ABSTRACT

In this paper, we attempt to enhance the overall recognition rate for view-invariant gait recognition. We propose a simple but efficient framework for this task with training gait sequences from multiple views. A most important problem in the framework is about the optimal choice for the training views, that is, how many views are enough to ensure a satisfying overall performance and how to combine these views to achieve the optimal performance. To solve this problem, we execute intensive experiments and give reasonable optimal choices based on the experimental results. Besides, the gait feature descriptor and the fusion method we develop for the framework also contribute to the promising results. We propose to use mean of Radon transforms of the silhouettes as the descriptor which is very competent for view-invariant application. Moreover, the combination of class correlation and view correlation is applied to score level fusion of results from different views. The CASIA database B which contains gait data from 11 views distributed uniformly in range of $[0^\circ, 180^\circ]$ is chosen in our experiments.

Index Terms— View-invariant gait recognition, radon transform, score fusion, linear discriminant analysis (LDA), multiple views.

1. INTRODUCTION

More and more public areas and important working places have used cameras for surveillance and monitoring purpose. And gait has its unique advantages for this particular application in comparison with other biometrics such as iris, fingerprint, face and so on. Firstly, gait data collection is none contact and unobtrusive. Secondly, gait is perceivable at a distance. Therefore, gait analysis for recognition has received significant attention ever since it is studied.

So far, a lot of algorithms have been developed for gait recognition which can be divided into view-dependent and view-invariant approaches. For view-dependent methods, most efforts have been put on shape feature analysis [1, 4, 11] and the corresponding shape descriptors are very easy to calculate. Because the projection (shape) of the 3D human body onto a 2D image plane varies when the walking direction w.r.t the camera changes (see the shape of the body through frontal view is totally different from the shape through side view), this kind of algorithms are sensitive to view change.

For view-invariant gait or action recognition, the algorithms proposed can be divided into image-based and model-based methods. In image-based methods, the main idea is to synthesis new feature from other views to the training view according to some criteria, such as optical flow [10], planar homography transformation [9], locally linear mapping [8]. For model-based approaches, the author in [7] proposed to perform 3D model reconstruction using images from multi-views, while in [6], the author used a model with five rectangles to fit the synthesized image and the joint angle between legs was used as the gait descriptor. This kind of methods can perform relatively well within a larger range of different testing views, but the computation is much more complicated for real application.

In this paper, we introduce a simple framework for view-invariant gait recognition. With training samples from multiple views available, we recognize a testing sequence against each of the training views using simple view-dependent recognition method and combine the results from each view using weighted sum fusion method. As the view-dependent approach can work efficiently within a small range of views, a significant recognition rate for all range of views should be achieved if we utilize gait data from multiple views. Then, another problem arises which is also our main concern in this paper: what is the least number of views and how to choose such views could we achieve the significant results. Optimal solution is provided based on intensive experiments. Besides, we propose a new gait feature descriptor called Radon transform based Energy Image (REI) which is very competent for our task. Experimental results on benchmark database show the superiority of the new feature as well as the framework for view-invariant gait recognition.

2. KEY PRINCIPLES IN THE FRAMEWORK

2.1. Overview

Assume gait data from $N$ different views $\theta_{g1}, \ldots, \theta_{gN}$ are available for training, our goal is to use the least $N$ with optimal views $\theta_{g1}^{op}, \ldots, \theta_{gN}^{op}$ to achieve significant recognition rate for all range of testing views $\theta_p$ (all $\theta_g$ and $\theta_p$ are w.r.t the...
camera). Fig. 1 displays the framework for our view-invariant gait recognition algorithm with \(N = 2\).

The captured videos first go through pre-processing steps, which contain walking person detection and tracking, silhouette extraction, silhouette alignment and normalization. The methodologies we used are described in [12]. Then the binary silhouettes in a gait sequence is represented by a REI.

For training REI features, they are used to determine subspaces that extract discriminative information for recognition and features from different views form different subspaces. And for testing REI features, they are projected onto each of the subspaces separately and then matched with the training features in the same subspace. At last, score level fusion is applied to the matching results and nearest neighborhood classifier is used to obtain the identity of the walking subject.

2.2. Gait Feature Representation

In this part, we will introduce a new gait feature descriptor to represent the aligned binary silhouettes in a gait sequence. It is based on Radon transform [4] and the idea of energy image [1].

The Radon transform has been demonstrated to be very useful in diverse fields. It has many different forms and a most popular one is defined as follows:

\[
R(\rho, \theta) = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} f(x, y) \delta(\rho - x \cos \theta - y \sin \theta) dx dy
\]

where \(f(x, y)\) is a 2D function, and \(R(\rho, \theta)\) is the integral of \(f\) along the line \(\rho = x \cos \theta + y \sin \theta\). For our work, the 2D function is the binary silhouette \(B(x, y)\), and the corresponding Radon transform \(R(\rho, \theta)\) is in its discrete form which is the summation of pixel intensities along the line. The center of the silhouette is defined as the reference point and the result is normalized for each silhouette. Both the definition and result of Radon transform are shown in Fig. 2.

Given a gait video sequence of \(M\) frames, the proposed new feature descriptor can be calculated as

\[
REI(\rho, \theta) = \sum_{i=1}^{M} R_i(\rho, \theta)
\]

We call it Radon transform-based Energy Image (REI). It is competent for view-invariant recognition task because it concerns not the position but the total number of bright pixels along a certain direction. Although the change of view direction can affect the shape of the silhouette, the normalization makes up for it in a way. Besides, the noise in a specific pixel seems also to have very little effect on the transform result.

2.3. Dimension reduction method and similarity measure

The size of Radon transform of a 64*44 image is 81*180 (=14580) which means the dimension of the original feature space is very high for REI representation. Besides, there are many pixels with zero values (see Fig. 2) which are redundant in similarity measure. Therefore, a subspace learning method is needed not only to extract discriminative information but also to reduce the dimension. We choose LDA [2] because it aims to find the projection directions that on the one hand maximize the distance between samples from different classes and on the other minimize the distances between samples from the same class, both of which are very useful for recognition.

Euclidean distance is used for similarity measure. We choose weighted sum fusion method to combine distances from different views and the weight should indicate the importance of each view during the combination. Let \(d_i(X, T_j)\) denote the distance between testing sample \(X\) from view \(\theta_p\) and the training sample \(T_j\) from view \(\theta_{gi}\) of the \(j\)th walking subject, then by combining distances from different training views we arrive at the final distance given as follows

\[
D(X, T_j) = \sum_{i=1}^{N} w_{ij} d_i(X, T_j)
\]

Our goal is to determine the weight \(w_{ij}\) which should be smaller if the angle between \(\theta_p\) and \(\theta_{gi}\) is very large and vice versa. Beside the view consideration, we also want to make sure that the weight \(w_{ij}\) is larger if the testing subject is the same person with the \(j\)th training subject and smaller otherwise. Therefore, we propose to calculate the view correlation and class correlation separately as follows

\[
C_{\theta_i} = Corr(X, \mu_{\theta_i}^c)
\]

\[
C_{ij} = Corr(X, \mu_{ij}^c)
\]

where \(Corr(x, y) = \frac{\sum (x - \overline{x})(y - \overline{y})}{\sqrt{\sum (x - \overline{x})^2 \sum (y - \overline{y})^2}} (l is the dimension)\), \(\mu_{\theta_i}^c\) is the mean of all samples from view \(\theta_{gi}\) and \(\mu_{ij}^c\) is the mean of all samples for the \(j\)th subject in view \(\theta_{gi}\).
Then, the weight $w_{ij}$ is calculated as

$$w_{ij} = C_{ij}^{C_v} \times C_v^i$$ (6)

### 3. EXPERIMENTS AND RESULTS

#### 3.1. Database

We use CASIA gait database (Dataset B) [5] to evaluate the performance of the framework for view invariant gait recognition. This database was taken with 11 cameras around the left hand side of the subject when he/she was walking, and the angle between two nearest view directions is $18^\circ$. This means the cameras are distributed uniformly in the range of $[0^\circ, 180^\circ]$. Gait data of 124 subjects were captured at last, among whom 93 were men and 31 were women. Every subject was asked to walk 10 times in the scene (6 normal + 2 with a coat + 2 with a bag). Thus there are a total of $10 \times 11 \times 124 = 13,640$ video sequences in the database. It is a large gait database in both number of subjects and number of views, so it is very convenient for different kind of research works on gait.

In our experiments, we only use the 6 normal walking sequences from each subject, 4 of which are treated as training set and the left 2 are treated as testing set. In each experiment, we choose the training gait sequences from the training set according to the number ($N$) and specific views that are assumed to be available ($124 \times 4 \times N$ sequences), and samples from each of the 11 views in the testing set ($124 \times 2$ sequences each view) are used for testing separately.

#### 3.2. Experimental results

##### 3.2.1. Reason for combination of multiple views

Fig. 3. Recognition power of the 11 different training views with different range of testing views. Here, we define "+" as clockwise and "-" anticlockwise w.r.t. the training view direction [13].

Firstly let’s look at the view-invariant recognition power of the 11 specific views defined in the database. Similarly as in [13], we use a fixed training view to recognize samples from other views and then regard the mean recognition rate as its view-invariant recognition power. We conduct experiments using different range of testing views and the results are shown in Fig. 3. From the figure we can see that if the angle between testing and training views is less than $18^\circ$, the recognition power of the 11 training views are all above 90%. However, the recognition power decreases greatly with the increase of the range of testing views. For the overall view range $[0^\circ, 180^\circ]$, even if we set training view as $126^\circ$ with the best overall recognition power 73.02\%, the minimum recognition rate among all the testing views is 33.47\% which is still very low.

Obviously, if gait sequences from only one view are available for training, significant recognition rates can be achieved only within a slightly view change w.r.t the training view. Therefore, it is a direct way to enhance the overall performance by using gait data from multiple views for training. And theoretically, the more number of views are available, the better overall performance is achieved. Then another problem arises: how many number of views are enough to ensure the requirement on minimum recognition rate and specifically how to choose these views to achieve the optimal performance? We will focus on this problem in the following part.

##### 3.2.2. How to choose the training views

Suppose gait data from $N$ different views are available for training, according to the rule of permutation and combination, there are $C_1^{11}$ different combinations for the choice. For each combination, we evaluate the overall performance on the 11 different testing views and record both the minimum and the mean recognition rate. The results for $N = 2$ and $N = 3$ are shown separately in Fig. 4.

![Fig. 4. Performance of each combination with minimum recognition rate and mean recognition rate for N=2 and N=3 separately.](image)

For $N=2$, the highest mean recognition rate 88.39\% is achieved by the combination of view $18^\circ$ and view $108^\circ$ with minimum recognition rate 68.15\%, while the highest minimum recognition rate 73.79\% is achieved by combination of views $108^\circ$ and $180^\circ$ with mean recognition rate 87.13\%. And for $N=3$, with gait data from views $18^\circ$, $108^\circ$ and $162^\circ$ the mean recognition rate reaches 90.69\% and minimum recognition rate 84.68\% which are both the highest among all combinations. It is obvious that the optimal performance should be achieved with training views distribute uniformly among the testing view range. From the results for $N = 2$ we can see that more than one combination may have the promise performance, therefore, it is the requirement from
Table 1. Results comparison with training views 18°, 108° and 162°.

<table>
<thead>
<tr>
<th>Difference from the proposed algorithm</th>
<th>0°</th>
<th>18°</th>
<th>36°</th>
<th>54°</th>
<th>72°</th>
<th>90°</th>
<th>108°</th>
<th>126°</th>
<th>144°</th>
<th>162°</th>
<th>180°</th>
</tr>
</thead>
<tbody>
<tr>
<td>Proposed Algorithm (REI+LDA+Weight)</td>
<td>87.9</td>
<td>96.37</td>
<td>94.35</td>
<td>87.1</td>
<td>84.68</td>
<td>91.94</td>
<td>97.18</td>
<td>88.31</td>
<td>87.9</td>
<td>97.18</td>
<td>85.08</td>
</tr>
<tr>
<td>GEI</td>
<td>41.94</td>
<td>99.6</td>
<td>44.35</td>
<td>7.26</td>
<td>37.9</td>
<td>70.97</td>
<td>95.56</td>
<td>27.02</td>
<td>3.23</td>
<td>96.77</td>
<td>38.31</td>
</tr>
<tr>
<td>WEI</td>
<td>39.52</td>
<td>96.37</td>
<td>19.76</td>
<td>6.85</td>
<td>64.11</td>
<td>48.39</td>
<td>95.56</td>
<td>6.45</td>
<td>8.87</td>
<td>94.76</td>
<td>35.48</td>
</tr>
<tr>
<td>Baseline</td>
<td>46.37</td>
<td>96.77</td>
<td>62.1</td>
<td>39.11</td>
<td>77.02</td>
<td>87.1</td>
<td>91.13</td>
<td>43.95</td>
<td>37.5</td>
<td>95.16</td>
<td>52.02</td>
</tr>
<tr>
<td>PCA</td>
<td>45.97</td>
<td>96.77</td>
<td>62.1</td>
<td>38.71</td>
<td>77.02</td>
<td>87.1</td>
<td>91.13</td>
<td>43.55</td>
<td>37.5</td>
<td>95.16</td>
<td>52.02</td>
</tr>
<tr>
<td>Min</td>
<td>86.69</td>
<td>96.37</td>
<td>94.35</td>
<td>87.1</td>
<td>83.48</td>
<td>91.54</td>
<td>97.18</td>
<td>87.5</td>
<td>86.69</td>
<td>96.77</td>
<td>85.08</td>
</tr>
<tr>
<td>Product</td>
<td>86.29</td>
<td>95.97</td>
<td>94.85</td>
<td>86.69</td>
<td>82.26</td>
<td>91.13</td>
<td>97.18</td>
<td>87.91</td>
<td>86.29</td>
<td>96.37</td>
<td>85.08</td>
</tr>
</tbody>
</table>

real application that decides which one to choose as optimal. Due to space limitation, the results for $N > 3$ will not be shown here, but it could be concluded that the performance of combination of 3 different views are already very promising.

3.2.3. Optimal performance

The optimal overall performance for $N = 3$ is shown in the first row of Table 1. As many factors can affect the performance of the framework, we will discuss them separately.

Firstly, we compare the influence of gait feature descriptor. Two other state-of-the-art descriptors (GEI[1], WEI[11]) replace REI separately in the proposed algorithm and results are shown in the second and third rows in Table 1.

We then evaluate the effect of different subspace learning methods and different fusion methods [3]. Both of the results are shown in Table 1(When we discusses the influence of one factor, we keep the other factors as the same with the proposed algorithm).

3.3. Discussion

We summarize here based on the above experiments:

(1) The simple shape feature descriptor is robust only to a slightly view change w.r.t. the training views. Therefore, it is reasonable to use gait data from multiple views for training to enhance the overall performance. Promising results can be achieved with combination of only three different views 18°, 108° and 162° respectively.

(2) Both the gait feature descriptor and the subspace learning method have much influence on the final results. The discriminative information extracted by the new feature REI is more competent for view-invariant recognition task than the other two. LDA can extract more discriminative information for recognition than PCA.

4. CONCLUSION AND FUTURE WORK

In this paper, we introduce a simple framework for view-invariant gait recognition based on gait sequences from multiple views. By using gait data from only three views for training, promising results can be achieved with mean recognition rate 90.69% and minimum recognition rate 84.68% among all testing views. Moreover, the new descriptor REI is found out to have much discriminative power for view-invariant gait recognition. In the near future, further study will be conducted to optimize the overall performance either through a new database or through theoretical analysis.

5. REFERENCES