Appears in the *17th IAPR International Conference on Pattern Recognition*, vol 2, pp 188–191, Cambridge, UK 2004.

Detecting Human Motion with Support Vector Machines

Hedvig Sidenbladh
Department of Data and Information Fusion
Division of Command and Control Systems
Swedish Defence Research Agency
SE-172 90 Stockholm, Sweden
hedvig@foi.se

Abstract

This paper presents a method for detection of humans in video sequences. The intended application of the method is outdoor surveillance. In such an uncontrolled environment, the appearance of humans varies hugely due to clothing, identity, weather and amount and direction of light. The idea is therefore to detect patterns of human motion, which to a large extent is independent of the differences in appearance. To this end, a Support Vector Machine is trained with dense optical flow patterns originating from humans. The subjects are moving in different angles to the camera plane, on different image scales. This trained SVM is the core of a human detection algorithm which searches optical flow images for human-like motion patterns.

1. Introduction

Detection of humans in image sequences is an active research area within computer vision. In many applications, such as human-computer interaction, the detection is a basis for tracking. In other, such as surveillance systems or safety systems in cars, the detection in itself is used to trigger some type of alarm. The intended application of the method presented in this paper is outdoor surveillance.

Most human detection systems (e.g. [4, 6, 8, 9, 10, 12, 13]) detect humans or faces in a single image. These approaches are based on the assumption that humans can be localized in each individual frame, based on a model of human *appearance* (shape, contrast, color). However, in an uncontrolled outdoor environment such as the one considered in our application, human appearance varies due to environmental factors such as light conditions, clothing, contrast, and identity. The image sequences could even be taken during the night with a light enhancing camera. Furthermore, the subjects can be camouflaged or masked. All these factors cause large variation in the appearance of both the human and the scene, thereby obscuring the interesting features for human/non-human classification.

The approach in this paper is to detect human *motion*, which is a more discriminative cue for our problem than appearance (Figure 1). Even though the visual motion of a human varies with orientation towards the camera and limb configuration, it is much less dependent of the environmental factors described above [2, 7]. Furthermore, camouflage strives to alter the appearance of the subject, while it is much more difficult for a human to camouflage motion patterns. Thus, a detection approach which relies on models of human appearance is less efficient in our application.

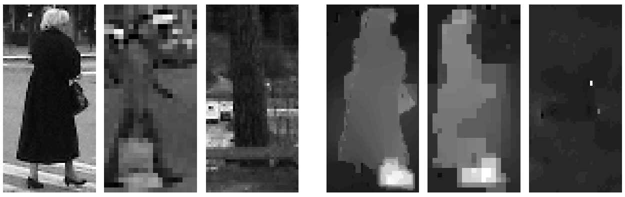

(a) Image
(b) Horizontal optical flow

Figure 1: Motion is characteristical for humans. (a) From left to right: two different video frames with humans in the same posture, and an video frame with a tree. The intra-human differences are large. At the same time, the tree might be mistaken for a person (upright rod-like structure) when viewed at a low resolution. (b) Detected horizontal flow (scaled plot: white - largest flow in right direction, black - largest flow in left direction) in the same frames. The flow depends on the movement itself and is largely uncorrelated to differences in appearance.

The motion cue used as input to the detection is robustly estimated dense optical flow [1] $\mathbf{u} = [u, v]$, where u is the horizontal flow and v the vertical flow between a pair of consecutive images in a sequence. A set of examples of human and non-human flow patterns is collected manually to serve as input to the training (Section 3).

The human flow pattern examples lie on a non-linearly shaped manifold in the high-dimensional space of flow patterns. Due to the high dimensionality of the state-space and the relatively low number of training samples, we use a Support Vector Machine (SVM) [3, 5] to learn the human/non-human classification from the examples (Section 3).

New flow patterns can now be compared to the trained

SVM and classified as human or non-human. A complete detection process involves a linear search over position and scale in each flow image. This is discussed in Section 4.

Preliminary experiments on video sequences of city scenes (Section 5) shows that the method is able to detect humans of different orientation and scale, and discriminate well between human and non-human motion.

2. Related Work

Most methods for human detection aim at detecting human appearance. Cues used are edges [8, 10], wavelet responses [12], color distributions [4], background subtraction [9] or a combination of multiple cues such as depth information, color and neural-net models of face patterns [6]. The detection is often used as an initialization step to tracking [4, 6, 9]. As stated in the introduction, the appearance cue is sometimes very weak in our application, leading us to instead use the image motion cue.

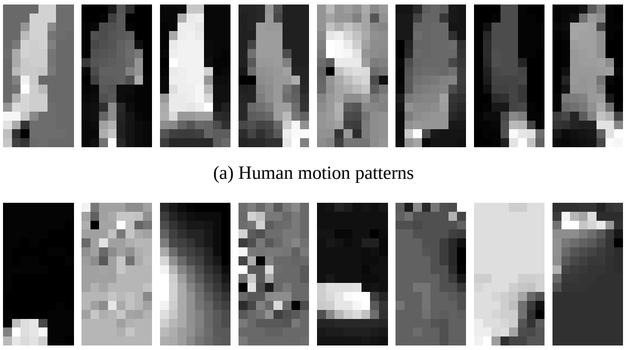
An approach based on motion is presented by Song et al. [14]. Here, feature points from two consecutive images in a sequence are compared to the corresponding points on a 2D kinematic model of a human. This approach does not entirely rely on motion information since there is an underlying assumption that one can find features corresponding to specific positions on the body. Viola et al. [15] use instead a filter-based approach to motion pattern recogntion. Using five filters for motion in different directions they are able to detect walking humans in low image resolution with a very low error rate. Optical flow is another representation of image motion. The model-based method of Fablet and Black [7] compares dense flow patterns with a generative model of human flow appearance. The method recovers both pose, orientation and position in the image but is computationally heavier than our pattern recognition approach.

Support Vector Machines (SVM) [3, 5] have previously proved efficient for face detection [13] and gender classification [11]. Pedestrian detection using SVM from appearance cues such as edge segments in the image [10] or wavelet responses [12] has also been reported. Our approach extends this work in that it detects human-like motion patterns instead of appearance patterns, making the detection more robust to difference in appearance due to environment, clothing and identity.

3. Training the SVM

The problem of learning a binary classifier can be expressed as that of learning the function $f : \Re^n \to \pm 1$ that maps patterns $\mathbf{x}$ onto their correct classification y as $y = f(\mathbf{x})$. In the case of an SVM, the function f takes the form [3, 5]

$$f(\mathbf{x}) = \sum_{i=1}^{N} y_i \alpha_i k(\mathbf{x}, \mathbf{x}_i) + b \;, \tag{1}$$


(a) Human motion patterns

(b) Non-human motion patterns

Figure 2: Examples from the training data, horizontal flow shown (scaled plot, see Figure 1). (a) Human motion patterns, different orientations from the camera. (b) Non-human motion patterns, mainly from foliage and cars.

where N is the number of training patterns, $(\mathbf{x}_i, y_i)$ is training pattern i with its classification, α_i and b are learned weights, and $k(.,.)$ is a kernel function. Here, we use a radial basis function $k(\mathbf{x}, \mathbf{x}_i) = e^{-\|\mathbf{x}-\mathbf{x}_i\|/2\sigma^2}$. The patterns for which $\alpha_i > 0$ are denoted *support vectors*.

The surface $f(\mathbf{x}) = 0$ defines a hyperplane through the feature space as defined by the kernel $k(.,.)$. The weights α_i and b are selected so that the number of incorrect classifications in the training set is minimized, while the distances from this hyperplane to the support vectors are maximized. This is achieved by solving the optimization problem [3, 5]

$$\mathit{Maximize}:$$

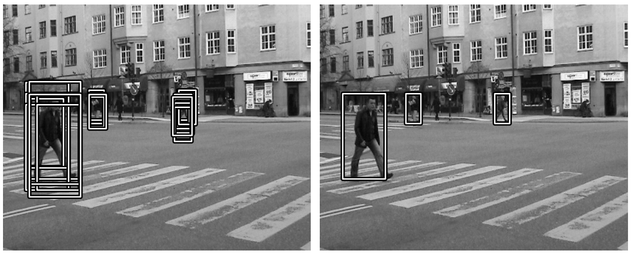
$$L_D \equiv \sum_{i=1}^{N} \alpha_i - \frac{1}{2} \sum_{i=1}^{N} \sum_{j=1}^{N} y_i y_j \alpha_i \alpha_j k(\mathbf{x}_i, \mathbf{x}_j) \tag{2}$$

$$\mathit{subject\ to}:$$

$$0 \leq \alpha_i \leq C \;, \quad \sum_{i=1}^{N} y_i \alpha_i = 0 \;. \tag{3}$$

The constant C affects the tolerance to incorrect classifications. Using the optimal parameters α_i, Eq (1) with any support vector (x_i, y_i) as indata can be used to find b. For a thorough description of the training process, see [3, 5].

Patterns. The training set consists of 443 human flow patterns and 11688 non-human flow patterns (Figure 2). A flow pattern is here defined as a vector $\mathbf{x} = [\mathbf{u}_1, \mathbf{u}_2, \ldots, \mathbf{u}_{mn}]$ where $\mathbf{u}_k = [u_k, v_k]$ is the flow in the k:th pixel in the rectangular pattern of size $m \times n$. Here, $m = 16$ and $n = 8$, which means that $\dim(\mathbf{x}) = 256$. The human flow patterns used for training are collected manually from dense flow images. These are computed [1] from pairs of consecutive images in video sequences with a large number of individuals in different types of environment. Non-human patterns are collected automatically from similar sequences without



(a) All detections

(b) Multiple candidates removed

Figure 3: Removing multiple detections of the same subject. (a) Original set of detections. (b) Set of detections after removal of multiple candidates (see text).

humans. Each example pattern $\mathbf{x}_i$ is resampled to a size of 16×8 pixels and assigned a label $y_i = 1$ if the pattern is human, otherwise $y_i = -1$.

Iterative training. The learning procedure has a computational complexity of $\mathcal{O}(N^2)$ where N is the number of training patterns, which means that the computations become infeasible for our large number of patterns. To enable efficient learning, the iterative strategy of Osuna et al. [13] is therefore employed.

After training, 247 human patterns and 499 non-human patterns were used as support vectors.[1]

4. Detection of Humans

A new pattern $\mathbf{x}$ can now be classified as human or non-human using the learned function $y = f(\mathbf{x})$ (Eq (1)).

The input to the detection is a flow image, obtained using the same robust flow algorithm [1] as for the training images. A linear search over positions and pattern heights is performed.[2] For each position and height, the corresponding image window is extracted and normalized to a size of 16×8 pixels. The resulting pattern $\mathbf{x}$ is then classified using the SVM. Positive answers are returned as detections.

Removing multiple detections. Figure 3(a) shows the result of a detection. Typically, several human pattern candidates corresponding to the same individual are found. To give a more accurate estimate of the number of people in the scene, hits overlapping more than 50% will be replaced by a single window whose position and height is the spatial weighted mean of the positions and heights of the overlapping windows. Figure 3(b) shows the result of such a pruning procedure.

5. Experimental Results

The detection algorithm was tested on images from a city scene with cars and people of different scales, moving in different directions. Images in the sequence were 360×288 pixels large and were captured in 25 Hz.

Figure 4 shows three frames from the same sequence. Three of the four persons moving in the scene is detected accurately, while the fourth (a person in black) is ignored. The reason for this could be that the flow response on that person is weak due to foreground-background similarity (Figure 4(b,d,f)). The vulnerability to flow estimation errors is a weakness of the method.

Figure 5 shows another sequence in which a car is present. Even though there are some human-like motion patterns in the car flow (Figure 5(b)), the method distinguishes correctly between human and non-human patterns. In both frame 0 and 10 (Figure 5(a,e)), humans are partly occluded by the car. Since the method can not presently detect partly occluded motion patterns, the detector ignores the occluded humans. This is also an important issue.

6. Conclusions

An SVM-based detector of human-like optical flow patterns was presented. To detect humans in a scene, dense optical flow was first computed from a pair of consecutive images in a video shot of the scene. Windows of different size and position in the image was then tested against an SVM which was previously trained with a large number of examples of human and non-human flow patterns. The method was tested on a number of images from a city scene.

Future work. The human training set is quite small, due to the manual pattern acquisition. To obtain a larger training set, more examples of human motion patterns could be collected by reinforcement learning, i.e. iterative user guided incorporation of detection results in the training set [12].

Furthermore, the robustness of the method to flaws in the computed optical flow should be investigated.

Another future direction of research is detection of partially occluded patterns. One option is to detect the motion of human body parts separately. Knowledge about spatial relations between the detected body parts are then used to reinforce or suppress the detections.

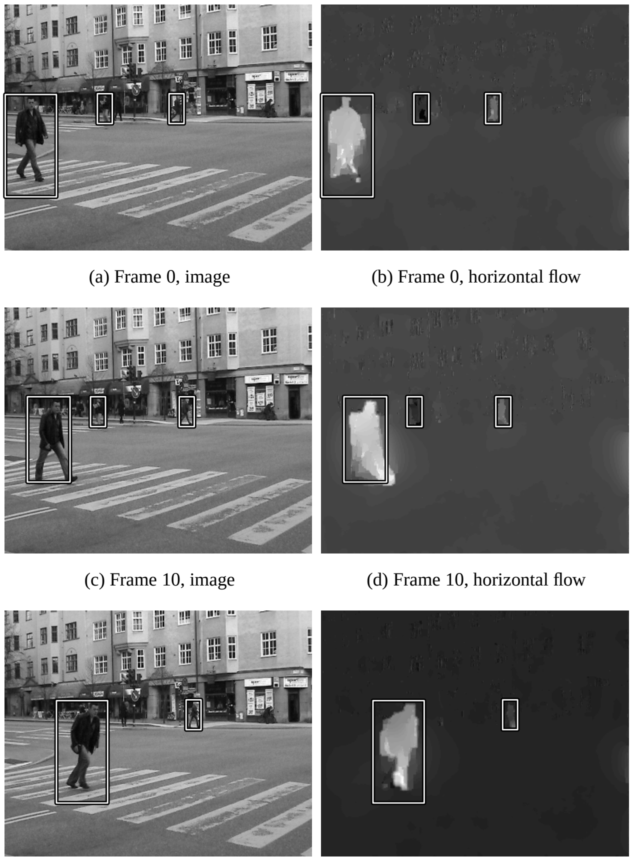
References

[1] M. J. Black and P. Anandan. The robust estimation of multiple motions: Parametric and piecewise-smooth flow fields. *CVIU*, 63(1):75–104, 1996.

[2] A. Bobick and J. Davis. The representation and recognition of action using temporal templates. *PAMI*, 23(3):257–267, 2001.

[1] The high proportion, 56%, of human patterns used as support vectors indicates that more human examples would be beneficial to the performance of the SVM. This is further discussed in Conclusions.

[2] The computational efficiency of the detection can easily be enhanced by introducing a threshold on the absolute amount of flow in the candidate pattern before the classification step begins.



(a) Frame 0, image

(b) Frame 0, horizontal flow

(c) Frame 10, image

(d) Frame 10, horizontal flow

(e) Frame 20, image

(f) Frame 20, horizontal flow

Figure 4: Four persons crossing the street. The person in black is ignored by the detector. (a,c,e) Frame 0, 10 and 20 in the sequence with detected humans marked. (b,d,f) The corresponding detected horizontal flow (scaled plot, see Figure 1).

[3] C. J. C. Burges. A tutorial on support vector machines for pattern recognition. *Data Mining and Knowledge Discovery*, 2(2):121–167, 1998.

[4] D. Comaniciu and V. Ramesh. Robust detection and tracking of human faces with an active camera. In *IEEE International Workshop on Visual Surveillance*, 2000.

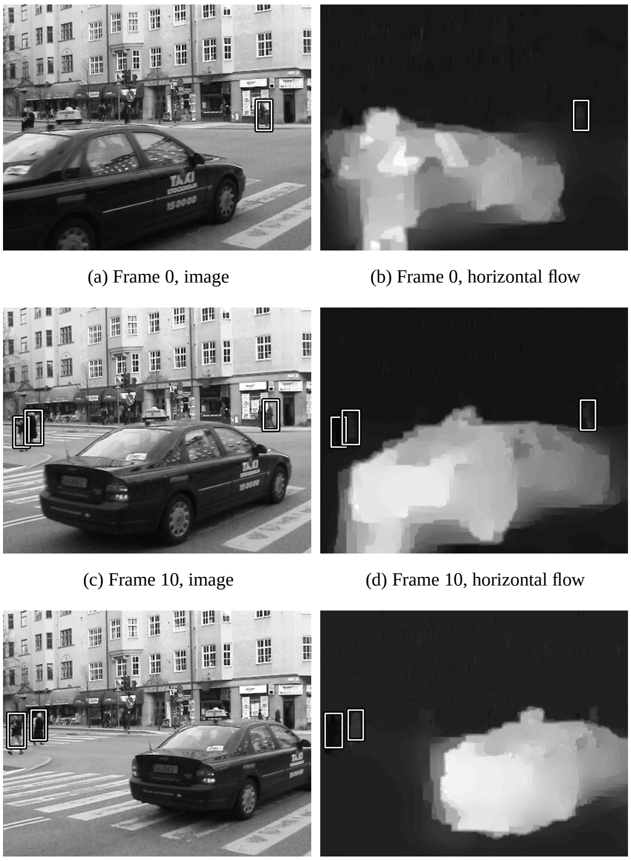
[5] N. Cristianini and J. Shawe-Taylor. *An Introduction to Support Vector Machines*. Cambridge University Press, Cambridge, UK, 2000.

[6] T. Darrell, G. Gordon, M. Harwille, and J. Woodfill. Integrated person tracking using stereo, color, and pattern recognition. In *CVPR*, pages 601–609, 1998.

[7] R. Fablet and M. J. Black. Automatic detection and tracking of human motion with a view-based representation. In *ECCV*, volume 1, pages 476–491, 2002.

[8] D. M. Gavrila. Pedestrian detection from a moving vehicle. In *ECCV*, volume 2, pages 37–49, 2000.

[9] I. Haritaoglu, D. Harwood, and L. Davis. W4: Real-time surveillance of people and their activities. *PAMI*, 22(8):809–830, 2000.



(a) Frame 0, image

(b) Frame 0, horizontal flow

(c) Frame 10, image

(d) Frame 10, horizontal flow

(e) Frame 20, image

(f) Frame 20, horizontal flow

Figure 5: Car and three persons. The detector separates well between human and non-human motion patterns. However, partly occluded people are not detected. (a,c,e) Frame 0, 10 and 20 in the sequence with detected humans marked. (b,d,f) The corresponding detected horizontal flow (scaled plot, see Figure 1).

[10] S. Kang, H. Byun, and S-W. Lee. Real-time pedestrian detection using support vector machines. *International Journal of Pattern Recognition and Artificial Intelligence*, 17(3):405–416, 2003.

[11] B. Moghaddam and M-H. Yang. Sex with support vector machines. In *Advances in Neural Information Processing Systems 13*, pages 960–966, 2001.

[12] M. Oren, C. Papageorgiou, P. Sinha, E. Osuna, and T. Poggio. Pedestrian detection using wavelet templates. In *CVPR*, pages 193–199, 1997.

[13] E. Osuna, R. Freund, and F. Girosi. Training support vector machines: an application to face detection. In *CVPR*, pages 130–136, 1997.

[14] Y. Song, X. Feng, and P. Perona. Towards detection of human motion. In *CVPR*, volume 1, pages 810–817, 2000.

[15] P. Viola, M. J. Jones, and D. Snow. Detecting pedestrians using patterns of motion and appearance. In *ICCV*, pages 734–741, 2003.