Gait Recognition Based on Relative Distance and Motion Features

S. Arseev[1], A. Konushin[1,2]
[1]Lomonosov Moscow State University, Moscow, Russia;
[2]NRU Higher School of Economics, Moscow, Russia

In this paper we present a human gait recognition system for human identification in video sequences. It uses the structural approach, i.e. pose estimation to extract features from the sequence. Three kinds of features are used: anthropometric features, based on the length of the skeleton segments; relative distance features, based on relative distances between the skeleton joints; and motion features, based on the movement of a joint between two frames. Two versions of the algorithm are presented: the first one uses the depth data alongside with the images while the other one uses only the video seqience. We performed training and evaluation on two datasets, experimental results are presented.

Keywords: *biometrics, gait recognition, gait features, anthropometric features, relative distance features, motion features.*

1. INTRODUCTION

Gait recognition task is a subset of the biometric human recognition task. Human recognition in video can be applied in many different areas, such as security and automatic dataset markup for further machine learning tasks. However, all existing methods have limitations, preventing them from being used in certain cases. For example, the most popular method of human recognition is the face recognition, but it cannot be performed when the face is covered or if the person is facing away from the camera. Gait recognition, i.e. recognition based on features acquired from analyzing the walking subject's movement, can be performed even on bad quality video, and the only requirement is the visibility of a silhouette.

Gait recognition methods use two main distinct approaches to the task: the silhouette approach and the structural approach. The silhouette is a binary mask which contains pixels that belong to the person. Methods that use the silhouette as a main information source typically have low computational complexity (and, therefore, are fast), but they are very sensitive to the silhouette change, so if the silhouette is changed by a bag or even a long coat, it can seriously affect the recognition result.

The structural approach is based on pose estimation and construction of a pose model, typically a skeleton, i.e. graph with vertices corresponding to the joints of a human body. These methods have higher complexity due to the model construction, but are much less sensitive to the silhouette change due to the fact that if the model is produced correctly, it is not affected by that.

2. PROPOSED METHOD

The proposed method is based on the method proposed by Ke Yang et al [3]. Each sequence is pre-processed to produce a pose estimation model for each frame (a skeleton) which is used by the identification algorithm itself. The skeleton produced by Microsoft Kinect sensor is marked as shown on Figure 1.

Three kinds of features are extracted from each sequence of skeletons: anthropometric features (AF), relative distance features (RDF) and motion features (MF). Then, these features are classified using an ensemble of K-nearest neighbour (KNN) classifiers.

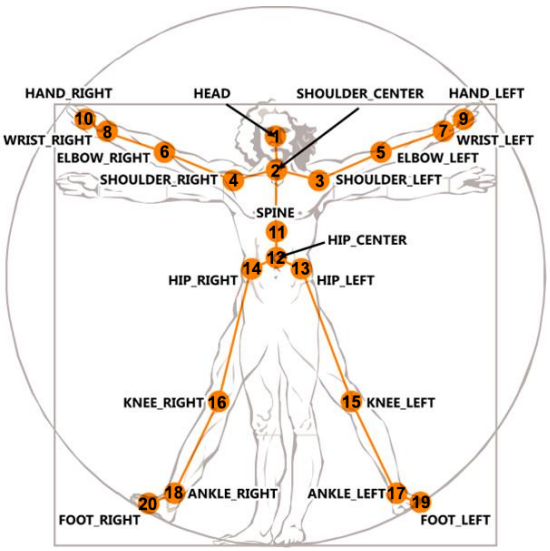

Fig. 1. Kinect skeleton.

2.1 Feature extraction

2.1.1 Anthropometric features

Anthropometric features represent skeleton segments lengths and the total height of the person. $Len(a,b)$ is the euclidean distance between joints a and b.

For the Kinect skeleton.

$$Height = Len(1,2) + Len(2,11) + Len(11,12) + \frac{Len(14,16) + Len(16,18) + Len(13,15) + Len(15,\ 17)}{2}$$

19 Len values and the $Height$ value produce the anthropometric features (AF) vector. To improve measurement precision, mean and standard deviation values are calculated for each AF component over the video sequence. Then all values which differ from the mean by more than two standard deviations are dropped out, and the mean values of remaining elements produce the final AF vector.

2.1.2 Relative distance features

Each joint a of the skeleton is described by its coordinates: x_a, y_a and z_a for skeletons with the depth map. Relative distance features represent the distances between skeleton joints. They are calculated for each axis separately, and the RDF vector consists of all distances represented as $|x_a - x_b|$, where a and b are the numbers of the paired joints, such as 3 and 4 or 15 and 16, or a is a joint in the pair and b is the corresponding middle joint (2 for a 3-10 and 12 for 13-20). It also includes features represented as $\left|x_c - \frac{x_a + x_b}{2}\right|$, where a and b are the numbers of the paired joints and c is either 1 (head joint) or 11 (spine joint). The mean and standard deviation

value of each distance over the sequence produce the final RDF vector.

For three-dimensional Kinect skeletons, the RDF vector consists of 240 elements: there are 16 pairs of joints and each pair produces 5 features for each of the 3 coordinate axis.

2.1.3 Motion features

Motion features are based on the joint shift between frames representing the joint movement speed. The shift is calculated in relative coordinates to ignore the overall movement of the person.

These features are calculated on al axis separately: x, y and z. In the two-dimensional version of the algorithm there are only two axis since there is no depth information.

$$Mx_{a,t} = \left|x_{a,t} - x_{b,t} - \left(x_{a,t-1} - x_{b,t-1}\right)\right|$$

Here, a is a joint on a limb (joints 3-10 and 13-20) and b is the corresponding middle joint (as in 2.1.2).

The mean (by t) and standard deviation values for each joint and each axis produce the MF vector. For Kinect skeletons, it consists of 96 elements: each of the 16 pairs produces 2 features for each of the 3 axis.

The difference between these three kinds of features extracted from a skeleton sequence is shown in Figure 2.

Fig. 2. Three types of features.

2.2 Sequence classification

The resulting feature vector, consisting of AF, RDF and MF vectors, is classified using an ensemble of 100 K-nearest neighbour classifiers with the city block metric. Due to the large size of the feature vector (356 features), each classifier uses a subset of 10 features.

Classifiers were selected to the enseble via an evaluation procedure. First, the KNN classifier was trained using a random subset of 10 features. Then, for each element of the testing set, five nearest neighbours from the training set were selected and the classifier was assigned a score if any of them had the same class as the test sample: 5 points if the nearest neighbour was of the same class, 4 points for the second nearest neighbour and so on with 1 point for the fifth nearest neighbour being of the same class as the test sample.

Then the classifier was compared to the classifiers that were already in the ensemble and penalized for each feature that was used by an another classifier in the ensemble to avoid using similar classifiers and neglecting features. After evaluating a batch of the random classifiers, the ensemble was reformed using 100 classifiers with best scores, including both the old ones that were already in the ensemble and the new ones. The final ensemble was built after evaluation of 2000 random classifiers.

Figure 3 shows the usage of the features by the ensemble for the three-dimensional version of the algorithm. The X axis represents the index of the feature in the feature vector with first 240 being RDF, next 20 being AF and the rest being MF. The Y axis represents the number of classifiers that used this feature. As seen from the diagram, almost all features are used in the ensemble, and some of them show clear prominence over the others. While the AF vector is the shortest of the three, its features are not dropped out and all contribute to the result.

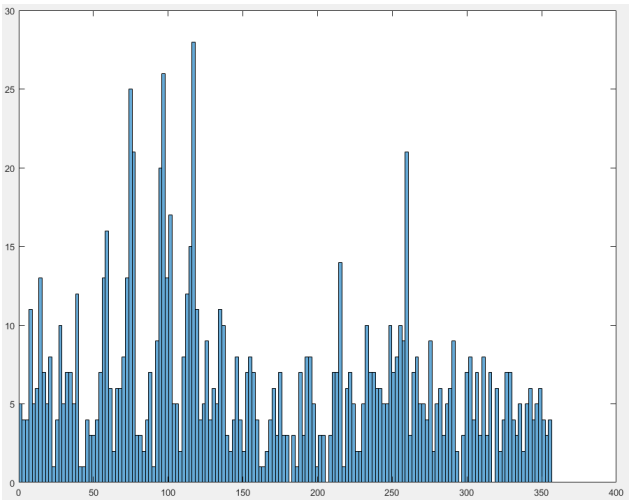

Fig. 3. Feature usage.

The classification results were produced using the weighted vote from the ensemble. Each classifier produced 5 most probable labels, and the class label was assigned a score of 5 for the first place, 4 for the second, and so on until the score of 1 if it was the fifth most probable. Then the total score of each class label was calculated, producing the ranked classification result.

3. EXPERIMENTAL EVALUATION

The algorithm has been evaluated on two datasets: the dataset used by Andersson and Araujo [1] and the TUM GAID dataset [2]. The first one was used to evaluate the three-dimensional version of the dataset and the second one was used to evaluate the two-dimensional version. The results are presented in the table.

	Rank 1	Rank 5
Kinect	93.9	100
GAID-N	95.8	98.7
GAID-B	76.5	93.2
GAID-S	87.8	94.5

Table 1. Experimental results

The results are given as the percentage of correctly classified samples. Rank-5 metric counts the sample as correctly classified if any of the 5 most probable classes was the correct one.

The “Kinect” line represents the dataset used in [1] and [3], and the following lines represent different subsets of the training set in GAID dataset: GAID-N are normal walking sequences,

GAID-B are sequences with backpack and GAID-S are sequences where the person is wearing coating shoes. As seen from the table, the algorithm shows good recognition rate both with the depth map and without it. That means that the recognition quality depends mostly on the skeleton extraction quality while the additional depth information is not strictly necessary for this method.

4. CONCLUSION

A new gait recognition algorithm based on the work [3] was presented in this paper. It uses three kinds of features for classification: anthropometric features, relative distance features and motion features. These features are then classified using an enseble of KNN classifiers, each of which uses a subspace of these features. The algorithm has been evaluated on two datasets and shows good recognition accuracy both with and without the additional information from the depth map.

5. REFERENCES

[1] Andersson V. O., Araujo R. M. Person Identification Using Anthropometric and Gait Data from Kinect Sensor // Proceedings of the Twenty-Ninth AAAI Conference on Artificial Intelligence. 2015. 1. P. 425-431

[2] Hoffman M., Geiger J., Bachmann S., Schuller B., Rigoll G. The TUM Gait from Audio, Image and Depth (GAID) Database: Multimodal Recognition of Subjects and Traits // Journal of Visual Communication and Image Representation, Special Issue on Visual Understanding and Applications with RGB-D Cameras. 2014. 25. N1. P. 195-206

[3] Yang K., Dou Y., Lv S., Zhang F., Lv Q. Relative Distance Features for Gait Recognition with Kinect // Journal of Visual Communication and Image Representation. 2016. 39. P. 209-217