"ALEXANDRU IOAN CUZA" UNIVERSITY OF IAȘI
FACULTY OF COMPUTER SCIENCE

MASTER'S THESIS

Human Gait Recognition

proposed by

Rareș-Alexandru Stan

Session: *July, 2019*

Scientific Coordinator

Lect. Dr. Ignat Anca

"ALEXANDRU IOAN CUZA" UNIVERSITY OF IAȘI
FACULTY OF COMPUTER SCIENCE

Human Gait Recognition

Rareș-Alexandru Stan

Session: *July, 2019*

Scientific Coordinator
Lect. Dr. Ignat Anca

Cuprins

1 Introduction

Gait is the movement pattern of the limbs during walking over a solid surface. It varies based on speed, terrain, maneuvering or efficiency of energy. This movement is unique for each human and can be used for recognizing persons from afar, without the need of their cooperation or physical contact, whereas fingerprint, iris or facial do need the physical access or their cooperation [1].

There are three main categories in which recognition could be classified, Machine Vision (MV), floor sensors and wearable sensors. MV is preferred because it is effective in continuous authentication and is the most non-intrusive approach.

We will create a system for human gait recognition using machine Vision and Convolutional Neural Networks, that accept a series of frames with the person walking.

2 State of the Art

Human gait is the movement pattern of the limbs during walking. It can vary depending on the persons age, weight, how tired he is and if he is carrying extra weight. A system for recognizing persons by their walking should take all of the situations from above, to correctly identify them.

There are three main approaches for identifying people by their gait, Machine Vision (MV), floor sensors and wearable sensors. Each of the three approaches have some disadvantages and advantages:

- MV:
  - it is cheaper to implement, no need to install extra sensors, just some video cameras;
  - can cover a wide area;
  - it is affected if the people are wearing voluminous clothes;
- Floor Sensors:
  - are not affected by the clothes worn by the user;
  - are more expensive to implement than MV;
  - limited area for recognizing people;
- Wearable Sensors:
  - are not limited by a specific area;
  - are not affected by the clothes worn by the user;
  - you need to have physical access or to have their cooperation.

In Machine Vision there are tow main approaches, model-free and model-based, where the first approach uses direct image sequences, whereas the latter needs more processing of the input sequence.

2.1 Model-Based Machine Vision

Molhema Mohualdeen and Magdi Baker [2] have proposed a model-based approach for the Gait Recognition problem using Region os Interest (ROI), Discrete Wavelet Transform (DWT), Edges, Gait Cycle and Neural Networks. ROI was used in the preprocessing phase to reduce data and extract the exact silhouette from each frame, by cropping.

Next, in the feature extraction phase, they used DWT for multi-scale analysis, using diagonal, horizontal and vertical details of the three levels low pass and high pass filters on two dimensions DWT. Beside 3L-2D-DWT they used Edge Detection for magnitude and orientation and box technique for step and cycle length, using the with of the bounding box. Estimating the Gait Cycle was done by combining the silhouettes between the tow main phases of Gait and combining them together for each person and measuring the combination area and the width of the white shape boundary represents the step length. Classification was done using a Back Propagation Neural Network (BPNN).

Munif Alotaibi and Ausif Mahmood [3] propose a different type of preprocessing with a Convolutional Neural Network for classification. The processioning is done using the Gait Energy Image (GEI), defined as: $GEI(x, y) = 1/s \sum_{t-1}^{s} F^t(x, y)$, where s is the total number of frames representing the Gait Cycle and $F^t(x, y)$ is the silhouette of the subject at the time interval t. For determining the Gait Cycle it is used the bounding box changes method and the silhouettes are then resized to $140 * 140$ pixels. The Neural Network has 4 pairs of Convolution and Pooling layers, each with eight $5 * 5$ filters and eight sub-sampling maps with pooling factor 2. For the activation function of the Convolutional layers it is used the Hyperbolic Tangent function. After the last Convolution and Pooling pair a Dense Layer with 124 nodes and SoftMax activation functions is used, to classify the data. For adding a new user to be recognized by the system, the old model is taken and froze the Convolutional and Pooling layers, so they are not changed during the new training period, and just the Dense Layer is modified, by adding a new node, and retrained.

Hazem El-Alfy, Ikuhisa Mitsugami and Yasushi Yagi [4] build a system in which the preprocessing is done using the Gauss Map of the silhouettes and classification they use Euclidean Distance on the feature vectors between the person to be recognized and the existing database. In more details, the Gauss Maps wore done on the silhouette's surface, evaluated locally, to overcome the lack of the third dimension and made all the normal vectors point outwards the silhouette. Gauss Mapping was done on a silhouette with it's boundary extracted then smoothed using a parametric cubic spline interpolation, for it's continuity at zero, first and second order with control points being every fifth pixel of the boundary. All of the silhouettes pixels are then Distance Transformed where the distance is calculated as follows $d = max(|x_1 - x_2|, |y_1 - y_2|)$, where (x_1, y_1) and (x_2, y_2) are two distinct pixels from the image and then are computed the contour lines of the distance map. After this the image is divided in a regular grid and for each cell in computed a histogram of all the normal vectors in that contour cell. All of the cell are combined in a feature vector and it is repeated for all contours in that image. Last all the feature vectors for that image are merged into the final feature descriptor, the NDM. All of the NDMs from a full gait cycle are integrated together, using their average for the aggregate cycle

descriptor. Over this aggregated feature vector the Euclidean Distance is calculated.

3 Contributions

There are two classes for the Gait Recognition method, using Machine Vision, Model-Base and Model-Free. In the former class there are methods that preprocess the input data or extract some other information from it, which is used for classification, whereas the latter uses the raw data or with very little preprocessing as input for classification. In this theses we have tried a Model-Based approach and propose a Model-Free one for Human Gait Recognition.

For training and testing data we have used the CASIA-B database [5][6][7] as it already has the silhouettes made, which gave us the possibility to focus on the recognition methods. For new we used only the 90° images, the rest will be used in a future test, and split the data in testing, nm01-nm04, for validation nm-05 and for testing nm-06.

We have tried two different approaches, one Model-Based, in which the input images were cropped, smoothed and then a surface curvature metric was extracted from the silhouette and then classified using a Convolutional Neural Network, and one Model-Free, where the input images were cropped and then feed to a Convolutional Neural Network for classification

Through surface curvature metric we refer to a function that can map the shape, position and altitude, from an image with three dimensional objects, to a numerical value, for each pixel in the original image.

Smoothing refers to the process of removing noise or fine-scale structures from images, resulting in a less jagged edges and sharp curves/transitions.

4 Approach

In the first part ot this chapter we will present the architecture and methods used for the Model-Based approach and after that the Model-Free system. At the start of each description there will be defined what the method uses for preprocessing and classification.

4.1 Model-Based Approach

The shape index is a measure of local curvature, derived from the eigen values of the Hessian, defined by Koenderink and van Doorn [8]. It can be used to find structures based on their apparent local shape. It maps to values in the range $[-1, 1]$, representing different shape types. It is defined as follows:

$$s = \frac{2}{\pi} * \arctan \frac{k_2 + k_1}{k_2 - k_1} \quad (k_1 \geq k_2)$$

Here (k_1, k_2) are polar coordinates described by $k_{1,2}^2 - 2H_{k_{1,2}} + K = 0$, where H is the mean curvature, measuring the spread of normals for the points of infinitesimal arcs, divided by the arch length and averaged over all surfaces, and K is the Gaussian curvature, specifying the spread of normals.

Neural Networks are a collection of connected nodes, name neurons, that for a node j, at the time interval t, with input $p_j(t)$, consist of:

- an activation $a_j(t)$,
- a threshold θ_j,
- an activation function f that computes the new activation at time $t+1$ from $a_j(t), \theta_j$ and the input $p_j(t)$, resulting $a_j(t+1) = f(a_j(t), p_j(t), \theta_j)$,
- and an output $o_j(t) = f_{out}(a_j(t))$

and with a propagation function, that computes the input $p_j(t)$, for node j, from $o_i(t)$ of predecessor neurons and has the form $p_j(t) = \sum_i o_i(t) * w_{ij} + w_{0j}$, where w_{0j} is a bias.

Learning in Neural Networks is done by first doing a feed forward, computing the output of the network, for the input $x, x \in \mathbb{R}^n$, where n is the size of the input vector, then computing the error of the entire network for the known input and propagating the error and updating the weights for each layer in the network like so

$$w_{ij}(t+1) = w_{ij}(t) - \eta \frac{\partial C}{\partial w_{ij}} + \xi(t)$$

where η is the learning rate, C is the cost function, and $\xi(t)$ a stochastic term. The cost function depends on the learning type and the activation function, usually in supervised learning the cost function is Cross Entropy. Backpropagation is done for each training sample for the desired number of epochs.

Convolutional Neural Networks beside the classic dense layers they have a convolution and pooling layer and they are primarily used for image and video related learning. Convolutional layers have multiple filters of fixed dimensions, randomly initialized in a uniform distribution. The $a * b$ filter is applied over the input x, using the dot product between the filter and sub-matrix of x, with a stride s, and the output is added with the bias term and then the activation function is applied, this happens for all the defined filters for a specific convolutional layer independently. Pooling layers reduce the size of the input by a specified factor, using a specific function, max, average or min, so that the following filters theoretically look at bigger feature of the input. Dense layers are usually used just for output labeling or additional classifying.

5 Conclusions

6 Bibliography

Bibliografie

[1] H. Srivastava, "A comparison based study on biometrics for human recognition," *IOSR Journal of Computer Engineering*, vol. 15(1), pp. 22–29, 2013.

[2] M. Mohualdeen and M. Baker, "Gait recognition based on silhouettes sequences and neural networks for human identification," *Indonesian Journal of Electrical Engineering and Informatics (IJEEI)*, vol. 6, p. 110 117, March 2018.

[3] M. Alotaibi and A. Mahmood, "Improved gait recognition based on specialized deep convolutional neural network," *Computer Vision and Image Understanding*, 2017.

[4] H. El-Alfy, I. Mitsugami, and Y. Yagi, "Gait recognition based on normal distance maps," *IEEE Transactions on Cybernetics*, 2018.

[5] S. Zheng, J. Zhang, K. Huang, R. He, and T. Tan, "Robust view transformation model for gait recognition," in *International Conference on Image Processing(ICIP)*, (Brussels, Belgium), 2011.

[6] S. Yu, D. Tan, and T. Tan, "A framework for evaluating the effect of view angle, clothing and carrying condition on gait recognition," in *Proc. of the 18'th International Conference on Pattern Recognition (ICPR)*, (Hong Kong, China), August 2006.

[7] R. He, T. Tan, and L. Wang, "Robust recovery of corrupted low-rank matrix by implicit regularizers," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, April 2014.

[8] J. JKoenderink and A. J. van Doorn, "Surface shape and curvature scales," *Image and Vision Computing*, 1992.