A Novel Shape Feature for Fast Region-based Pedestrian Recognition

Ali Shahrokni† † Computational Vision Group, University of Reading, UK
Darren Gawley‡ ‡ School of Computer Science, The University of Adelaide, Australia
James Ferryman†

Abstract

A new class of shape features for region classification and high-level recognition is introduced. The novel Randomised Region Ray \((\text{RRR})\) features can be used to train binary decision trees for object category classification using an abstract representation of the scene. In particular we address the problem of human detection using an oversegmented input image. We therefore do not rely on pixel values for training, instead we design and train specialised classifiers on the sparse set of semantic regions which compose the image. Thanks to the abstract nature of the input, the trained classifier has the potential to be fast and applicable to extreme imagery conditions. We demonstrate and evaluate its performance in people detection using a pedestrian dataset.

1. Introduction and Related Work

This paper introduces a new class of shape features for region classification and high-level recognition. In particular we address the problem of human detection using an abstract representation of the scene. Segmented images provide semantically meaningful components that form the basis of recognition for objects of interest in the scene. Our proposed method is based on the observation that humans can recognise and discern objects from their crude silhouette in poor visibility.

Traditionally, recognition based on region detection has been hampered by the sensitivity of feature extraction to segmentation error. However, recent advances in reliable image subregion extractions [2] has inspired region-based recognition methods. Notably, Gu et al. [5] recently introduced a unified framework for detection, segmentation and classification based on detected regions. In spite of the promising results of the above methods which exploit spatial semantics this area is vastly unexplored and remains an active research domain. To that end, we explore classification trees [6] which are established as as fast and reliable \textit{appearance} descriptors for object classification [4]. We investigate their application to \textit{shape-based} object recognition for the specific task of object detection. This leads to introduction of novel binary features which we refer to as \textit{Randomised Region Rays} \((\text{RRR})\) \(^1\) features and are used efficiently for region classification. This methodology is novel and unique in its approach to recognition through specific shape characteristics.

Instead of processing individual pixels or patches around geometric features for detection, we base our approach on the concept of semantically meaningful components of image such as superpixels [3]. The main contribution of this paper is to design dedicated features for recognition based on a crude representation of the scene. We therefore do not rely on pixel values for training classifiers, instead we design and train specialised classifiers on the set of superpixels which is far more sparse than the set of pixels in the image.

The advantage of such a system is two fold. First it enables classification and object detection (such as human body) based on a crude representation of the scene. This is of essential importance in low visibility situations and when the camera is moving in a rapid and unpredictable manner where traditional background models would fail. The second advantage of this approach is that it would naturally lead to a dramatic reduction of the computational costs of the algorithm due to the smaller amount of input data to process.

In the remainder of this paper we introduce our novel shape features based on segmented input images. We then discuss training and inference using these features and present experimental results and evaluations by comparing our approach with a state-of-the-art people detector.

\(^1\)Pronounced ‘Arrr’

\(^*\)This work was supported in part by BAE systems (Operations) Limited.
The RRR features defined above are very simple and fast to compute on superpixels. Furthermore, their definition based on relative length of the rays makes them invariant to scale. We use Bresenham’s algorithm to compute the rays, \( r \), efficiently. In the next section we show how these primitive features can be used to classify superpixels and enable object recognition.

## 2. Parts-based Human Body Recognition

We train randomised trees on different classes of superpixels which represent essential parts of the object class in question, i.e. human body. We subdivide the body into three parts (upper body, middle body and lower body) and train classifiers to distinguish between different body parts as well as background superpixels. The number of parts can be a parameter and vary depending on system requirements. For instance if the body size is small in the image and typically a few superpixels covers the body, then the body can be modelled by one or two parts only. In general if there are \( n \) parts in the object class model, the classifier will be trained on \( n + 1 \) classes (including background).

For the training and evaluation of the developed algorithm, we use the Penn-Fudan Pedestrian database which is a publicly available pedestrian image dataset \(^2\) with ground truth masks. This dataset consists of 170 images with 345 labeled pedestrians. The extracted superpixels in the labelled training database provide different instances of each of these main parts. The labelled ground truth parts are further used to obtain a geometrical distribution model for the body parts with respect to each other. This model will encode the spatial relationship between object parts. Specifically in the case of the 3-part human body, it models the distribution of the upper, middle and lower body parts as 2-D Gaussian distributions with respects to the middle body mean centroid. This model will be used for the global inference of the body part positions given the individual responses of super pixels.

### 2.2. Training

We start by extracting a set of superpixels, \( S = \{S_i \mid i = 1, \ldots, N\} \), in the input image. The next step is to train a set of randomised trees, \( F = \{F_k \mid k = 1, \ldots, K\} \), to classify superpixels into known object parts or background. Similar to [7], \( F_k = \{f_\sigma(k, 1), \ldots, f_\sigma(k, D)\} \) represents the \( k^{th} \) tree in the "forest" \( F \) and \( f_\sigma(k, j) \) is a set of random RRR features, \( \sigma \), at depth \( j \) of the \( k^{th} \) tree.

\(^2\)http://www.cis.upenn.edu/~jshi/ped.html/

![Figure 1](image.png)

Figure 1: (a) RRR feature defined on superpixels using two angles and rays. (b) Training database is used to build classification trees for foreground objects.

## 2. Recognition Framework

Randomised trees [6] and ferns [7] have been successfully used in detection and tracking of patches and have been applied to real-time tracking and recognition of textured objects with large motion and appearance changes. Inspired by the basic idea of randomised classifiers employed in such methods, we propose a new classifier based on shape features of uniform regions in the scene defined by superpixels [3]. These superpixels form the input to our classification system. We therefore do not rely on pixel values in the classification and recognition stage. We then combine the classification results in a joint probability framework to infer the possible locations of objects of interest in the scene. While our approach is generic and can be applied to any object category, for the purpose of this work we focus on human body detection. The details of our proposed model are explained in the next section.

### 2.1. RRR Shape Features

We introduce Randomised Region Ray (RRR) features which are in the form of binary questions that can be used to collectively describe the geometrical form of superpixels. Each RRR feature casts two random rays from the centre of the superpixel to form a binary decision that can be used to train decision trees on the shape characteristics of the superpixels. One such feature is illustrated in Fig. 1. angles \( \alpha \) and \( \beta \) are randomly determined from the axis of reference (the vertical line). Binary trees are then trained to classify superpixels into different classes based on the comparison of the length of the two rays at angles \( \alpha \) and \( \beta \). Here we denote a ray defined on superpixel \( S_i \) with angle \( \theta \) by \( r(S_i, \theta) \). Each RRR feature at superpixel \( S_i \) can be expressed as:

\[
RRR(S_i, \alpha, \beta) = \begin{cases} 1 & \text{if } r(S_i, \alpha) > r(S_i, \beta) \\ 0 & \text{otherwise} \end{cases} \quad (1)
\]
Examples of the forest classification for individual superpixels are shown in Fig. 4-c. Likelihoods of superpixels belonging to body parts are computed using the trained forest of classification trees on superpixels shown in Fig. 4-b. While the colours in Fig. 4-b are randomly selected, the colours in Fig. 4-c encode the probability of each superpixel belonging to upper body (red), middle body (green) and lower body (blue) and the most likely body part label is shown in Fig. 4-d. It can be noted that areas with human presence is highlighted by the higher likelihood value which is independently computed for each superpixel using the trained decision forest. This implies that the RRR features are capable of learning the distinctive characteristics of each body part. In the next section we show how these individual responses for each superpixel can be aggregated for higher level inference.

2.4. Foreground Inference

Target inference is done by applying the Generalised Distance Transform [2] to the classifier outputs for individual superpixels. For each detected superpixel $S_i$, we obtain a class distribution for each body part and the background. Using the learnt Gaussian geometric distribution model of body parts of Section 2.2, Generalised Distance Transform can then be used efficiently to compute the maximum a posteriori probability of the objects (MAP) given by $P(O_l | S_1, \ldots, S_N) = P(S_1, \ldots, S_N | O_l)P(O_l)$, where $O_l$ is part $l$ of the object category, i.e. human body.

3. Experimental Results

We have tested the proposed algorithm and the developed RRR features using a leave-one-out test on the Penn-Fudan Pedestrian Database. The classification forest is composed of 40 trees with 9 levels of depth. These trees are trained on $K - 1$ images and tested on one image at a time, where $K$ is the number of images in the dataset (i.e. $K = 170$). This enables us to independently evaluate the algorithm 170 times. Fig. 2-a shows the overall response of the part-based model computed by Generalised Distance Transform on individual superpixel classifications by the trained forest. Higher value of the red colour component corresponds to higher probability of human body presence. Fig. 2-b shows the detection result of the 3 body parts of the highest rank detection (upper body, middle body and lowerbody centroids are marked). We can use non-maximum suppression or similar algorithms to detect multiple people in the image as shown in Fig. 5.

The ROC curve of performance on the Penn-Fudan dataset was also computed to plot the ratio of true positives vs. the fraction of false positives for training dataset as well as the leave-one-out test experiment. Both these experiments include 170 results and the classification score is used to draw the curve. For the purpose of comparison, we have also computed the performance of Histograms of Oriented Gradient (HOG) descriptors with linear SVM classifier [1] using the same dataset. To that end we used the OpenCV implementation of HOG-SVM and adjusted the hit threshold to 0.5 and group threshold to 0 to improve the performance without grouping the detections. The results are shown in Fig. 3 and show comparable performance and improvements using our introduced RRR features and superpixels without relying on pixel-level data for inference.

4. Conclusion and Future Work

This work is motivated by the importance and challenges of development of specific classification features that are suitable for the loosely segmented input views. We introduce a new approach for object classification
and detection which works on an abstract representation of the input image and uses novel Randomised Region Ray features and binary decision trees for object classification. The RRR features are very easy to compute and are scale invariant. The input image can be a superpixel segmentation or any crude representation of the scene. These representations can either come directly from the sensing device through built-in processor/filters or can be efficiently computed prior to classification. The trained classifier has the potential to be fast and applicable to extreme videography conditions where the camera is mounted on a mobile platform such as UAVs or has poor visibility. As a result, the computational costs of the RRR-based superpixel classification are substantially lower due to the simplicity of the RRR features themselves as well as the sparsity of the superpixels compared to pixel-level cues and classification algorithms. The typical non-optimised processing time of a superpixel image by the RRR classification forest is around 500ms.

The results obtained on the Penn-Fudan Pedestrian database suggest that the approach is promising and is capable of detecting humans using only a sparse set of superpixels as input. Furthermore, we can see that the RRR-based classification has comparable performance to existing algorithms that use pixel-level information for classification.

As the results indicate, the classifier performs better on lower and middle body parts. This might be due to the fact that the upper body (head area) is less significant in size in relation to the other parts. Another issue is that some superpixels bleed into other parts of the image and can have negative impact on the learning process. Possible solutions might involve modification of the underlying superpixel-computation technique to obtain more well-defined superpixel inputs. These issues would be addressed in the future work.

References


