Computer Science and Artificial Intelligence Laboratory

Technical Report

MIT-CSAIL-TR-2003-013
AIM-2003-020
CBCL-230

August 27, 2003

Direction Estimation of Pedestrian from Images

Hiroaki Shimizu and Tomaso Poggio

massachusetts institute of technology, cambridge, ma 02139 usa — www.csail.mit.edu

Abstract

The capability of estimating the walking direction of people would be useful in many applications such as those involving autonomous cars and robots.

We introduce an approach for estimating the walking direction of people from images, based on learning the correct classification of a still image by using SVMs. We find that the performance of the system can be improved by classifying each image of a walking sequence and combining the outputs of the classifier.

Experiments were performed to evaluate our system and estimate the trade-off between number of images in walking sequences and performance.

Research support

This report describes research done at the Center for Biological & Computational Learning, which is in the Department of Brain & Cognitive Sciences at MIT and which is affiliated with the McGovern Institute of Brain Research and with the Artificial Intelligence Laboratory.

This research was sponsored by grants from: Office of Naval Research (DARPA) Contract No. N00014-00-1-0907, Office of Naval Research (DARPA) Contract No. N00014-02-1-0915, National Science Foundation (ITR/IM) Contract No. IIS-0085836, National Science Foundation (ITR/SYS) Contract No. IIS-0112991, National Science Foundation (ITR) Contract No. IIS-0209289, National Science Foundation-NIH (CRCNS) Contract No. EIA-0218693, and National Science Foundation-NIH (CRCNS) Contract No. EIA-0218506.

Additional support was provided by: AT&T, Central Research Institute of Electric Power Industry, Center for e-Business (MIT), DaimlerChrysler AG, Compaq/Digital Equipment Corporation, Eastman Kodak Company, Honda R&D Co., Ltd., ITRI, Komatsu Ltd., The Eugene McDermott Foundation, Merrill-Lynch, Mitsubishi Corporation, NEC Fund, Nippon Telegraph & Telephone, Oxygen, Siemens Corporate Research, Inc., Sony MOU, Sumitomo Metal Industries, Toyota Motor Corporation, and WatchVision Co., Ltd.

1 Introduction

In recent years many applications for automatically detecting visual objects such as obstacles, people and faces were introduced. There are, however, only a few attempts focusing on estimating the walking direction of people. In this report we describe an approach to the problem.

We consider the challenge as similar to estimating posture of a human, of a face and of hands. In all these problems, there are two basic kinds of approaches: model-based and learning-based.

A model-based approach attempts to recover a pose by analyzing input images and comparing them to available models. One of the most popular model-based approach is to construct 2D ellipsoid or stick models which are then used in a comparison driven by features obtained from input images [1][2][3]. The deformable surface in XYT space is used as a feature for analyzing gait [1]. In the work by Guo et al. [2], the skeleton of the silhouette of a walking human is obtained and then compare to a 2D stick model. In Chang [4], ribbons corresponding to arms and legs are used for analyzing gait. A statistical description of blobs is used in the people detection system developed by Wren et al. [5]. This 2D model-based approach usually requires the segmentation of the body parts of a human from the background; it also requires sequences of images in order to track the parts of the human body.

Another popular model-based approach is to use an accurate 3D model with information about the kinematic and the shape properties of the human body [6][7]. This approach is usually quite difficult since it requires an accurate prior model.

Learning-based approaches estimate directly the parameters of the pose of the human body. In these approaches, it is not always necessary to segment a explicit shape of body parts. In many cases, low-level 2D features such as shape, motion, color and position of the points of interest are used by learning-based classifiers.

In the work by Freeman [8], the x-y image moments and orientation histogram of the shape are used. Low-level optical flow induced by the motion of humans can be also used [9]. Deformable shape models are applied to the tracking of pedestrian in work by Baumberg [10]. Image pixels are sometimes used as input directly. Darrell et al.[11] use image pixels directly for pose estimation of hands. Quite a few papers deal with faces. The local models obtained from a large database of examples are used for estimating a pose of a human upper body [20]. Kumar [17] uses a linear morphable model to estimate the opening of the mouth directly from the image. In the work by Heisele et al.[19], a component-based method is used to detect faces in still images. The parameters of the rotation can be estimated by using the geometry of each of the face components. The work by Oren [12] uses wavelet coefficients as low-level features and applies Support Vector Machine classifier to them to detect pedestrians. Other classification methods as decision tree [13] and nearest neighbors [9] are also popular.

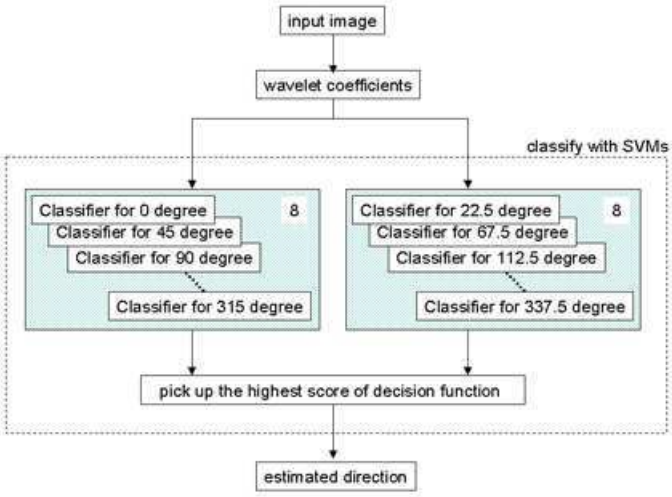

Figure 1: Overview of our direction estimation method by a single image

The approach described here starts from a single image for direction estimation and allows any background. We choose a learning-based method since the model-based methods require automatic segmentation of body parts for pose recovery. We choose a regularization technique such as Support Vector Machines because it was successfully used in many computer vision applications and well founded in statistical learning theory [14]. We use frame sequences only for improving the direction estimation. In this case we apply the same technique to each image and decide the final direction by majority vote among the classifications of each image in the sequence. In this project we do not consider the detection of people in images and assume that they have been already detected [15].

2 System Overview

We decribe the algorithm for estimating walking directions (see Fig.1). We use Haar wavelets to generate feature vectors of the input images and train 16 individual classifiers each one corresponding to certain walking direction. Before training, we separate the training data into two groups - one consisting of 8 directions such as $45.0 \times i$ $(i = 0, \ldots, 7)$ and the other consistingt of the other 8 directions such as $45.0 \times i + 22.5$ $(i = 0, \ldots, 7)$. Each individual classifier is trained on one direction. At run time, each of the trained classifiers produces a real-valued output. The system choses the most likely direction by a decision

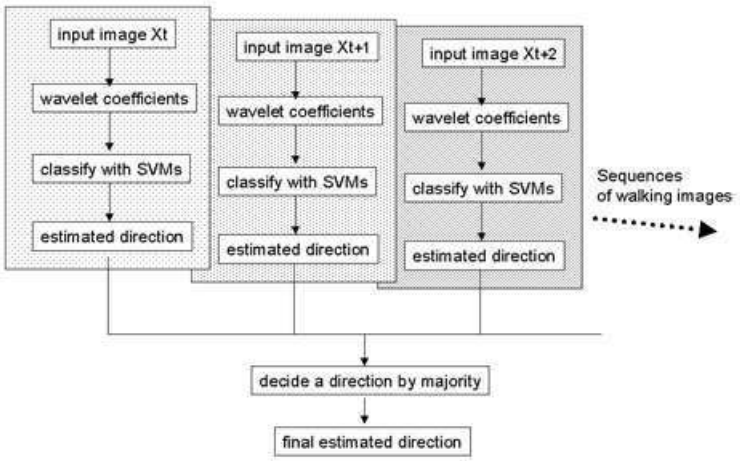

Figure 2: Overview of our method by multiple images

function which is based on the outputs of a classifier for the direction and of the two classifiers corresponding to the neighboring directions.

In order to estimate directions more accurately we apply this technique to each image of walking sequences and combine the individual classifications (see Fig.2). We explain the details in the following sections.

3 Feature extraction

Haar wavelet coefficients (8x8 pixels) are used to generate feature vectors for each image. The wavelets represent an overcomplete set at each scale since they overlap 75 percent with the neighboring wavelets in the vertical and horizontal directions [16][17]. We use three different orientations(i.e. horizontal, vertical and diagonal) of Haar wavelets. This method results in a thorough and compact representation of the input images (See Fig.3).

4 Classification

We use Support Vector Machines to classify the feature vectors resulting from the Haar wavelet representation. The choice of the kernel function usually plays an important role on the overall performance of SVM-based classification. From the results of our experiments, we chose a linear kernel function for our system.

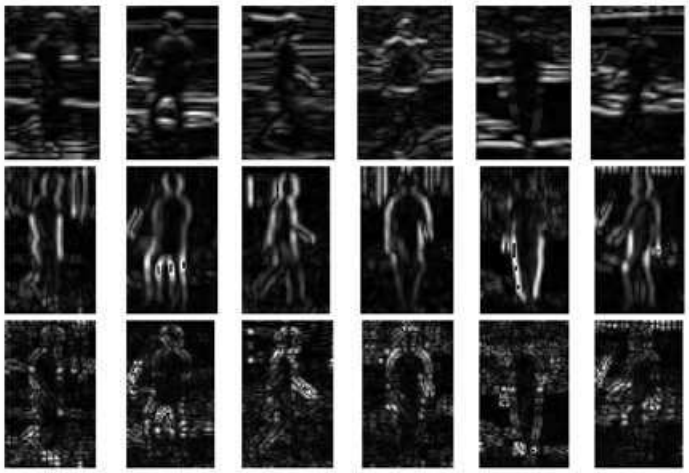

Figure 3: Samples of wavelet coefficients

In our approach we decided to classify the walking direction into one of 16 directions (i.e. 0, 22.5, 45.0, ..., 315.0, 337.5, eg every 22.5 degrees). To achieve our goal we trained 16 individual classifiers, each corresponding to one of the directions. When the system attempts to classify an image with a walking direction which does not correspond to any of the trained 16 directions, the classifier closest to the unknown direction is supposed to produce the largest output of any of the classifiers. If this assumption were true, we could assign any new direction to one of the trained directions.

We separate the training data into two groups - 8 directions such as 0, 45, 90,..., 315, and the other 8 directions such as 22.5, 67.5, 112.5,..., 337.5. Each classifier is trained on either of the groups corresponding to the appropriate direction. We chose this approach since it gives better estimation results than the alternative method of training each of the 16 classifier on a single group of 16 directions (See Fig.5).

At run time, each of the 16 classifiers of the system produces 16 outputs. The following decision function based on the outputs of the 16 classifiers is used to decide the estimated direction:

- let i be ith target direction correspond to one of 16 directions,
- let s_i be an output of ith classifier
- and let N_i be the closest neighboring directions of ith target direction.

- For any of the 16 directions, the decision function $f(s_i)$ is defined as follows:

$$f(s_i) = \omega \times s_i + \sum_{j \in N_i} s_j,$$

where ω is a weight of a target direction.

- We pick up k correspond to the highest value of the decision functions:

$$k = arg \max_i (f(s_i); i = 1, 2, \ldots, 16)$$

In the case of evaluating the 45 degrees direction, we use the outputs of the classifier for 45 degrees as a target direction and those for 22.5 and 67.5 degrees as neighboring directions:

$$f(s_{45}) = \underbrace{\omega \times s_{45}}_{target} + \underbrace{(s_{22.5} + s_{67.5})}_{neighbors}$$

According to our experiments (see Fig.6), more than 90% of testing data were classified in terms of the correct direction or one of the two closest neighboring directions. This result suggest that we may achieve more accurate results by using multiple images in a walking sequence. To estimate the direction from a sequence of images, we apply the above procedure to each image in the sequence and decide the final direction by choosing the most frequent direction (see Fig.2). When more than two directions have the same frequency, we calculate the sum of all output scores of each direction and chose the direction corresponding to the greatest sum.

5 Experiments

The training examples were obtained from the pictures of walking people taken under different lightning and in different places. The height of people in all training images were normalized to the same size. The size of each training image was 95×151 pixels. As we described in Section.4, we separated the training images of 16 directions into two groups. All of the classifiers were trained with 1000 positive and 7000 negative examples in each group. The positive examples contain the images of the direction correspond to the classifier and the negative examples contain those of the other 7 directions in the same group. For instance, the classifier for 45 degrees was trained with the images of 45 degrees as positive examples and with those of the other 7 directions (i.e. 0,90,135,180,...,315) as negative examples. The trained classifiers were run over 2400 testing images (150 images for each direction). As shown in Fig.4, one walking cycle consisted of 5 to 6 images. There is no overlap between the testing and training images.

We evaluate the recognition rates of our system as a function of the number of frames (between 1 and 10) in the walking sequences. Thus we tested 0 to

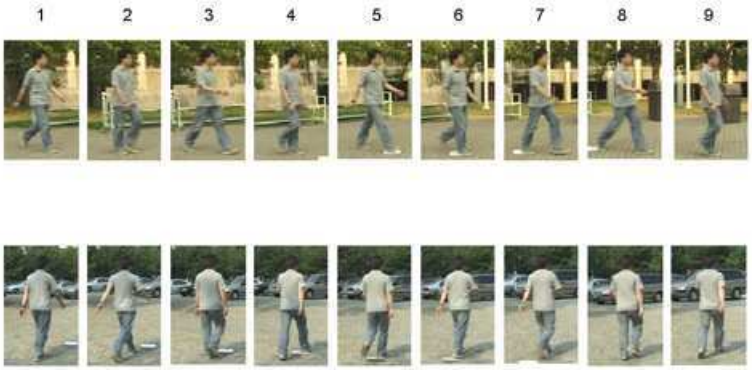

Figure 4: Two sample sequences of walking images

2 cycles of the walking sequcences. The result of our experiments is shown in Fig.7 and Fig.8. In Fig.7, we can see that 5-6 frames, which correspond to about 1 cycle of the walking sequence, is necessary and sufficient: increasing the number of frames beyond 6 does not improve the estimate of the direction. If accuracy is estimated in terms of the correct direction *and* the two neighboring ones, performance with $5-6$ frames is about the same as with a single frame, which is not surprising since the latter is already quite high. Performance in this case seems to improve with 10 frames (see Fig.8).

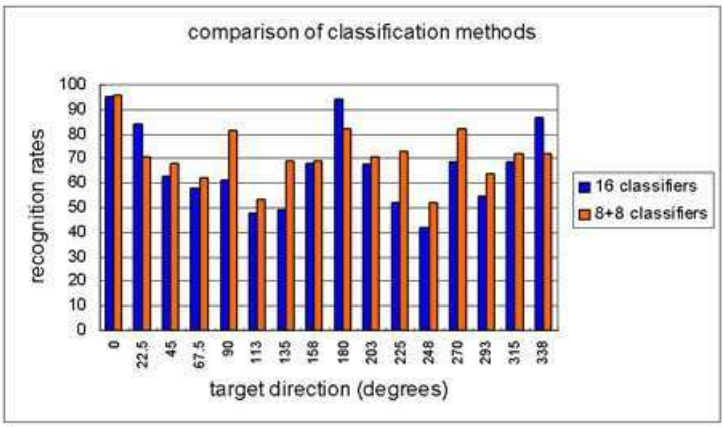

Figure 5: Recognition rates of two kinds of classification methods

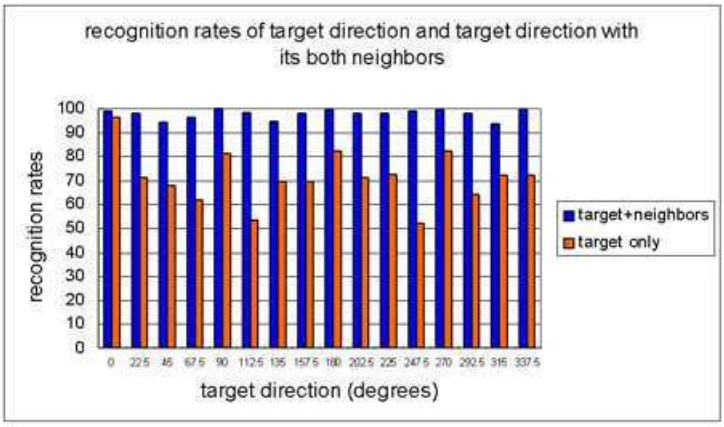

Figure 6: Recognition rates of a target only and a target + closest neighbors

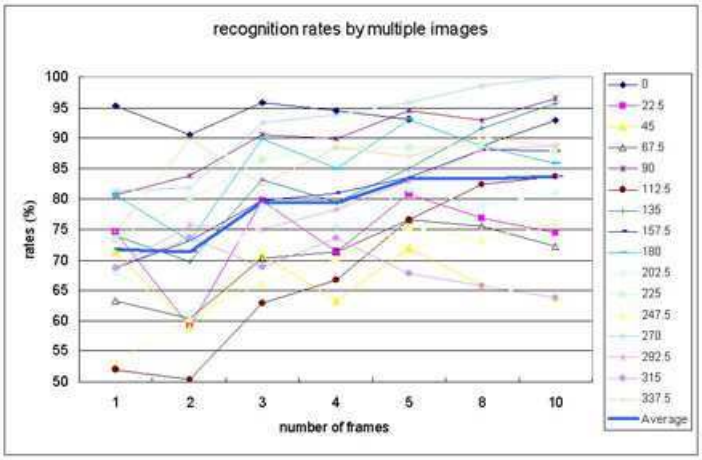

Figure 7: Recognition rates of the estimation of target directions by multiple images

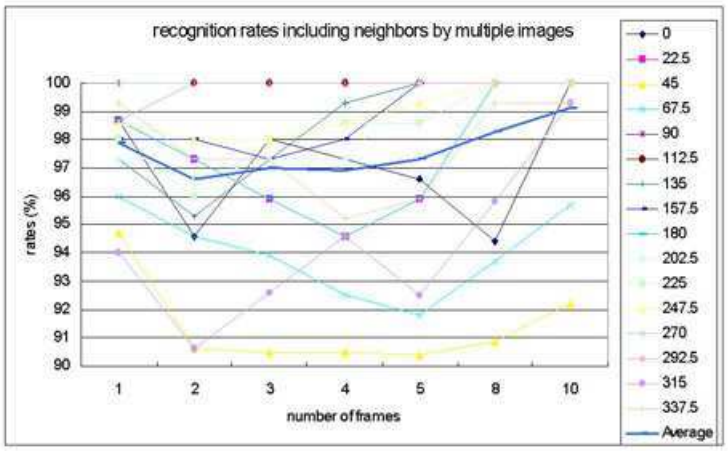

Figure 8: Recognition rates of the estimation of target and neighboring direction by multiple images

6 Conclusion

In this paper, we presented a method for estimating the direction of walking by a human from a single image. We extended this method to image sequences by applying the same technique to each frame and combining the classification results. Our approach is capable of handling variations in lightning and image background; it is capable of estimating walking direction even when only a single image is available. This may be an advantage in cases in which the system fails to track the pedestrians in a video for several frames.

We found the interesting result that a cycle of walking sequence improves direction estimation; longer sequences do not help.

As shown in Fig.5, our approach can classify more than 90% of test images into the correct direction and neighboring directions from a single image.

Acknowledgements

We would like to thank Bernd Heisele for helping us at various stages of this project and Sanmay Das for helping us to publish this paper. We would also like to thank all the members of CBCL who helped us.

References

[1] S. Niyogi and E. Adelson. Analyzing and Recongizing Walking Figures in XYT. *Proceedings of Conference on Computer Vison and Pattern Recognition, pp496-474*, 1994.

[2] Y. Guo, G. Xu and S. Tsuji. Understanding Human Motion Patterns. *Proceedings of International Conference on Pattern Recognition, pp325-329*, 1994.

[3] N. Howe, M. Leventon and B. Freeman. Bayesian Reconstruction of 3D Human Motion from Single-Camera Video. *Proceedings of Neural Information Processing Systems Conference, pp820-826*, 1999.

[4] I.-C. Chang and C.-L. Huang. Ribbon-Based Motion Analysis of Human Body Movements. *Proceedings of International Conference on Pattern Recognition, pp436-440*, 1996.

[5] C. Wren, A. Azarbayejani, T. Darrell and A. Pentland. Pfinder: Real-Time Tracking of the Human Body. *IEEE Transactions on Pattern Analysis and Machine Intelligence, Vol.19, No.7, pp780-785*, 1997.

[6] D. Gavrila and L. Davis. 3-D Model-Based Tracking of Humans in Action: Multi-View Approach. *Proceedings of Conference on Computer Vision and Pattern Recogniton, pp73-80*, 1996.

[7] D. Metaxas and D. Terzopoulos. Shape and Nonrigid Motion Estimation through Phisics-Based Synthesis. *IEEE Transactions on Pattern Analysis and Machine Intelligence, Vol.15, No.6, pp580-591*, 1993.

[8] W. Freeman, K. Tanaka, J. Ohta and K. Kyuma. Computer Vision for Computer Games. *Proceedings of International Conference on Automatic Face and Gesture Recongnition, pp100-105*, 1996.

[9] R. Polana and R. Nelson. Low Level Recognition of Human Motion. *Proceedings of IEEE Workshop on Motion of Non-Rigid and Articulated Objects, pp77-82*, 1994.

[10] A. Baumberg and D. Hogg. An Efficient Method for Contour Tracking Using Active Shape Models. *Proceedings of IEEE Workshop on Motion of Non-Rigid and Articulated Objects, pp194-199*, 1994.

[11] T. Darrell and A. Pentland. Space-Time Gestures. *Proceedings of Conference on Computer Vision and Pattern Recognition, pp335-340*, 1993.

[12] M. Oren, C. Papageorgiou, P. Sinha, E. Osuna and T. Poggio. Pedestrian Detection Using Wavelet Templates. *Proceedings of Conference on Computer Vision and Pattern Recognition, pp193-199*, 1997.

[13] L. Goncalves, E, Di Benardo, E. Ursella and P. Perona. Monocular Tracking of the Human Arm in 3D. *Proceedings of International Conference on Computer Vision, pp794-770*, 1995.

[14] V. N. Vapnik Statistical Learning Theory. *Wiley*, New York, 1998.

[15] A. Mohan, C. Papageorgiou and T. Poggio. Example-Based Object Detection in Images by Components. *IEEE Transactions on Pattern Analysis and Machine Intelligence, Vol.23, No.4, pp349-361*, 2001.

[16] C. Papageorgiou, M. Oren and T. Poggio. A General Framework for Object Detection. *Proceedings of International Conference on Computer Vision, pp555-562*, 1998.

[17] V. Kumar and T. Poggio. Learning-Based Approach to Estimation of Morphable Model Parameters. *AI Memo No.1696*, 2000.

[18] B. Heisele, T. Poggio and M. Pontil. Face Detection in Still Gray Images. *AI Memo No.1687*, 2000.

[19] B. Heisele, T. Serre, M. Pontil and T. Poggio. Component-Based Face Detection. *Proceedings of Conference on Computer Vision and Pattern Recognition, pp657-662*, 2001.

[20] G. Shakhnarovich, P. Viola and T. Darrell. Fast Pose Estimation with Parameter Sensitive Hashing. *AI Memo No.2003-009*, 2003.