# Comparative Study of People Detection in Surveillance Scenes

Joint IAPR International Workshops on Statistical Techniques in Pattern Recognition (SPR) and Structural and Syntactic Pattern Recognition (SSPR)

SSPR /SPR 2006: Structural, Syntactic, and Statistical Pattern Recognition pp 100-108

- A. Negre (1)
- H. Tran (1)
- N. Gourier (1)
- D. Hall (1)
- A. Lux (1)
- J. L. Crowley (1)

1. Institut National Polytechnique de Grenoble, Laboratory GRAVIR, INRIA Rhone-Alpes, France

Conference paper



Part of the Lecture Notes in Computer Science book series (LNCS, volume 4109)

## Abstract

We address the problem of determining if a given image region contains people or not, when environmental conditions such as viewpoint, illumination and distance of people from the camera are changing. We develop three generic approaches to discriminate between visual classes: ridge-based structural models, ridge-normalized gradient histograms, and linear auto-associative memories. We then compare the performance of these approaches on the problem of people detection for 26 video sequences taken from the CAVIAR database.

## Preview

es%20by%20neural%20networks%3A%20A%20review%20of%20linear%20auto-associator%20and%20principal%20component%20approaches&author=D..%20Valentin&author=H..%20Abdi&author=A..%20O%E2%80%99Toole&journal=Journal%20of%20Biological%20Systems&volume=2&pages=413-429&publication_year=1994)

12. Zhao, L.: Dressed Human Modeling, Detection, and Part Localization. PhD thesis, The Robotics Institute Carnegie Mellon University (2001)
Google Scholar (https://scholar.google.com/scholar?q=Zhao%2C%20L.%3A%20Dressed%20Human%20Modeling%2C%20Detection%2C%20and%20Part%20Localization.%20PhD%20thesis%2C%20The%20Robotics%20Institute%20Carnegie%20Mellon%20University%20%282001%29)

# Copyright information



# About this paper

Cite this paper as:

Negre A., Tran H., Gourier N., Hall D., Lux A., Crowley J.L. (2006) Comparative Study of People Detection in Surveillance Scenes. In: Yeung DY., Kwok J.T., Fred A., Roli F., de Ridder D. (eds) Structural, Syntactic, and Statistical Pattern Recognition. SSPR /SPR 2006. Lecture Notes in Computer Science, vol 4109. Springer, Berlin, Heidelberg

- DOI (Digital Object Identifier) https://doi.org/10.1007/11815921_10
- Publisher Name Springer, Berlin, Heidelberg
- Print ISBN 978-3-540-37236-3
- Online ISBN 978-3-540-37241-7
- eBook Packages Computer Science

- About this book
- Reprints and Permissions

# Personalised recommendations

# Comparative study of People Detection in Surveillance Scenes

A. Negre, H. Tran, N. Gourier, D. Hall, A. Lux, and J. L. Crowley

Institut National Polytechnique de Grenoble, Laboratory GRAVIR, INRIA Rhone-Alpes, France

**Abstract.** We address the problem of determining if a given image region contains people or not, when environmental conditions such as viewpoint, illumination and distance of people from the camera are changing. We develop three generic approaches to discriminate between visual classes: ridge-based structural models, ridge-normalized gradient histograms, and linear auto-associative memories. We then compare the performance of these approaches on the problem of people detection for 26 video sequences taken from the CAVIAR database.

## 1 Introduction

Many video-surveillance systems require the ability to determine if an image region contains people. This problem can be considered as a specific case of object classification in which there are only two object classes: person and non-person. Object classification in general is difficult because it has to face different kinds of imaging conditions. People detection is even harder due to the high variation of human appearance, gait, as well as the small size of human region which prevents face or hand recognition. Numerous efficient appearance-based approaches exist for object recognition [9, 3]. However, such techniques tend to be computationally expensive. Video-surveillance systems must run at video-rate and thus require a trade-off between precision and computing time.

To speed up the classification, simpler methods have been proposed. In [5], the authors only use compactness measure computed on the region of interest to classify car, animal or person. This measure is simple but sensitive to scale and affine transformations. Moreover, this method highly depends on segmentation, which remains a primitive problem. In [1] and [13], the contour is used to modelize deformable shapes of a person. However, the person must be represented by a closed contour. These methods strongly depend on contour detection or segmentation techniques.

This paper presents three methods for determining the presence of people in an imagette. Two methods use ridges as structural features to model people: the structural method uses a set of main human components like legs, torso, and the statistical method describes humans by modified SIFT based descriptor. The third method uses global appearance information of the detected region to discriminate between person and non-person. This method inherits strong points of

appearance based vision: simplicity and independance from the detection technique. In the following, we expose each method and compare their performance. Our objective is to show the advantages as well as drawbacks of appearance-based object classification approaches and structural feature based approaches, experimented in case of people. This comparative study motivates the use of a multi-layer object classifier to improve the detection rate.

## 2 Local Feature Extraction in Scale-Space

Everyday objects typically exhibit significant features at several different scales. To describe such structures of different sizes, images must be analysed in scale space. The scale-space representation of an image is a continuous space $L(x, y, \sigma)$ obtained by convolution of the image $I(x, y)$, with a Gaussian $G(x, y; \sigma)$: $L(x, y, \sigma) = G(x, y; \sigma) * I(x, y)$ where $G(x, y; \sigma) = \frac{1}{2\pi\sigma^2} e^{-(x^2+y^2)/2\sigma^2}$.

Natural interest points are local extrema in Laplacian scale space. Such points correspond to the center of blob-like structures and are widely used as key-points for scale invariant indexing and matching. Such a description provides a reliable method for object detection and description. However, natural interest points are well suited for compact objects, but tend to become unstable in the presence of elongated objects.

We extend natural interest points to describe elongated objects with natural interest lines. In addition of providing a more reliable scale normalization, natural interest lines also provide local orientation information and affine normalization. As with natural interest points, the value of $\sigma$ for the maximal scale corresponds to the half-width of the object. At this scale, the amplitude of the Laplacian exhibits a ridge. The mathematical definition of a ridge point on a surface is as follows: given a scale space $L(x, y, \sigma)$, a ridge point at scale $\sigma$ is a point at which the signal $L(x, y, \sigma)$ has a local extremum in the direction of the largest surface curvature. The ridge detection method used in this paper is described in full detail in [10].

## 3 Human recognition based on structural model

To represent a person in a structural manner, some authors use silhouettes [1, 5], or skeletons [6] and study changes of the model (like head, hand, legs, ...) in the time to analyse person movement. This representation strongly depends on the segmentation algorithm which is a primitive problem in computer vision. Ridges represent centerlines of an oblong structure. At an appropriate scale, it represents a skeleton of the object. Ridges at several scales capture more information about the object.

Figure 1 shows imagettes of a person extracted from a walking sequence of the CAVIAR[1] database. On these imagettes, we overlay ridges and blobs (extrema of Laplacian in 3 dimensions) detected in the region of interest. It is interesting

[1] http://homepages.inf.ed.ac.uk/rbf/CAVIAR/caviar.htm

**Fig. 1.** Different configurations of a person represented by ridges (lines) and blobs (circles) at scale $\sigma = 4\sqrt{2}$.

to see that ridges not only represent torso, legs and other significant parts of a person, but also changes in configuration of the person. We propose to model a person by using ridges representing person parts, more precisely torso and legs.

### 3.1 Extracting ridges in region of interest

Given a region of interest, we want to know at which scale ridges should be detected. If the region perfectly fits the person, the scale to detect ridges corresponding to torso is exactly equal to the half of the region width and the scale to detect ridge corresponding to legs is quarter width. This is straightforward for a rectangle. If the region is defined by a contour, the width and the height of a region are deduced from its second moments.

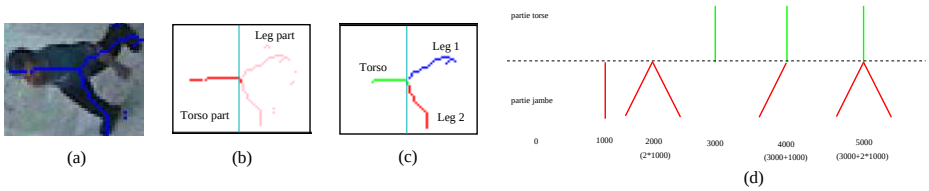


**Fig. 2.** (a) Ridges detected at scale related to the width of region. (b-c) Seleted ridges corresponding to torso and legs of person. (d) 5 configuration possibilities for each person.

Experimentation on ridge detection shows that with the use of the Laplacian, some ridges representing the same structures of objects are repeated at several scales. This also happens with persons: ridges detected at torso scale in the leg part represent well the legs (as we see in figure 1). Therefore, we propose to begin with ridges detected only at torso scale. In this manner, we only work at the scale corresponding to the size of the person.

### 3.2 Determining major ridges corresponding to torsos and legs

Knowing the orientation of a person, we cut the region into two parts by the smaller main axis (figure 2b) and take for torso part the longest ridge, the second

longest for leg part (figure 2c). The detected ridges have to be significant in energy and length. Only ridges having length and average Laplacian bigger than a threshold are considered. There may be no ridge satisfying the above condition in torso part or there is only zero/one ridge in leg part. This is the case of a person wearing a T-shirt or a trouser of same colour as the background or a partially hidden person. It is not important because it makes the model robust to partial occlusion. Using ridges, a person can be in one of the configurations presented in figure 2d.

### 3.3 Constructing descriptors

We represent a configuration of a person by a vector of 10 components determined from 3 ridges detected previously: (N, $\theta_1$, $len_1$, $dis_1$, $\theta_2$, $len_2$, $dis_2$, $\theta_3$, $len_3$, $dis_3$). The first component is the number $n$ of ridges we take from torso part and leg part of the region of interest. $n$ can be 0, 1, 2, 3. As n = 1 (torso ridge or leg ridge) and n = 2 (torso ridge + leg ridge or 2 leg ridges) do not represent an unique configuration. We assign a weight to each ridge in the model in function of its importance (for example 1 for leg ridge and 3 for torso ridge). $n$ is now converted into a sum of weighted ridge number. This means {0, 1, 2, 3, 4, 5}.

The nine following components are 3 triplets (angle between ridge and main axis, ridge length normalized to scale, distance from ridge center to region center normalized to scale). Among the ten components in the descriptor, the first component is the most significant because it represents the configuration of a person. For this reason, we give a strong weight to the first component (1000 in our experimentation), and normalize all other components by their maximal values. These values are learnt from the groundtruth: $\theta_{max} = 2\pi$, $len_{max} = 35, dis_{max} = 17$.

## 4 Ridge normalized gradient histograms

Based on observation that human silhouette can be represented by a long ridge, we propose an another approach that describes human region by a SIFT based descriptor. More precisely, we extract the main ridge to obtain a local reference invariant to orientation and scale. A gradient histogram is computed in this reference system.

### 4.1 Computing ridge properties

The first step consists in detecting and separating each ridge structure in scale space. We begin to compute ridges at each scale level as seen in the previous section. In order to obtain video-rate performance, a pyramidal algorithm described in [2] is used to compute the Laplacian scale space. Ridge structures are obtained by connected component analysis in this scale space.

We then obtain a set of ridge points $X_{n=1..N} = (x_n y_n s_n)^T$ where $x_n$ and $y_n$ represent the position in the image and $s_n$ represent the scale. In order to obtain a local reference of the ridge, we compute the first and second moments of these feature points. For more robustness, each point is weighted by its Laplacian. As we work in a down-sampled pyramid, we weight each point by $2^{k_n}$ where $k_n$ represents the stage in the pyramid. The result of ridge description is a set of ridge lines, characterized by the position of the center of gravity of the ridge points, as well as the orientation of the ridge $(x, y, \sigma, \theta)$. In the following section, we will see how to use such a representation to describe and to recognize objects.

### 4.2 Statistical Description of Ridges

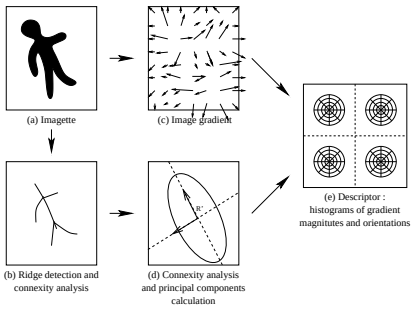


**Fig. 3.** Calculation of the ridge descriptor : ridge extraction and connectivity analysis are computed to obtain a set of ridge objects (b). The main ridge is selected and the first and second order moments are computed to obtain a local reference (d). The descriptors are then created by computing the image gradient (c), rotated by the principal direction of the ridge. The gradient orientation and magnitude are then accumulated into histograms (e).

We experiment a statistical description of ridges inspired by the SIFT descriptor [7] and Gaussian Receptive Field Histograms [8]. The descriptor is based on an array of gradient histograms. Our original contribution is to normalize each gradient measure using the intrinsic scale and the orientation of the most contrasted ridge in the imagette (cf. fig.3). After building a local reference from ridge parameters, the gradient $(L_x, L_y)$ is computed for each pixel in the imagette at a scale $\sigma_c = \alpha\sigma_i$ where $\sigma_i$ is the average scale of the ridge and $\alpha$ is a constant. A typical value of $\alpha$ is 0.5. This scale is used because we it corresponds to the boundary information of the structure described by the ridge.

Gradient magnitude is normalized by the average amplitude of the Laplacian of the ridge in order to correct for variations in illumination. The gradient orientation is rotated relatively to the orientation of $\mathcal{R}'$. This normalized gradient field of the imagette is divided into four regions, and the statistics of the gradient magnitudes and orientations for each region is collected in a histogram (fig.3(e)).

A Gaussian weighting function $\gamma$ is used to assign more importance to centered points. The function $\gamma$ is defined by the ridge properties :

$$\gamma(x,y) = e^{-\frac{x^2_{\mathcal{R}'}}{2\sigma_1^2} - \frac{y^2_{\mathcal{R}'}}{2\sigma_2^2}}$$

Where $(x_{\mathcal{R}'}, y_{\mathcal{R}'})$ are the position of the point considered in the reference $\mathcal{R}'$ and $\lambda_1$ is the greatest eigenvalue of the ridge covariance matrix. When the histogram is computed, a four-point linear interpolation is used to distribute the value of the gradient in adjacent cells, in order to minimize boundary effects. Moreover, to make comparisons, the gradient histogram is normalized in each region.

## 5 Recognizing People using Linear Auto-associative Memories

As a global approach, auto-associative memories use the entire appearance of the region of interest. The main advantage of this kind of approach is that no landmarks or model has to be computed, only the objects has to be detected. Global approaches can also handle very low resolutions. A popular method for template matching is PCA [11], but this tends to be sensitive to alignment, and the number of dimensions has to be specified. Neural nets also have been used. However, the number of cells in hidden layers is chosen arbitrarily.

We adapt auto-associative memory neural networks by using the Widrow-Hoff learning rule [12]. As in ridge extraction, the tracker detects bounding boxes and main orientation for each object in the scene. We use these informations to create grey value imagettes normalized in size and orientation as in [4]. This normalization step provides robustness to size, chrominance, alignment and orientation.

### 5.1 Linear Auto-associative Memories

Linear auto-associative memories are a special case of one-layer linear neural networks where input patterns are associated with each other. Each cell corresponds to an input pattern [12]. Auto-associative memories aim to associate each image with its respective class, and to recognize learned images when input images are degraded or partially occluded. We describe a grey-level input image by a normalized vector $x = \frac{x'}{\|x'\|}$. m images of n pixels of the same class are stored into a n x m matrix $X = \big(x_1, \ldots, x_m\big)$. The linear auto-associative memory of the class k is represented by the connexion matrix $W_k$. The reconstructed image $y_k$ is obtained by computing the product between the source image x and the connexion weighted matrix $W_k : y_k = W_k \cdot x$. We measure the similarity between the source image and a class k of images by taking the cosine between $x$ and $y_k$ : $cos(x,y) = x \cdot y^T$. A score of 1 corresponds to a perfect match. The connexion matrix $W_k^0$ is initialized with the standard Hebbian learning rule $W_k^0 = X_k \cdot X_k^T$. Reconstructed images with Hebbian learning are equal to the first eigenface of image class. To improve recognition abilities of the neural network, we learn $W_k$ with the Widrow-Hoff rule.

### 5.2 Widrow-Hoff Correction Rule

The Widrow-Hoff correction rule is a classical local supervised learning rule. It aims to minimize the difference between desired and given responses for each ell of the memory. At each presentation of an image, each cell modifies its weights from the others. Images X of the same class are presented iteratively with an adaptation step $\eta$ until all are classified correctly. This corresponds to a PCA with equalized eigenvalues. As a result, the connexion matrix $W_k$ becomes spherically normalized. The Widrow-Hoff learning rule can be described by:

$$W_k^{t+1} = W_k^t + \eta \cdot (x - W_k^t \cdot x) \cdot x^T$$

In-class images are little degraded by multiplying with the connexion matrix. In opposite, extra-class images are strongly degraded. Imagettes of the same class are used for training an auto-associative memory using the Widrow-Hoff correction rule. Prototypes of image classes can be recovered by exploring the memory. In opposite, prototypes can not be recovered with non-linear memories. Auto-associative classification of different class is obtained by comparing input and reconstructed images. The class which obtains the highest score is selected. We train two auto-associative memories for classes 0 and $n \geq 1$ persons.

## 6 Comparative Performance Evaluation

We evaluate the three techniques in the context of video-surveillance by determining if an image region contains people or not. Our training database consists of 12 video sequences which contain about 20000 people whose regions of interest are labelled in CAVIAR database. The two ridge-based methods compute human descriptors from imagettes in the training sequences and learn the descriptors by using KMeans algorithm. 34 human descriptors have been learnt in the first method and 30 in the second. The third method based on associative memories needs to learn people examples as well as non-people examples. For this, we created two sequences of the background and taken random imagettes from these sequences. Two matrices have been learnt and they are considered as people model and non-people model. For test, we use 14 sequences including 12 other sequences in CAVIAR database and 2 background sequences. These sequences contain 9452 people and 4990 non-people regions. Ridge-based methods measure the similarity as the euclidian distance between two vectors of descriptors in the first method and the $\chi^2$ distance in the second method. The third method computes directly the cosine between the imagette with the recontructed imagettes. The three similarity measures are normalized and thresholded to dertemine the presence of people.

Table 1 shows the perfomance of 4 human classification techniques: three techniques presented in the previous sections and one technique using SIFT descriptor computed at the most significative interest point detected in the imagette. This method uses the same technique for learning and testing than the

| Method | People | | Others | |
|---|---|---|---|---|
| | Recall | Precision | Recall | Precision |
| Ridge based Structual Model | 0.80 | 0.90 | 0.80 | 0.70 |
| Ridge based Normalized Histogram | 0.90 | 0.93 | 0.80 | 0.73 |
| Linear Auto-associatives Memories | 0.99 | 0.96 | 0.70 | 0.90 |
| Modified SIFT | 0.77 | 0.90 | 0.75 | 0.51 |

**Table 1.** Comparaison of recognition methods

second method. We can observe that the technique based on associative memories performs best. The reason is that this method has learnt person examples as well as non-person examples as the two first methods based on ridge learnt only person examples. If we do not train a non-people class, it gives the worst result because this method used only one model to represent all variations in the human classe. So it can not discriminate non-peolple from people. This method is good for people identification and can help for split-merge detection.

The statistical descriptor computed on ridge region gives better results than the structural descriptor. This is explained by the fact that the first method considers also one ridge as human model. Consequently, all regions containing one ridges are classified as people regions. This method requires more parameters and human knowledge than ridge histograms, but can recover people configuration. The second method gives good result in general case but presents some drawbacks when human is partially occulted or affected by light or shadow. In these cases, the detected ridge does not correspond to the global shape of the human. Therefore, the descriptor is built on nearby region but not centered on human region. Modified SIFT performs worst, because interest points are less stable than ridges for representing elongated structure like human shape. Linear auto-associative memories are disrupted when people walk through shadow areas, but can recognize configurations which do not exhibit ridges, such as people crouching down.

## 7 Conclusion

We proposed 3 different approaches for entity recognition in video sequences. Two approaches are based on local features: the ridge configuration model and the ridge normalized gradient histograms. The third one, linear auto-associative memories, is based on global appearance. Ridge normalized gradient histograms are robust to illumination changes, whereas auto-associative memories are sensitive to it. Ridge configuration models are robust to global illumination changes, but are disrupted in case of local changes. Ridge normalized gradient histograms also provide an estimation of the size and orientation of the object. As a global approach, auto-associative memories do not need to compute a model for persons and run at video-rate, but have to learn a 0 person class to be efficient.

Ridge-based approaches can be disrupted by neighboorhoods of pixels, whereas auto-associoative memories are robust to partial changes in the imagette.

We believe all three approaches can be extended to other cognitive vision problems. Ridge configuration models can be useful for gait and number of people estimation. However, this method requires specific adaptation to other object categories. Ridge normalized gradient histograms are well-suited to the discrimination of other objects, provided that these objects exhibit a main ridge. We can improve the recognition process by combining all three methods: Ridge-based methods localize objects and detect their size and main orientation using their main ridge. The image region can be normalized into a fixed size imagette to be compared to appearance prototypes constructed by linear auto-associative memories or ridge normalized gradient histograms. People configuration and gait can be described by ridge structural model.

## References


1. A. M. Baumberg and D. C. Hogg. Learning flexible models from image sequences. Technical report, University of Leeds, October 1993.
2. J. L. Crowley and O. Riff. Fast computation of scale normalised gaussian receptive fields. In *Scale Space Methods in Computer Vision*, pages 584–598, Skye, UK, June 2003.
3. J. L. Crowley D. Hall and V. Colin de Verdière. View invariant object recognition using coloured receptive fields. *Machine GRAPHICS and VISION*, 9(2):341–352, 2000.
4. N. Gourier, D. Hall, and J.L. Crowley. Estimating face orientation from robust detection of salient facial features. In *Proceedings of Pointing 2004, ICPR International Workshop on Visual Observation of Deictic Gestures*, pages 17–25, August 2004.
5. I. Haritaoglu, D. Harwood, and L. S. David. Hydra: Multiple people detection and tracking using silhouettes. In *Second IEEE Workshop on Visual Surveillance*, Fort Collins, Colorado, 26 June 1996.
6. M. K. Leung and Y. H. Yang. First sight: A human body outline labeling system. *IEEE Transaction on Pattern Analysis and Machine Intelligence*, 17(4):359–377, April 1995.
7. D. G. Lowe. Distinctive image features from scale-invariant keypoints. In *IJCV*, volume 60, pages 91–110, 2004.
8. B. Schiele and J.L. Crowley. Recognition without correspondence using multidimensional receptive field histograms. 36(1):31–50, January 2000.
9. C. Schmid. *Appariement d'images par invariants locaux de niveaux de gris*. PhD thesis, Institut National Polytechnique de Grenoble, 1996.
10. H. Tran and A. Lux. A method for ridge detection. In *Asean Conference on Computer Vision*, pages 960–966, Jeju, Korea, January 2004.
11. M. Turk and A. Pentland. Eigenfaces for recognition. *Cognitive Neuroscience*, 3(1):71–96, 1991.
12. D. Valentin, H. Abdi, and A. O'Toole. Categorization and identification of human face images by neural networks: A review of linear auto-associator and principal component approaches. *Journal of Biological Systems*, 2:413–429, 1994.
13. L. Zhao. *Dressed Human Modeling, Detection, and Part Localization*. PhD thesis, The Robotics Institute Carnegie Mellon University, 2001.