes of objects using the principle of maximizing the margin. Because of this principle, the generalization error of SVM is theoretically independent from the number of feature dimensions. We used LIBLINEAR SVM library [25] for classification.

### 3. EXPERIMENTS AND RESULTS

The performance of the proposed method is evaluated on three popular benchmark gait recognition databases: TUM GAID database [17], CASIA B database [26] and CASIA C dataset [27]. In all experiments, the local descriptors on each video sequence are computed using dense trajectories. The codebook size $K$ is empirically computed and set to 32.

#### 3.1. Results on TUM GAID database

TUM GAID is one of the largest gait databases comprising 3,370 gait sequences of 305 subjects. It was recorded in two seasons, winter and summer, using Microsoft Kinect at 30 frames-per-second (fps). A subset of 32 subjects participated in both seasons. Therefore, there is a substantial variation in the clothing of the participants which makes it a challenging gait database. Ten walk sequences were captured for each subject, namely normal walk ($N$), walk with backpack ($B$) and walk with coating shoes ($S$). Each subject in the common subset of 32 people has 10 more sequences referred to as normal walk after time ($TN$), walk with backpack after time ($TB$), and walk with coating shoes after time ($TS$).

The gallery and probe set division is done similarly to [17]. The first four recordings of $N$ (i.e., $N_1 - N_4$) for each person are used as gallery set, and the sequences $N_5 - N_6$, $B_1 - B_2$ and $S_1 - S_2$ are used in probe set, giving three experiments namely $N$, $B$ and $S$. In the next set of experiments labeled as $TN$, $TB$ and $TS$, the sequences $N_7 - N_8$, $B_3 - B_4$ and $S_3 - S_4$ are used in probe set, while the gallery set is the same. The recognition results achieved by the proposed algorithm and other gait recognition methods on TUM GAID database are presented in Tab. 1. The proposed algorithm achieves the best results on $N$, $B$, $S$, and $TB$ experiments. In $TN$ and $TS$ experiments PFM [28] and CNN-SVM [29] performs better than our method respectively. On average the proposed algorithm achieved the best recognition rate 96.5%.

#### 3.2. Results on CASIA-B database

The CASIA-B gait database contains the walk sequences of 124 subjects, recorded from 11 different viewing angles in a well controlled laboratory environment at 25 fps. Three different variations in walking style namely normal walk ($nm$), walk with bag ($bg$) and walk with coat ($cl$) are recorded for each person. There are 10 walking sequences for each subject: 6 of normal walk, 2 of walk with carrying bag and 2 of walk with wearing-coat. In experiments, the first 4 out of 6
nm sequences of each subject are used in gallery set. Three different experiments are conducted using the remaining two sequences of nm, bg and cl in probe set separately. Performance comparison of the proposed method with the state-of-the-art methods on CASIA-B database is outlined in Tab. 2. The results show that on experiment cl, SDL [32] performs better than our algorithm, while on experiments nm and bg our method achieves the best results, with the highest average recognition rate 95.6%.

3.3. Results on CASIA-C database

The CASIA-C database contains the gait sequence of 153 subjects with four variations: normal walk (fn), slow walk (fs), fast walk (fq), and walk with a backpack (fb). The videos were captured at night using a low resolution thermal camera at 25 f/s. Each subject has 4 sequences of fn and 2 sequences of each fs, fq and fb. A total of four experiments are conducted. In the first experiment, 3 sequences of fn are used as gallery set and the fourth fn sequence is placed in probe set. In the next three experiments, fs, fq and fb forms the probe set, while the gallery set is same. The results achieved by the proposed algorithm and the state-of-the-art methods are presented in Tab. 3. In experiments fn, fq and fb our methods achieve the best results, whereas in fs RSM [35] performs marginally better than our algorithm. Our method achieves the best average recognition rate 99.8%.

The results presented in Tables 1-3 confirm the effectiveness of the proposed gait recognition algorithm. On all three databases, the proposed algorithms has shown very convincing results outperforming the state-of-the-art in most experiments. In particular, the average recognition performance of the proposed algorithm is the highest on all three gait databases.

4. CONCLUSION

In this paper, we presented a novel model-free gait recognition algorithm which exploits the spatiotemporal characteristics of a human motion. In contrast to most existing gait recognition methods, the proposed solution does not involve any human body segmentation. The proposed method extracts dense trajectories by tracking a set of points in the successive frames of the walk sequence. Local descriptors based on MBH and HOG features are computed and encoded using Fisher vector encoding. The classification is performed using linear support vector machine. The experimental results on three popular gait benchmark databases reveal that the proposed algorithm is highly accurate.

5. REFERENCES