TOWARDS DETECTING PEOPLE CARRYING OBJECTS
A Periodicity Dependency Pattern Approach

Tobias Senst, Rubén Heras Evangelio, Volker Eiselein, Michael Pätzold and Thomas Sikora
Communication Systems Group, Technische Universität Berlin, Sekr. EN-1, Einsteinufer 17, 10587 Berlin, Germany

Keywords: Gait analysis, Periodicity analysis, People carrying objects, Scene interpretation, Pattern recognition.

Abstract: Detecting people carrying objects is a commonly formulated problem which results can be used as a first step in order to monitor interactions between people and objects in computer vision applications. In this paper we propose a novel method for this task. By using gray-value information instead of the contours obtained by a segmentation process we build up a system that is robust against segmentation errors. Experimental results show the validity of the method.

1 INTRODUCTION

The number of video surveillance cameras is increasing notably. This leads to a growing interest in algorithms for automatically analyzing the huge amount of video information generated by these devices. The development of such algorithms is being further boosted by the increasing processing power that modern CPU architectures offer.

Detecting people carrying objects is a commonly formulated problem. Results can be used as a first step in order to monitor interactions between people and objects, like depositing or removing an object. Most of the current approaches aiming to detect people carrying objects are based on contours. One of the first methods was Backpack, proposed in (Haritaoglu et al., 1999; Haritaoglu et al., 2000). Backpack is a two step algorithm. At first, the algorithm detects the persons in a frame and analyses the symmetry of their silhouette assuming that the silhouette of a person is symmetric. Non-symmetric parts of the person are labeled as potential carried baggage. After that, they analyse the frequency of the parts labeled in the first step to discard those of them showing a periodicity, which are assumed to be the arms and legs.

In (Damen and Hogg, 2008) the authors analyse the silhouette of a person too, but, instead of using the non-symmetric parts of it, they match the silhouette to a 3D model of a person. In order to cope with different camera positions, scale, translation and rotation, the model must be adjusted to the detected persons. Carried baggage is then detected by the salients obtained with the best model.

In (Abdelkader and Davis, 2002) the authors introduce a body model inspired by the human physiognomy and use simple constraints for the detection of carried objects. They partition the detected persons in four blocks and calculate the periodicity and the amplitude of the blocks over time. Those persons whose features do not fit the properties observed for the gait of persons walking without baggage are classified as persons carrying an object. Gait analysis has already been formulated in (Ekinci and Aykut, 2007) and (Abdelkader et al., 2002).

All these algorithms rely on a precise object segmentation which is difficult to achieve in video surveillance sequences. To avoid this prerequisite, we propose a novel method based on periodicity analysis of those regions containing a person. Therefore, we use grayvalue information instead of segmentation masks. This makes our algorithm more robust against failures in the segmentation step.

2 A PERIODICITY DEPENDENCY PATTERN APPROACH

The proposed method consists of three steps. First we compute an aligned spatio-temporal bounding box for each tracked person to cope with the noisy results of person detection and tracking. The alignment is based on a foreground segmentation mask.

In the second step, the bounding box is partitioned into a set of regions. Based on the self-similarities of these regions over time, we define a periodicity de-

Senst T., Heras Evangelio R., Eiselein V., Pätzold M. and Sikora T. (2010).
TOWARDS DETECTING PEOPLE CARRYING OBJECTS - A Periodicity Dependency Pattern Approach.
In *Proceedings of the International Conference on Computer Vision Theory and Applications*, pages 524-529
DOI: 10.5220/0002845505240529

pendency (PD) descriptor. The aim is to obtain a pattern describing the spatial dependency of human motion, e.g. synchronous arm and leg motion.

In the third step, the analysed person is classified as carrying an object or not based on a set of previously learned PD patterns.

The algorithm is shown schematically in Figure 1.

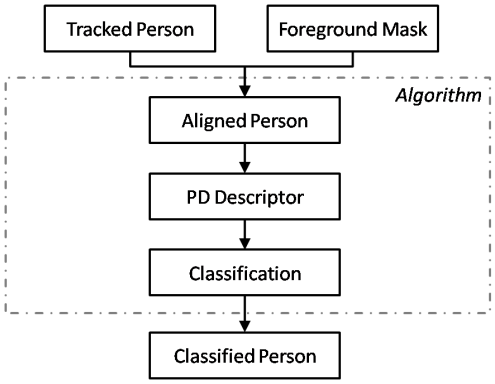

Figure 1: Schema of the proposed method. Starting with a detected and tracked person, we compute a spatio-temporal volume which is the base data for our descriptor. The descriptor is classified using a support vector machine trained with annotated sample sequences.

2.1 Spatio-temporal Alignment of the Bounding Box

The proposed method starts by computing an aligned spatio-temporal bounding box for every tracked person. This preprocessing step is needed because a person must always be located in approximately the same place within the bounding box. For this purpose we take trajectories and foreground masks as input for the system.

The trajectories used in this step were extracted manually with the viPER-tool[1]. They could also be obtained by means of a tracking algorithm, e.g. a Kalman filter-based approach as described in (Pathan et al., 2009) but this is beyond the scope of this paper.

The foreground masks used to align the bounding boxes containing the persons were obtained by background subtraction with a gaussian mixture background model as defined in (Stauffer and Grimson, 1999) combined with a shadow removal technique as proposed by (Horprasert et al., 2000).

The aim is to fit the bounding box with centre at (x_{bb}, y_{bb}) and size (w_{bb}, h_{bb}) to the foreground mask. Therefore we recompute the size of the bounding box $(w_{bb}^{new}, h_{bb}^{new})$ as the smallest bounding box fitting the mask of the person. Since we are using short video sequences, we can temporally smooth them as shown in the following equations

$$w_{bb}^{t} = \begin{cases} \frac{1}{t+1} \cdot w_{bb}^{new} + \frac{t}{t+1} \cdot w_{bb}^{t-1}, & \text{if } t < T \\ \frac{1}{T+1} \cdot w_{bb}^{new} + \frac{T}{T+1} \cdot w_{bb}^{t-1}, & \text{otherwise} \end{cases} \quad (1)$$

$$h_{bb}^{t} = \begin{cases} \frac{1}{t+1} \cdot h_{bb}^{new} + \frac{t}{t+1} \cdot h_{bb}^{t-1}, & \text{if } t < T \\ \frac{1}{T+1} \cdot h_{bb}^{new} + \frac{T}{T+1} \cdot h_{bb}^{t-1}, & \text{otherwise} \end{cases} \quad (2)$$

where T is a predefined value determining the learning rate of the system. During the initialization phase, the weight for the samples of the size of the bounding box is rated differently not to bias the resulting size with the first sample. After the number of samples is greater than T, the learning rate is set to the constant value $(T+1)^{(}-1)$. The selection of T can be made dependent of the trajectories that the persons follow in the video to account for the scene geometry. A smaller value of T can be chosen for the alignment of persons walking towards the camera or away from it so that the size of their bounding box can change quickly. In our experiments we took a constant value of $T = 100$ since the people appearing in the videos are mostly walking parallel to the camera.

To center the bounding box we define a sector overlapping with the upper part of the body. Then we compute the centre of the foreground contained in this sector of the bounding box. The idea is to center the bounding box in the more stable parts of the body (i.e. shoulder and neck). Finally, we normalize the bounding box so that the posterior frequency analysis is independent of the size of a person.

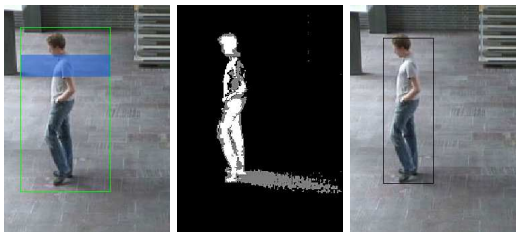

Figure 2: Left: alignment of the bounding box with the sector overlapping with the upper body part (blue). Middle: foreground mask with detected shadow regions (gray). Right: aligned bounding box.

Experiments with our dataset have shown that shadow removal produces partially sparse foreground masks so that the center alignment can be unstable. To overcome this problem we first take the foreground mask without shadow removal for the calculation of the centre (x_{bb}, y_{bb}) and then we remove the shadows to calculate the size of the box. Since the bounding box is defined to be the maximum width and height

[1] http://sourceforge.net/projects/viper-toolkit

of the foreground mask, it is not critical if some foreground points in the body of the person are removed during the shadow removal process.

2.2 Periodicity Dependency Descriptor

Human gait is structured in space and time, due to the symmetry of the human body. In (Webb et al., 1994) model a human as a set of two pendula oscillating with a phase delay of a half period.

Based on the observation that the gait of a person carrying an object appears differently than someone who does not carry an object, we analyse and learn the motion information of people to detect those of them carrying objects.

The movement of different body parts is highly depending on each other since the human body forms a kinematic chain. To model this dependency we partition each bounding box into N blocks B_n with $n = 0 \ldots N-1$ and analyse their periodicity. Therefor we use the similarity plot between the images I_{t_1} and I_{t_2} as defined by (Haritaoglu et al., 1999) for each block in an image stack of the size Δt:

$$S_n(t_1,t_2) = \min_{d\mathbf{p}\in\Omega} \sum_{B_n} |I_{t_1}(\mathbf{p}+d\mathbf{p}) - I_{t_2}(\mathbf{p})|, \quad (3)$$

where I_{t_1} and $I_{t_2} \in B_n$, $\mathbf{p} = (x,y)^T$, Ω defines the search window and n is the block number.

As a result, a spatial map of self-similarities between the blocks within the bounding box is obtained from which the periodicity is calculated. The similarity plot of those blocks containing body parts exhibitting a cyclic motion is a periodic signal while the signal generated by those blocks containing static body parts is quasi-linear. Indeed we observed that those blocks containing carried objects oft generate a cyclic motion but with minor amplitude than those blocks containing the body extremities, e.g. a person carrying a hand bag. That is why we use a search window defined by Ω to compute the similarity plot of a given block. On the other hand this makes the algorithm more robust against alignment errors. In a postprocessing step we smooth the similarity plot to remove the high frequencies from the information signal.

People have different gaits depending for instance on their walking speed. This makes it necessary to use a relative measurement. Furthermore we need a measurement that is not affected by inversion or translation of the considered signals. Therefore we define the periodicity dependency (PD) of two blocks as the maximum of the absolute relative correlation of the similarity plot. The PD of a block n to another block m is:

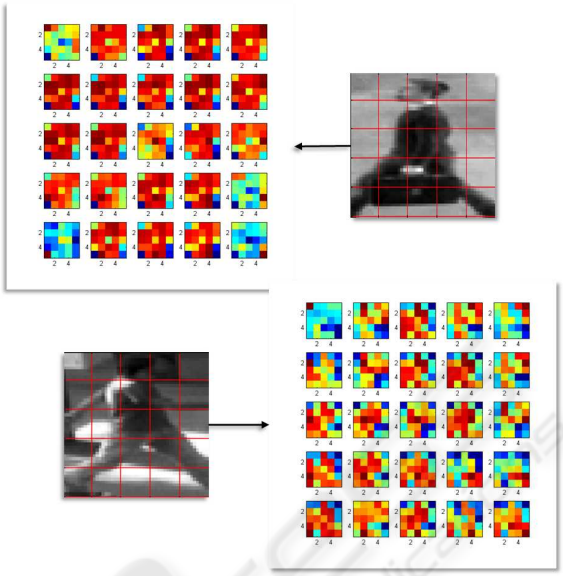

Figure 3: Top: color-coded Periodicity Dependency (PD) map for a walking person with a briefcase (red=1, blue=0). Bottom: color-coded PD map for a walking person without baggage. Both examples were generated using 5x5 blocks.

$$PD_n(m) = max \left| \frac{\sum_{\Delta t} (S_n - \overline{S}_n) \cdot (S_m - \overline{S}_m)}{\sqrt{\sum_{\Delta t} (S_n - \overline{S}_n)^2 \cdot \sum_{\Delta t} (S_m - \overline{S}_m)^2}} \right| \quad (4)$$

The result of computing the PD over all blocks is a spatial PD map as shown in Figure 3. To avoid redundancy we save the values of the PD map in a PD descriptor (PDD) defined as follows:

$$PDD_t = \begin{bmatrix} PD_1(2) \\ \vdots \\ PD_1(N) \\ PD_2(3) \\ \vdots \\ PD_2(N) \\ \vdots \\ PD_{N-1}(N) \end{bmatrix} \quad (5)$$

Figure 3 shows an example of a PD map of a person carrying a briefcase and a person not carrying anything. The PD pattern of the lower blocks of a person carrying an object is significantly different to the one of a person not carrying anything. The lower left and right corner blocks, which correspond to the position of feet and legs, correlate much less with all other blocks for those persons carrying a piece of baggage. The reason is that the carried object occludes the legs and their movement is not detected by the similarity plot. The correlation between other blocks changes in

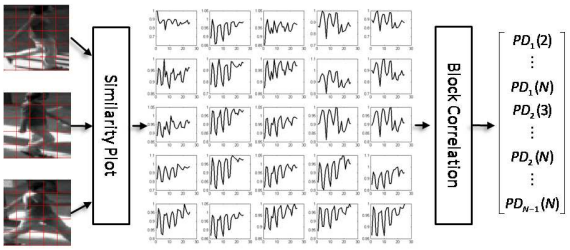

Figure 4: Periodicity Dependency Descriptor (PDD) schema for a person walking without baggage and 5x5 blocks.

a similar fashion. Figure 4 shows schematically the computation of a PDD.

2.3 Classification using PDD

The proposed human model described by the periodicity dependency descriptor (PDD) allows to model the correlation of the movement in different blocks of a tracked person. For the classification we use a support vector machine (SVM). For each frame we assign a binary label indicating the probability that a person is carrying an object or not. These labels are used to classify a tracked person over a short video sequence. The classification is done as a voting system which classifies the whole video sequence in favor of the majority of the framewise assigned labels. Mathematically formulated, this yields

$$L_{hasBaggage} = \begin{cases} 1, & \text{if } N_{pos} > N_{neg} \\ 0, & \text{otherwise} \end{cases} \quad (6)$$

with N_{pos} being the number of positively classified samples, N_{neg} the number of negatively classified samples and $L_{hasBaggage}$ the decision for a whole sequence.

We assumed here that people do not leave their baggage while walking. This assumption can be fulfilled by choosing short tracks of the persons.

3 EXPERIMENTS AND RESULTS

The method was tested with 31 annotated indoor sequences with 2328 frames taken from a fixed camera viewpoint at 25 fps and an image size of 720x432. There were 17 scenes with people carrying various objects as briefcases, trolley bags and unusual objects as e.g. a tripod case and 14 with persons moving without baggage. Throughout the sequences the illumination changes heavily yielding hard shadows and a differently illuminated background. Therefore a good segmentation of the scene is hard to achieve.

Figure 5: The training dataset contains 6 scenes, 3 of them with people carrying objects. The overall number of frames in these sequences is 571.

Firstly, the support vector machine (SVM) must be trained to recognize PD patterns of people carrying objects. In order to improve the results, a comparison of different descriptor parameters was performed. The SVM was trained both with a linear kernel and a radial basis function (rbf) kernel. The results can be seen in Table 1. For the training of the SVM the dataset was split into a training set of 6 sequences (571 sample images) shown in Figure 5 and a recall dataset of 25 sequences including 1757 frames. This is an approximate ratio of 1:3 training data to recall data. The annotated persons were normalised to a size of 50x50 pixels.

A smaller number of blocks generates worse results, because different body parts can be included in a block and therefore the motion information is blurred. Effectively, in this case local motions are superimposed in one block.

A finer block grid can decrease the robustness against alignment errors. The probability that a local motion signal is perceivable in several blocks increases with the number of blocks. As a result, the intra-class variation of the data increases and the classification can be hindered.

Comparing different values for the value of Ω, it can be seen that a bigger search window enhances the robustness of the system because alignment errors within the bounding box can be compensated. Nonetheless, computing the similarity map is a time-demanding step and it is further slowed down by a greater Ω. The parameters for the evaluation are therefore chosen as a 5x5 block grid with a search window of 10 pixels. For this configuration a linear kernel for the SVM provided better results than a rbf-kernel.

The results produced by the proposed method are summarized in Table 2.

Figure 6: Width series and their corresponding classification functions. A-E: correctly classified and F: wrongly classified.

Table 1: Results of the classification experiments by testing different PDD settings and SVM kernels. Best results were achieved with 5x5 blocks, a search window of 10 and a linear svm kernel.

	Precision	Recall	Accuracy
3x3 linear Ω 5	0.575	0.993	0.583
3x3 rbf Ω 5	0.562	1.000	0.562
3x3 linear Ω 10	0.644	0.913	0.667
3x3 rbf Ω 10	0.615	0.941	0.636
5x5 linear Ω 5	0.527	0.466	0.465
5x5 rbf Ω 5	0.570	0.634	0.526
5x5 linear Ω 10	**0.816**	**0.658**	**0.725**
5x5 rbf Ω 10	0.659	0.589	0.598
7x7 linear Ω 5	0.563	0.550	0.507
7x7 rbf Ω 5	0.534	0.640	0.485

Figure 6 illustrates the foreground segmentation and detection results for some sequences[2]. The severe illumination changes and bad contrast hinder the foreground segmentation as can be seen in Figure 6C. The proposed method is quite robust to that difficulty since, while the contour of the person is disturbed, we use the periodicity observed in the whole bounding box.

We experimented also with persons viewed from different angles. Moving away and towards the camera has another pattern as moving side-ways. Thus the method produces less robust results since the SVM was trained with people moving parallel to the camera (Figure 6D).

Figure 6F shows a wrong classification result. In this sequence the person is swinging the handbag in the same frequency and amplitude as its leg so the resulting pattern is the same as a pattern learned with no baggage. Therefore an incorrect decision was taken.

Table 2: Detection carried objects result on 25 sequences.

	no Baggage	Baggage
total	11	14
false detected	0	2
correctly detected	11	12

4 CONCLUSIONS

It this paper we proposed a novel method for detecting people carrying objects. Although the results obtained in our first experiments are very promising, the system can be further improved in many directions. A more elaborated background model can allow the achievement of more precise bounding boxes. Semantic information could be used to determine the start and endpoint for the trajectories for each bounding box. Furthermore, the computation load of the similarity plot can be reduced by using other measures based for instance on PCA.

[2]available at http://www.nue.tu-berlin.de/menue/ mitarbeiter/tobias_senst/

REFERENCES

Abdelkader, C. B. and Davis, L. (2002). Detection of people carrying objects: A motion-based recognition approach. *Automatic Face and Gesture Recognition, IEEE International Conference on*, pages 378–383.

Abdelkader, C. B., Davis, L., and Cutler, R. (2002). Motion-based recognition of people in eigengait space. *Automatic Face and Gesture Recognition, IEEE International Conference on*, pages 267–272.

Damen, D. and Hogg, D. (2008). Detecting carried objects in short video sequences. In *ECCV '08: Proceedings of the 10th European Conference on Computer Vision*, pages 154–167, Berlin, Heidelberg. Springer-Verlag.

Ekinci, M. and Aykut, M. (2007). Human gait recognition based on kernel pca using projections. *J. Comput. Sci. Technol.*, 22(6):867–876.

Haritaoglu, I., Cutler, R., Harwood, D., and Davis, L. S. (1999). Backpack: Detection of people carrying objects using silhouettes. In *Computer Vision and Image Understanding*, pages 102–107.

Haritaoglu, I., Harwood, D., and Davis, L. S. (2000). W4: Real-time surveillance of people and their activities. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 22(8):809–830.

Horprasert, T., Harwood, D., and Davis, L. S. (2000). A robust background subtraction and shadow detection. In *In Proceedings of the Asian Conference on Computer Vision*.

Pathan, S., Al-Hamadi, A., Senst, T., and Michaelis, B. (2009). Multi-object tracking using semantic analysis and kalman filter. In *ISPA '09: Proceedings of 6th International Symposium on Image and Signal Processing and Analysis*, pages 271–276.

Stauffer, C. and Grimson, W. E. L. (1999). Adaptive background mixture models for real-time tracking. In *CVPR*, pages 2246–2252.

Webb, D., Tuttle, R. H., and Baksh, M. (1994). Pendular activity of human upper limbs during slow and normal walking. *American Journal of Physical Anthropology*, 93:477–489.