A survey of techniques for human detection from video

Neeti A. Ogale

Department of Computer Science
University of Maryland, College Park, MD 20742
neeti@cs.umd.edu

Abstract. Recent research in computer vision has increasingly focused on building systems for observing humans and understanding their appearance, movements, and activities, providing advanced interfaces for interacting with humans, and creating realistic models of humans for various purposes. In order for any of these systems to function, they require methods for detecting people from a given input image or a video. In this paper, we discuss a representative sample of techniques for finding people using visual input. These techniques are classified with respect to the need for pre-processing (background subtraction or direct detection), features used to describe human appearance (shape, color, motion), use of explicit body models, learning techniques, ...

1 Introduction

A important stream of research within computer vision which has gained a lot of importance in the last few years is the understanding of human activity from a video. Understanding human activity has applications in various fields, the most important of which is surveillance. Other applications include character animation for games and movies, avatars for teleconferencing, advanced intelligent user interfaces, biomechanical analysis of actions for sports and medicine, etc. Before the complexity of human activity can be understood, we first need automatic methods for finding humans in an image or a video. Once the human is detected, depending on the application, the system can do further processing to go into the details of understanding the human activity. This paper selects a representative sample of papers from the broad literature on full-body human detection, and presents a review and classification of the various methods. It is not intended to be comprehensive, and does not deal with specialized domains such as detection of faces, gestures or characterizing human activity, each of which possess an extensive literature of their own.

2 Classification overview

Our problem is to find people in a given video (or an image). The relevant literature can be divided into techniques which require background subtraction or segmentation and techniques which can detect humans directly from the input without such pre-processing.

Background subtraction techniques usually find the foreground object from the video and then classify it into categories like human, animal, vehicle etc., based on shape, color, or motion or other features. Here, we review the following techniques which perform human detection after background subtraction (see Table 1).

Table 1. Methods using background subtraction

Paper	Background subtraction	Human feature
Wren et al. [1997]	Color/Ref. image	Color,contour
Beleznai et al. [2004]	Color/Ref. image	Region model
Haga et al. [2004]	Color/Ref. image	F1-F2-F3
Eng et al. [2004]	Color/Ref. image	Color
Elzein et al. [2003]	Motion/Frame diff.	Wavelets
Toth and Aach [2003]	Motion/Frame diff.	Fourier shape
Lee et al. [2004]	Motion/Frame diff.	Shape
Zhou and Hoang [2005]	Motion/Frame diff.	Shape
Yoon and Kim [2004]	Motion + Color	Geom. Pix. Val.
Xu and Fujimura [2003]	Depth	Motion
Li et al. [2004]	Depth	Shape
Han and Bhanu [2003]	Infrared	IR+color
Jiang et al. [2004]	Infrared	IR+color

Direct techniques operate on (features extracted from) image or video patches and classify them as human or non-human. We can also classify techniques based on the features which are used to classify a given input as human or not. These features include shape (in the form of contours or other descriptors), color (skin color detection), motion, or combinations of these. Here is a list of the techniques we review (see Table 2).

Table 2. Methods based on direct detection

Paper	Human model	Classifier
Cutler and Davis [2000]	Periodic Motion	Motion similarity
Utsumi and Tetsutani [2002]	Geom. Pix. Val.	Distance
Gavrila and Giebel [2002]	Shape template	Chamfer dist.
Viola et al. [2003]	shape+motion	Adaboost cascade
Sidenbladh [2004]	Optical flow	SVM (RBF)
Dalal and Triggs [2005]	Hist. of gradients	SVM (Linear)

3 Techniques using background subtraction

3.1 Wren et al. [1997]

This paper describes the real-time Pfinder system for detecting and tracking humans. The background model uses a gaussian distribution in the YUV space at each pixel, and the background model is continually updated. The person is modeled using multiple blobs with spatial and colors components and the corresponding gaussian distributions. Since the blob is dyanamically changing, its spatial parameters are constantly estimated with a Kalman filter. Then, for each image pixel, the method evaluates the likelihood that it is part of the background scene or the blob. Each pixel is then assigned to the blob or the background in the maximum a posteriori (MAP) sense, followed by simple morphological operations. After this step, the statistical models for the blob and background texture are updated. Person blob models are initialized using a contour detection step which attempts to locate the head, hands and feet. Hand and face blobs are initialized with skin color priors. This system is geared toward finding a single human, and makes several domain-specific assumptions. It was tested in several HCI scenarios and is real-time.

3.2 Beleznai et al. [2004]

This paper treats the intensity difference between an input frame and a reference image as a multi-modal probability distribution, and mode detection is performed by using mean shift computation. The mean shift computation is performed in a fast way using integral images or summed area tables, which gives the method real-time performance in a manner which is independent of the size of the

window used. The mode detection procedure is able to locate isolated humans, but for separating partially occluded and grouped humans, a model-based validation process is used. The human model is very simple and consists of three rectangular regions. Within each cluster of humans, a maximum likelihood configuration of humans is identified.

3.3 Haga et al. [2004]

In this paper, a moving object is classified as human based on the spatial uniqueness of the image motion (called criterion F1 by the authors), temporal uniqueness of the human motion (F2), and the temporal motion continuity (F3). First, the moving object is detected by background subtraction, and then F1, F2, and F3 are evaluated. The spatial uniqueness of image motion is a measure of uniformity of local motion within a region. Temporal uniqueness is correspondingly defined in the time direction. A linear classifier separates human and non-human data in the F1-F2-F3 space, and is used to classify new input data.

3.4 Eng et al. [2004]

This paper proposes a combination of a bottom-up approach based on background subtraction and a top-down approach incorporating a human shape model as a solution to the problems of detecting a partially occluded human and multiple overlapping humans. First, a region-based background model is constructed under the assumption that each region has a multi-variate gaussian probability distribution over the colors. The background models are constructed in a simple manner using a set of background frames which is separated into square blocks using a k-means algorithm. Pixels in a new input image are compared with this background model and classified as foreground or background. The missing foreground parts are added by using color-based head and body detection. Then, a bayesian formulation is applied based on a simple model of the head and body as two ellipses, and all head and body pairs are determined based on the maximum a posteriori. The experiments presented in this paper deal only with a specific domain involving surveillance of a swimming pool.

3.5 Elzein et al. [2003]

The method in this paper first detects moving objects by computing optic flow only in regions selected by frame differencing. The optic flow velocity is then used to compute a time to collision with respect to a fixed reference point in the image. This is done because the goal is to detect regions which will potentially collide with the vehicle on which the camera is mounted, which is treated as the reference point. Pixels with a small time-to-collision are selected using thresholding, and morphological operations are used to construct groups or blobs of connected pixels. The resulting blobs are reshaped into rectangular regions which are then used for further processing. To determine if a selected rectangular region is a person, the authors train a classifier using wavelet-based features and a template matching scheme. Using a training database of pedestrian images, templates are constructed which are basically a normalized table of wavelet coefficients. The final template consists of a 49 dimensional feature vector, which is compared against a similar feature constructed for each input rectangles. If the number of similar coefficients are greater than a threshold, the rectangle is classified as a pedestrian. Clearly, since the input rectangles may be of different size, the matching is performed at several scales. The proposed method does not have real-time performance.

3.6 Toth and Aach [2003]

The method presented in this paper first performs illumination-invariant background subtraction by using frame differencing, window-based sum of absolute differences (SAD) aggregation, and an adaptive threshold. The authors use a Gibbs-Markov random field to create spatially varying thresholds which lead to smooth foreground shapes. The foreground blobs are identified using connected components, and the fourier transform is applied to the boundary shape. By retaining the first ten fourier components, a compact fourier shape descriptor is obtained. The classification of the blobs into human, vehicle or clutter is performed by a four layer feedforward neural network, with fourier descriptors as inputs and classes as outputs. The network is trained using 400 human examples and 400 vehicle

examples. The implementation uses OpenCV to achieve near real-time performance.

3.7 Lee et al. [2004]

In this paper, a shape-based approach for classification of objects is used following background subtraction based on frame differencing. The goal is to detect the humans for threat assessment. The target intruder is classified as human or animal or vehicle based on the shape of its boundary contour. The system classifies the contour of the object into different categories using the following procedure. The data points on the contour are reduced by a curve evolution technique which uses a relevance measure to remove vertices from the contour. By this method, the contour is reduced to 60 data points, which basically amounts to a polygon approximation expressed as bend angle vs. normalized length. The similarity between contours is measured using the L2 norm. For this, a new fast matching algorithm is developed, which can be used to classify the object as human, animal or vehicle.

3.8 Zhou and Hoang [2005]

This paper presents a method to detect and track a human body in a video. First, background subtraction is performed to detect the foreground object, which involves temporal differencing of the consecutive frames. After this step, the classification of the object is based on two approaches: the first is a codebook approach, and the second involves tracking of the object and if the object can be tracked successfully, it is considered to be a human. For the first approach, the size of foreground blob is normalized to 20x20, and then a shape feature vector of the foreground object is created. In order to create the shape vector of object, the mask image and boundary of human body are created. The distance from the boundary of human body to the left side of bounding box is used as feature vector. This is compared against the feature vectors of the human images in the codebook. The minimum of all distortions for the all the features vectors in the codebook is found, and that if that is less than threshold, then the object is classified as human. Tracking is based on color

histograms, motion and size of the foreground blob. False alarms due to static oscillatory motions are also detected and removed, to handle objects like shaking trees. Other features of the technique include shadow removal.

3.9 Yoon and Kim [2004]

A composite approach is proposed in this paper for human detection, which uses skin color and motion information to first find the candidate foreground objects for human detection, and then uses a more sophisticated technique to classify the objects. After the candidate human regions are detected, they are size normalized based on the distance of a point to the centre of gravity in the vertical direction and the starting and ending x-y position in the horizontal direction. The paper only considers the human upper body appearance for the detection. The upper body consist of clothes which have different colors and textures. Thus only a color based approach cannot be used for further classification and a geometric pixel value structure approach is used. The size normalized image is divided into non overlapping subparts and Mahalanobis distance between the blocks is calculated, similar to Utsumi and Tetsutani [2002]. Then, from the distance map images, PCA is performed to reduce the dimensionality and a SVM based classifier is used to classify humans and non-humans. However, no details are given about these steps.

3.10 Xu and Fujimura [2003]

The authors present a novel approach to detect the pedestrians, which is shown to work well in a indoor environment. They make use of a new sensing device, which gives depth information along with image information simultaneously. From the depth image, the part of the image between the specified depth values (Dmin and Dmax) is selected. After this, preprocessing on this image is performed which eliminates background areas like walls. This is done based on the fact that these background objects are large textureless areas of the image and they are usually partially present in the selected area between Dmin and Dmax. A split and merge algorithm is then used to perform segmentation by depth slicing which splits the depth layers

and then regions are merged based on depth continuity. By this step, objects including humans and other objects like foreground furniture are detected. To classify the detected objects, an ellipse is fitted to the objects. This eliminates the non-human objects and also for the humans detects the torso, eliminating the arms etc. The ellipse is iteratively shrunk till it is completely fitted inside the silhouette. To differentiate between humans and other object like carts, a heuristic based on the movement is used. In case of a human, the top part of the ellipse moves slowly with small fluctuations, which is not the case with other objects.

3.11 Li et al. [2004]

The authors describe the process of object-oriented scale-adaptive filtering (OOSAF) for finding objects of interest and apply it to the problem of detecting humans close to a camera, and to the problem of detecting multiple people in crowds. The OOSAF method uses the disparity map obtained from a stereo camera setup to estimate the scale at which filtering will be performed. For finding humans close to the camera, a histogram of the disparity is used as an input to OOSAF to select a scale, followed by filtering which isolates blobs. Bounding boxes are placed around the blobs followed by the application of a standard deformable template to prune the blobs. The method is applied in a similar manner to detect heads in a crowd scenario.

3.12 Han and Bhanu [2003]

In this paper, the authors propose to use infrared (IR) camera in conjunction with a standard camera for detecting humans. The cameras are mounted close to each other and observe the same scene from a similar viewpoint. Actually, the approach is not human specific, but will detect any moving object which has a thermal signature. First, background subtraction is performed independently in the color camera and the IR camera by using a gaussian probability distribution to model each background pixel. The detected foreground from the two cameras is registered using a hierarchical genetic algorithm, and the two registered silhouettes are then fused together into the final estimate.

3.13 Jiang et al. [2004]

This approach is based on fusion of infrared (IR) images with images from a regular camera. Humans display a characteristic signature in IR images due to their skin temperature, but these images typically have low contrast. They can be fused with images from a standard camera to obtain superior detection results. The proposed method first computes pixel saliencies in the two images (IR and visible) at multiple scales, and fusion is performed based on relative saliencies in the two images (called the perceptual contrast difference in the paper).

4 Direct detection

4.1 Cutler and Davis [2000]

The techniques in this paper focus on detecting periodic motions and is applicable to the detection of characteristic periodic biological motion patterns such as walking. The video from a moving camera is first stabilized and frame differencing and thresholding is performed to detect independently moving regions. Morphological operations are then used to obtain a set of tracked objects. Each segmented object is aligned along the time axis (to remove translation, and its size is also made constant across time. The object's temporal self-similarity matrix is computed by using similarity measures (such as correlation) which is periodic for periodic motions. Time-frequency analysis based on the short-time Fourier transform (STFT) is applied and autocorrelation is used for robust periodicity detection and analysis. A lattice fitting method is used to classify human, animal and vehicle, and the experiments demonstrate that the technique can distinguish the motion of a human from a dog. Not only is the system capable of detecting periodic human motion, but it also has knowledge of the period which is useful for extracting more information about gait such as stride length. The system performance is real-time.

4.2 Utsumi and Tetsutani [2002]

This paper uses the fact that the relative positions (geometric distances) of various body parts are common to all humans, although

the pixel values may vary because of the clothes or the illumination. The technique uses a structure known as the distance map which is built by taking an image of a human and breaking it into MxN blocks. A distance matrix of size MN x MN is then computed in which each element expresses the distance between color distributions present in a pair of blocks. Then, using such distance maps for a large database of human and non-human images, a statistical model is built for distance maps of each type, which consists of the average and covariance matrix for each block. The two distributions are compared using Mahalanobis distance and are found to be very similar except for a few elements. These few elements specify a data projection matrix which is the model used for recognition. Given a new input image, image patches at multiple locations and scales are compared to the model and a threshold is used to classify a patch as human or non-human.

4.3 Gavrila and Giebel [2002]

This paper deals with the challenging scenario of a moving camera mounted on a vehicle. Shape-based template matching is performed based on the Chamfer distance. A hierarchical tree of templates is constructed from a set of templates, which allows for efficient matching. This hierarchy is constructed automatically using partitional clustering, and each cluster is represented by a prototype. While matching, the process starts at the root and works its way towards the leaves to find a best matching template based on the chamfer distance. If the distance is greater than a set threshold for a given node, the search does not propagate to its child nodes. Thus, the matching is efficient. The authors also include a second verification state based on a neural network architecture which operates on rectangular patches detected by the previous template matching stage. The method also includes a Kalman filter based tracker for taking advantage of the temporal information for filling in missed detections. The paper reports results on a large testbed.

4.4 Viola et al. [2003]

This paper deals with the direct detection of humans from static images as well as video using a classifier trained on human shape

and motion features. The training dataset consists of images and videos of human and non-human examples. The paper restricts itself to the case of pedestrians (where humans are always in upright walking poses). The static detector uses images as inputs and efficiently extracts simple rectangular features using integral images. A cascade of classifiers is created to achieve superior detection and low false positives. Each stage of the classifier is trained on true and false positives from the previous stage using Adaboost to select weak classifiers (which are the simple rectangular features mentioned earlier). The dynamic detector is similarly trained using a combination of static and motion rectangular features. Both detectors are fast and provide good detection results on a large pedestrian dataset.

4.5 Sidenbladh [2004]

This paper focuses on human motion patterns for robust detection since they are relatively independent of appearance and environmental factors. The authors also observe that it is harder for a person to camouflage motion but easier to change appearance. The technique is based on collecting examples of human and non-human motion and computing optic flow. A support vector machine (SVM) with a radial basis function (RBF) kernel is trained on the optic flow patterns to create a human classifier. The resulting classifier can be applied to a new input video at multiple positions and scales, followed by pruning of detections with large overlap. The method is not suitable for detecting partially occluded humans.

4.6 Dalal and Triggs [2005]

The highlight of this paper is that it uses a histogram of gradients as the feature space for building a classifier. It uses the fact that the shape of a object can be well represented by a distribution of local intensity gradients or edge directions. This is done by dividing the image in small spatial parts (cells) and finding the histograms of edge orientations over all the pixels of the cell. The combined histogram entries form the feature representation after local contrast normalization in overlapping descriptor blocks. The authors experiment with several orientation and spatial binning resolutions and normalization

schemes to obtain the maximum performance. For classification, a dataset of human and non-human examples is created, and a linear classifier is trained using SVM on the gradient histogram features from the two classes. This classifier can then be applied to a new input image at several scales for detecting humans.

5 Summary

We have discussed several methods in the recent literature for human detection from video. We have organized them according to techniques which use background subtraction and those which operate directly on the input. In the first category, we have ordered the techniques based on type of background subtraction used and the model used to represent a human. In the second category, we have ordered the techniques based on the human model and classifier model used. Overall, there seems to be an increasing trend in the recent literature towards robust methods which operate directly on the image rather than those which require background subtraction as a first step.

Bibliography

C. Beleznai, B. Fruhstuck, and H. Bischof. Human detection in groups using a fast mean shift procedure. *International Conference on Image Processing*, 1:349–352, 2004.

R. Cutler and L. S. Davis. Robust real-time periodic motion detection, analysis, and applications. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 22(8):781–796, 2000.

Navneet Dalal and Bill Triggs. Histograms of oriented gradients for human detection. *IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pages 1063–6919, 2005.

H. Elzein, S. Lakshmanan, and P. Watta. A motion and shape-based pedestrian detection algorithm. *IEEE Intelligent Vehicles Symposium*, pages 500–504, 2003.

H. Eng, J. Wang, A. Kam, and W. Yau. A bayesian framework for robust human detection and occlusion handling using a human shape model. *International Conference on Pattern Recognition*, 2004.

D. M. Gavrila and J. Giebel. Shape-based pedestrian detection and tracking. *IEEE Intelligent Vehicle Symposium*, 1:8–14, 2002.

T. Haga, K. Sumi, and Y. Yagi. Human detection in outdoor scene using spatio-temporal motion analysis. *International Conference on Pattern Recognition*, 4:331–334, 2004.

Ju Han and B. Bhanu. Detecting moving humans using color and infrared video. *IEEE International Conference on Multisensor Fusion and Integration for Intelligent Systems*, 30:228–233, 2003.

Lijun Jiang, Feng Tian, Lim Ee Shen, Shiqian Wu, Susu Yao, Zhongkang Lu, and Lijun Xu. Perceptual-based fusion of ir and visual images for human detection. *International Symposium on Intelligent Multimedia, Video and Speech Processing*, pages 514–

517, 2004.

D. J. Lee, P. Zhan, A. Thomas, and R. Schoenberger. Shape-based human intrusion detection. *SPIE International Symposium on Defense and Security, Visual Information Processing XIII*, 5438: 81–91, 2004.

Liyuan Li, Shuzhi Sam Ge, T. Sim, Ying Ting Koh, and Xiaoyu Hunag. Object-oriented scale-adaptive filtering for human detection from stereo images. *IEEE Conference on Cybernetics and Intelligent Systems*, 1:135–140, 2004.

H. Sidenbladh. Detecting human motion with support vector machines. *Proceedings of the 17th International Conference on Pattern Recognition*, 2:188–191, 2004.

D. Toth and T. Aach. Detection and recognition of moving objects using statistical motion detection and fourier descriptors. *International Conference on Image Analysis and Processing*, pages 430–435, 2003.

Akira Utsumi and Nobuji Tetsutani. Human detection using geometrical pixel value structures. *Fifth IEEE International Conference on Automatic Face and Gesture Recognition*, page 39, 2002.

P. Viola, M. J. Jones, and D. Snow. Detecting pedestrians using patterns of motion and appearance. *IEEE International Conference on Computer Vision*, 2:734–741, 2003.

C. R. Wren, A. Azarbayejani, T. Darrell, and A. P. Pentland. Pfinder: real-time tracking of the human body. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 19(7):780–785, 1997.

Fengliang Xu and Kikuo Fujimura. Human detection using depth and gray images. *EEE Conference on Advanced Video and Signal Based Surveillance*, pages 115–121, 2003.

Sang Min Yoon and Hyunwoo Kim. Real-time multiple people detection using skin color, motion and appearance information. *In-*

ternational Workshop on Robot and Human Interactive Communication, pages 331–334, 2004.

Jianpeng Zhou and Jack Hoang. Real time robust human detection and tracking system. *IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, 3:149 – 149, 2005.