A Modified Mahalanobis Distance for Human Detection in Out-door Environments

Yan Chen, Qiang Wu, Xiangjian He, Wenjing Jia,Tom Hintz
Faculty of Information Technology, University of Technology, Sydney
{jade, wuq, sean, wejia, hintz}@it.uts.edu.au

Abstract

This paper proposes a novel method for human detection from static images based on pixel structure of input images. Each image is divided into four parts, and a weight is assigned to each part of the image. In training stage, all sample images including human images and non-human images are used to construct a Mahalanobis distance map through statistically analyzing the difference between the different blocks on each original image. A projection matrix will be created with Linear Discriminant Method (LDM) based on the Mahalanobis distance map. This projection matrix will be used to transform multi-dimensional feature vectors into one dimensional feature domain according to a pre-calculated threshold to distinguish human figures from non-human figures. In comparison with the method without introducing weights, the proposed method performs much better. Encouraging experimental results have been obtained based on MIT dataset and our own dataset.

1. Introduction

Human detection is a significant step for human activity understanding and it has become more and more important in computer vision. Generally, there are two categories of approaches for human detection: approaches based on motion detection or background subtraction, and approaches that detect human directly from static images. The first category is suitable for detection of moving persons from video but fails if the person does not show any movement. The second category is generally suitable for a broader range of applications that request detecting human either from a video (image sequence) or from a single static image. Over last decades, there have been various methods proposed for human detection. Dalai and Triggs [2] adopted gradient histograms as the features to construct their classifier, which was based on the fact that the shape of human object can be well represented by a distribution of edge directions. Zhu et al. [3] improved the work of Dalai and Trigg [2] and achieved a faster human detection than Dalai and Trigg did. Viola et al. [4] used a classifier trained on human shape and motion features to detect human in static images as well as in videos. However, it was restricted to detect pedestrians with upright walking posture. Sidenbladh [6] focused on human motion patterns for robust detection because the author believed that it is harder for a person to camouflage motion but easier to change appearance. Utsumi and Tetsutani [1] observed the fact that the relative positions of various body parts are common to all humans although the pixel values may vary because of variety of clothes and illumination. They set up a Mahalanobis distance map for each image and then find the distribution based on statistics analysis. Significant elements in distance map were selected to identify human subjects against non-human subjects.

This paper proposes a modified Mahalanobis distance by introducing weights to pixels. Each pixel in the image has different contribution to the image, and the weight enlarges the gap of contribution. This will help to distinguish the object (e.g., human) from non-object more easily. That is, after considering pixel weight, pixels with relatively high contribution will be given higher weights to show their higher contribution; on the other hand, pixels with relatively low contribution will be assigned smaller weights to show their lower contribution.

Table 1 presents a summary of related work and comparison with the work presented in this paper.

The remaining sections are organized as follows. Section 2 presents our human detection scheme. Section 3 demonstrates the experiment results and discussion. The conclusions can be found in Section 4.

Table 1. Techniques used for direct human detection.

Paper	Human features	Classification	Data
Dalai et al[2].	▪ Histogram of gradients (static image)	Linear SVM	▪ Pedestrian database ▪ Different view angles
Viola et al.[4]	▪ Shape ▪ Motion (Video)	Adaboost	▪ Street Pedestrian video
Utsumi et al[1]	▪ Pixel structure based on Mahalanobis distance (static image)	Pre-calculated threshold	▪ In-door images ▪ Front view angle mainly
This paper	▪ Pixel structure based on modified Mahalanobis distance (static image)	Pre-calculated threshold	▪ Out-door image of various illumination ▪ Different view angles

2. Human Detection Based on Appearance Model

Pixel values contain significant information for object detection. But if the color or the illumination of the targets varies, the detection using pixel values can not be effective. A human figure belongs to this category because human can have various clothes. Conversely, the geometrical locations of human body parts are similar for all human figures. This paper is based on the idea that the relative positions of human body parts are common.

2.1 Weight Assignment

The contribution of each pixel to one image is different. Some pixels have high contribution while others only give little contribution. Therefore, weight can be introduced to distinguish the contribution of the pixels. Let v_{ori} denote the original pixel value, w denote the weight of the pixel, then the new pixel value v_{new} will be

$$v_{new} = v_{ori} \times w \tag{1}$$

It requires large amount of time to calculate the new pixel values if each pixel has a different weight value. It is believed that pixels in neighbourhood have similar contribution, so each image can be divided into several blocks and pixels in the same block share a single weight value. Images can be divided into blocks according to their semantics. Suppose that an image has $m \times n$ pixels, $P = \begin{bmatrix} p_{11} & p_{12} & \cdots & p_{1n} \\ p_{21} & p_{22} & \cdots & p_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ p_{m1} & p_{m2} & \cdots & p_{mn} \end{bmatrix}$ is the original matrix of the image, and $W = \begin{bmatrix} w_{11} & w_{12} & \cdots & w_{1n} \\ w_{21} & w_{22} & \cdots & w_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ w_{m1} & w_{m2} & \cdots & w_{mn} \end{bmatrix}$ is the corresponding weight matrix of the image. Then, the new image pixel matrix $P' = \begin{bmatrix} p_{11}' & p_{12}' & \cdots & p_{1n}' \\ p_{21}' & p_{22}' & \cdots & p_{2n}' \\ \vdots & \vdots & \ddots & \vdots \\ p_{m1}' & p_{m2}' & \cdots & p_{mn}' \end{bmatrix}$ will be computed by

$$P' = P \bullet W \tag{2}$$

The new value of each pixel p_{ij} ' is obtained according to Equation (1). If pixel p_{ij} and pixel p_{xy} are in the same block, the weights of these two pixels will be the same. That is, $w_{ij} = w_{xy}$. Thus a new image matrix is obtained, and the following steps are based on the new image matrix. In our current work, the blocks are labeled manually based on pre-defined semantics.

2.2 Distance Map

A revised distance map which is based on the new pixel values is constructed to describe the pixels structure of an input image which is different from the previous work shown in [1]. The new image of size $m \times n$ is divided into $M \times N$ non-overlapping blocks as pictured in Figure 1. Each block consists of $p \times q$ pixels. In our experiments, only grey values of the pixels are considered. Let $X_l (l = 1,2,\cdots MN)$ denote a matrix $p \times q$ representing every block. Each component of the matrix X_l is the new grey value of the corresponding pixel in the original image.

The difference between each pair of blocks i and j $(i, j = 1,2,\cdots MN)$ (denoted by $d_{i,j}$) is measured based on Mahalanobis distance. Let us denote the average and the covariance of all new pixel values in

the block by $\overline{x_l}$ and $\sum_l$ respectively. Then the modified Mahalanobis distance $d_{i,j}$ is computed by

$$d_{i,j} = \frac{\left(\overline{x_i} - \overline{x_j}\right)^2}{\sum_i + \sum_j} \tag{3}$$

The modified Mahalanobis distance map of the image (denoted by D) is computed, which is an $MN \times MN$ matrix.

$$D = \begin{bmatrix} 0 & d_{1,2} & \cdots & d_{1,MN} \\ d_{2,1} & 0 & \cdots & d_{2,MN} \\ \vdots & \vdots & \ddots & \vdots \\ d_{MN,1} & d_{MN,2} & \cdots & 0 \end{bmatrix} \tag{4}$$

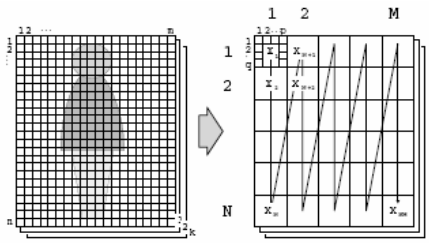


Fig. 1. Separation of image blocks [1]

2.3 System Training

In the training stage, the distance maps of all sample images are constructed based on Equations (2)-(4). Let us assume that there are K human samples and K non-human samples. Let $d_{i,j}^{obj,k}$ be the (i,j) element of k-th Mahalanobis distance map with human and $d_{i,j}^{bck,k}$ be the (i,j) element of k-th Mahalanobis distance map without human respectively, where $i,j = 1,2\cdots MN \times MN$ and $k = 1,2\cdots K$. Then the average distance of element (i,j) is computed as Equation (5) and Equation (6), where $d_{i,j}^{obj}$ is the average for human samples and $d_{i,j}^{bck}$ is the average for non-human samples. Equation (7) calculates the covariance of K human samples and Equation (8) calculates the covariance of K non-human samples. The difference at element (i,j) of the distance map between human images and non-human images is computed as Equation (9).

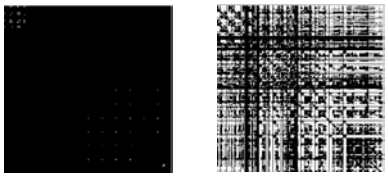
$$d_{i,j}^{obj} = \frac{1}{K}\sum_{k=1}^{K} d_{i,j}^{obj,k} \tag{5}$$

$$d_{i,j}^{bck} = \frac{1}{K}\sum_{k=1}^{K} d_{i,j}^{bck,k} \tag{6}$$

$$\sigma_{obj(i,j)}^2 = \frac{1}{K}\sum_{k=1}^{K}\left(d_{i,j}^{obj,k} - d_{i,j}^{obj}\right)^2 \tag{7}$$

$$\sigma_{bck(i,j)}^2 = \frac{1}{K}\sum_{k=1}^{K}\left(d_{i,j}^{bck,k} - d_{i,j}^{bck}\right)^2 \tag{8}$$

$$w_{i,j} = \frac{\left(d_{i,j}^{obj} - d_{i,j}^{bck}\right)^2}{\sigma_{obj(i,j)}^2 + \sigma_{bck(i,j)}^2} \tag{9}$$



(a) (b)

Fig. 2. Average distance maps. (a) human images; (b) non-human images

Two average distance maps are obtained, which are for all sample human images and all sample non-human images (see Figure 2). It can be easily observed that texture patterns of these two average maps are very different. A difference map $W = \lfloor w_{i,j} \rfloor_{MN \times MN}$ between human images and non-human images is calculated according Equation (5) to Equation (9). However, it will be computationally complicated if the whole distance maps ($MN \times MN$ elements) are used directly to differentiate human objects from non-human objects. To avoid such computation problems, a special refining procedure is designed as shown in [1] to select the most significant elements in distance maps, which show a clear difference between distance map of human images and distance map of non-human images. Initially, r largest elements are selected from the difference map W. After a coarse selection, the Linear Discriminant Method (LDM) is adopted to iteratively refine the elements selected during the coarse selection [1]. Finally, R elements are selected for human detection in the later stage. At the same time, a $1 \times R$ projection matrix A_R is created during LDM processing, which will be used to transform the multi-dimensional feature vector representing the modified Mahalanobis distance map to a one-dimensional score value for each probe image. The subject (i.e., human) classification will be carried out on the one dimensional score value domain based on a pre-calculated threshold.

2.4 Recognition Process

Given a probe image, its new pixel values will be calculated first according to the weight to each pixel and its distance map will be created based on the new pixel values. Then, R elements are selected from the distance map, of which the positions correspond to R positions located by the training stage (see Subsection 2.3 above). The score value of the probe image can be computed according to Equation (10),

$$score = A_R \times D_R^{\Pr obe} \tag{10}$$

where A_R is the projection matrix obtained in the training stage and $D_R^{\Pr obe}$ is the refined feature vector with R elements of distance map of the probe image. When the score is larger then a pre-calculated threshold, the corresponding probe image is identified as a human image. Otherwise, it is a non-human image. To calculate the threshold, all training samples consisting of human images and non-human images are fed into the recognition procedure. Obviously, the scores obtained based on Equation (10) will be clustered into two categories on the one dimensional data domain. A threshold is carefully selected to better distinguish these two classes, which maximizes the ratio of inter-class difference and intra-class difference according to Fisher's Linear Discriminant (FLD)[5].

3. Experiments

Two datasets are used to verify the performance of the proposed algorithm. One dataset is MIT's pedestrian dataset, the other one is the dataset collected by ourselves. In our own dataset, 1600 images are used for the experiment consisting of 800 human images and 800 non-human images. All these images are taken under various out-door illumination situations. The human postures are upstanding. However, there are various view angles including front view, side view and back view. The human images are against complex backgrounds including crowded. The 1600 images are divided into two parts for training and testing respectively. Each part has 800 images including 400 human images and 400 non-human images. Figure 3 shows some examples images used in the experiments. At training stage, all sample images are normalized into the size of 36×72. For human images, the human is located in the middle of the image area. In order to implement the current method on the images with different resolution, the interested area is labeled on testing images are slightly larger or smaller than the size of training samples but retain the aspect ratio. Before calculating the new pixel value and creating the distance map, the probe images are scaled down or up to match the size of training samples.

All human images including training samples and testing images are divided into four non-overlapping parts according to human body parts. They are human heads, human torsos, human legs and background. The locations of these four areas are estimated based on empirical experiments by observing many randomly selected human images. Different weight values are assigned to different parts of the images. Non-human images are also divided into four parts based on the corresponding positions of human images. Weight values are assigned by experience and those have great performance are selected. Table 2 shows some weight values and their corresponding ratio of inter-class and intra-class based on 3×3 block size and 50 selected significant elements. When all the weight values are assigned to 1, the modified Mahalanobis distance becomes to the common Mahalanobis distance. The experiments result shows that the higher weight value assigned to head area, the better performance is achieved.

(a)

(b)

Fig. 3. Examples of images used in the experiments. (a) human images; (b) non-human images

Table 2. Ratio of inter-class and intra-class.

Weight of background	Weight of head	Weight of torso	Weight of leg	Inter-class/ Intra-class
1	1	1	1	2.012
0.3	1	0.2	0.2	3.0925
0.3	1	0.3	0.3	2.9264
0.5	1	0.2	0.2	2.7194
0.5	1	0.5	0.5	2.3588
0.5	0.9	0.5	0.5	2.2718
0.2	0.5	0.9	0.5	0.9054

Besides weight values, there are other two parameters are adjusted in the experiments. One is the size of block (see Subsection 2.2), the other is the number of significant elements selected by LDM (see Subsection 2.3). By changing the above parameters, the different

recognition rates and ratios of inter-class scatter and intra-class scatter are recorded, which are calculated based on the common Mahalanobis distance and modified Mahalanobis distance.

Figure 4 shows the score distribution of human images and non-human images with the block size respectively) even when the block size is 6×6. Therefore, the modified Mahalanobis distance outperforms the common methods for distinguishing human images from non-human images, and it results in higher recognition rate and lower false positive rate.

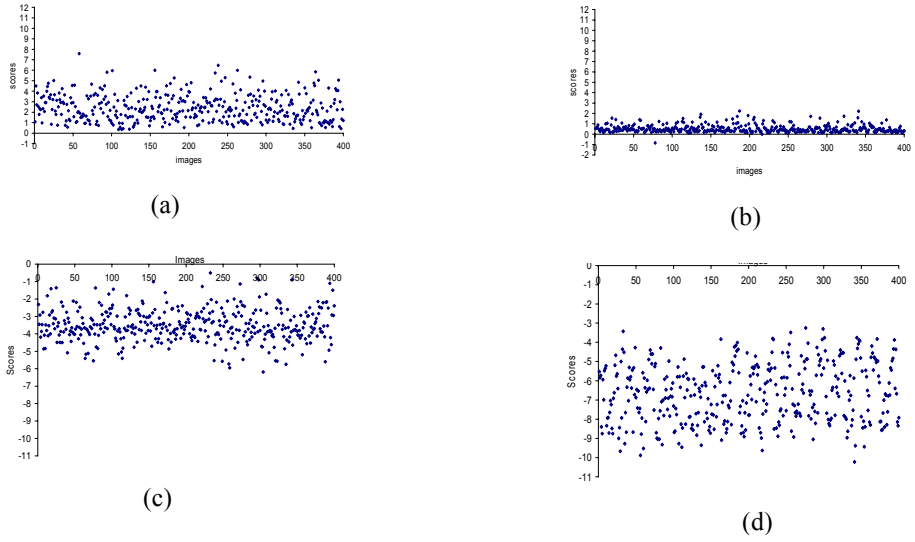


Fig. 4. Score distribution obtained by different methods. (a) human images with common Mahalanobis distance; (b) non-human images with common Mahalanobis distance; (c) human images with modified Mahalanobis distance; (d) non-human images with modified Mahalanobis distance

3×3 and 50 significant elements selected by LDM from the distance map. It is easy to see that score difference between human images and non-human images is larger when using modified Mahalanobis distance than using common Mahalanobis distance. Figure 5 compares the ratios of inter-class scatter and intra-class scatter between using common Mahalanobis distance and using Modified Mahalanobis distance (weights for head, torso, leg and background are 1, 0.2, 0.2 and 0.3 respectively). The result shows that the ratio of inter-class scatter and intra-class scatter is higher when using modified Mahalanobis distance than using original Mahalanobis distance except when the block size is 6×6 and the significant elements number is 30 due to the limitation of current experimental data. Figure 6 compares the recognition rates between using common Mahalanobis distance and using modified Mahalanobis distance. The result shows that the recognition rate is higher when using modified Mahalanobis distance (with weights for head, torso, leg and background equal to 1, 0.2, 0.2 and 0.3

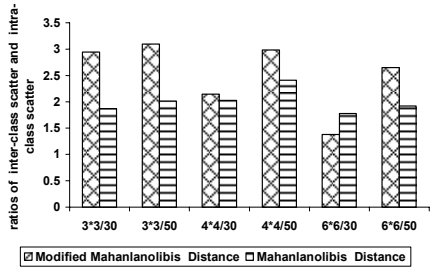


Fig. 5. Ratios of inter-class scatter and intra-class scatter obtained under different parameters with different methods

MIT's dataset contains 924 pedestrian images. It is used for recognition in this experiment. The weight performing best in our own dataset are used for MIT's dataset recognition. The weights are 1, 0.2, 0.2 for head, torso and leg respectively. Table 3 shows the results of Mahalanobis and our proposed method on MIT dataset.

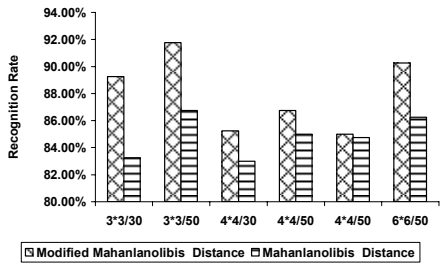


Fig. 6. Recognition rate obtained under different parameters with different methods

Table 3 Experiment results on MIT's dataset

	Recognition Rate	Ratios of inter-class scatter and intra-class scatter
Mahalanobis	85.05%	2.8769
Proposed method	89.75%	3.2107

According to the above experimental results, the modified Mahalanobis distance performance is better than the common Mahalanobis distance. The main reason is that the weight introduced to each part of an image enhances the essential pixel structure difference between human images and non-human images.

4. Conclusions

This paper has proposed a pixel structure based on a modified Mahalanobis distance to detect humans in static images. Comparison has been made to show the different performance using common Mahalanobis distance and the modified Mahalanobis distance when computing the distance map of an image. We conclude that using the modified Mahalanobis distance performs much better than the common Mahalanobis distance.

In the future, a dynamic method will be developed to obtain the weight value for each part of image. Moreover, in addition to the grey value, other pixel features such as colour and motion information will be investigated to construct more efficient geometrical pixel structure.

References

[1] A. Utsumi and N. Tetsutani, "Human detection using geometrical pixel value structures," in Proceedings. Fifth IEEE International Conference on Automatic Face and Gesture Recognition 2002, pp. 34-39.

[2] N. Dalai and B. Triggs, "Histograms of oriented gradients for human detection," in Computer Vision and Pattern Recognition, CVPR 2005. IEEE Computer Society Conference on, 2005, pp. 886-893 vol. 1.

[3] Q. Zhu, M.-C. Yeh, K.-T. Cheng, and A. S., "Fast Human Detection Using a Cascade of Histograms of Oriented Gradients," in Computer Vision and Pattern Recognition, CVPR 2006, IEEE Computer Society Conference on, 2006, pp. 1491-1498.

[4] P. Viola, M. J. Jones, and D. Snow, "Detecting pedestrians using patterns of motion and appearance," in Computer Vision, 2003. Ninth IEEE International Conference on, 2003, pp. 734-741 vol.2.

[5] T. Cooke, "Two variations on Fisher's linear discriminant for pattern recognition," Pattern Analysis and Machine Intelligence, IEEE Transactions on, vol. 24, pp. 268-273, 2002.

[6] H. Sidenbladh. "Detecting human motion with support vector machines," in. Proceedings of the 17th International Conference on Pattern Recognition, ICPR 2004. vol.2, pp.188-191, 2004.