Korc F., Hlavac V.: Detection and Tracking of Humans in Single View Sequences Using 2D Articulated Model, in the book edited by Rosenhahn, B. and Klette, R. and Metaxas, D.: Human Motion, Understanding, Modelling, Capture and Animation, Springer Verlag, Heidelberg, Germany, Series Computational Imaging and Vision, Vol. 36, 2007, p. 105-130, ISBN 978-1-4020-6692-4.

Detection and Tracking of Humans in Single View Sequences Using 2D Articulated Model

Filip Korč[1] and Václav Hlaváč[2]

[1] University of Bonn, Department of Photogrammetry
Nussallee 15, Bonn, 53115, Germany

[2] Czech Technical University in Prague, Center for Machine Perception
Karlovo naměstí 13, Prague 2, 121 35, Czech Republic

Abstract. This work contributes to detection and tracking of walking or running humans in surveillance video sequences. We propose a 2D model-based approach to the whole body tracking in a video sequence captured from a single camera view. An extended six-link biped human model is employed. We assume that a static camera observes the scene horizontally or obliquely. Persons can be seen from a continuum of views ranging from a lateral to a frontal one. We do not expect humans to be the only moving objects in the scene and to appear at the same scale at different image locations.

1 Introduction and Problem Formulation

Detection of walking or running humans in surveillance video sequences and tracking them is an appealing computer vision problem. The problem has been approached by many researchers. A satisfactory solution, however, has not been presented yet.

This work suggests a model-based approach to tracking in a video sequence captured from a single camera view. Matching of a human model to video stream is performed in two dimensions (2D) on the image frame level. 2D tracking aims at following the image projections of humans. The 3D displacement of a human is perspectively projected into a planar displacement that can be modeled as a 2D transformation. An adaptive model is required to handle appearance changes due to perspective effects or due to the change of the body parts position relative to each other. Two different postures in 3D may result in identical 2D projection. This ambiguity in a 2D model matching approach makes the tracking task rather difficult. The 2D-3D pose estimation problem is further discussed in Klette **??**.

It has been previously argued by Gavrila [1] that the choice of a solution is, to a great extent, application-driven. Our goal is to develop a method which could track humans in surveillance sequences, i.e. the main focus of our work is in tracking humans when precise pose recovery is not critical. We assume that a single static camera observes the scene horizontally or obliquely. We do not consider a top view. Model-based approach could in longer perspective help to overcome the assumption of a static observing camera. The walking/running

persons can be seen from a continuum of views ranging from a lateral to a frontal one. We do not expect humans to be the only moving objects in the scene and to appear at the same scale at different image locations. The method copes with a slowly changing background. It is assumed that the whole human body can be seen in the sequence. A short occlusion of a tracked human by other objects in the scene is admissible too.

2 Our Approach Informally

Our approach is based on a similar philosophy as the seminal paper by Hogg [2]. We also employ an articulated 2D model of a human consisting of several rectangles modeling a simplified 2D view of a human body.

The biped model of a human is composed of six rectangles which are fit to corresponding parts of a human body: head, torso, left/right thigh and left/right calf, see Figure 1a. Our extended model also takes into account data in the model neighborhood.

Our detection and tracking algorithm is designed as a loop consisting of three steps: (a) detecting human candidates, (b) validating model of a human, (c) tracking of the model in consequent frames.

Input of our algorithm is the outcome of pixel-based motion detector (background subtraction). When detecting human candidates in step (a), a coarse model is used to direct attention to places in the motion image where humans could be. The outcome is a region of interest (ROI). Detector of ROIs has to find almost no false negatives.

In the validation step (b) a slightly more sophisticated extended biped human model, recall Figure 1a, is employed within ROI to verify the appearance of a human.

The model tracking process (c) is launched after successful model initialization. If the model fails to explain image data satisfactorily in the tracking process then the algorithm is stopped and restarted in the next frame.

The coarse detection of ROI could be viewed as employing a simple model allowing the search of human candidates in a computationally efficient manner. After the ROI is found, a model with a more refined structure better explaining the observation could take over.

Such methodology opens a future way to employing models with a varying refinement of model structure appropriate to available image data resolution at which a human in the scene is present. This would lead to tracking a person at a proper level of detail, attempting to track body parts only when possible, for example. Another potential is in adapting computational demand to changing degree of certainty in the tracked human. This work reports results in which only ROI detection and the extended biped human model are used.

3 Related Work

There has been a number of works adopting a model-based approach and an articulated biped model. A survey of publications in human motion analysis can be found in [1]. This survey identifies a number of applications and provides an overview of developments in the domain of visual analysis of human movement preceding the year 1999. Many of the recent works, however, were devoted to more precise body pose recovery and subsequent action recognition in high resolution videos. We feel that there is a number of issues as self-occlusion which need to be further addressed to merely achieve better robustness in tracking of articulated objects in low resolution surveillance videos.

One of the early monocular model-based approaches to human detection and tracking may be found in Rohr [3]. They matched gray value edge lines to volume model contours and employed Kalman filter to estimate the model parameters in the consecutive frames. A person moving parallel to the image plane was assumed to reduce the complexity of the recognition task.

As far as the model is concerned, our work is close to Ju et al. [4]. They adopted a method for tracking articulated motion of human limbs in a sequence using parameterized models of optical flow. They made the assumption that a person can be represented by a set of connected planar patches: "cardboard person model". Constraints introduced between the patches enforce articulated motion. Their approach is limited to constant viewpoint.

Our model is also similar to Zhang et al. [5]. They apply a 2D five-link biped model to the problem of gait recognition using human body movements. Their work is constrained to a human body observed laterally. We took over their model and extended it significantly.

The model posture evaluation and the idea of a rough calibration of a scene to estimate the size of the model in Beleznai et al. [6] was an inspiration for us. Their method adopts a simple human model described by three rectangles to detect individual humans within groups and to verify their hypothesized configuration. The approach is capable of real time operation, and handles multiple humans and their occlusions Beleznai et al. [7].

The model and the focus of our work relates to Lan [8]. They use a pictorial structure model and aim at handling both self-occlusions and changes in the viewpoint. This way of exploiting constraints provided by walking may be considered as a possible improvement in our initialization step. Lan [9] extended the approach by taking limbs coordination into account. Both publications work with silhouette data captured from a single camera viewpoint.

The idea of our model is also related to Lim [10]. They present a multi-cue tracking procedure. Their shape model is comprised of several silhouettes learned offline from training sequences. The shape model is used along with an appearance model learned online for a given individual. The shape model is learned from a built collection of normalized foreground probability maps of humans. These probability maps are clustered into sets using K-means clustering algorithm. A mean image is then built for each set creating a representation of a pose. Tracking is then formulated as finding a set of warp parameters which map a foreground

blob in a frame onto one of the silhouettes in the shape model. Their appearance model could be an inspiration for a further enhancement of our approach.

Another source of inspiration may be found in Howe [11]. The author aims at applying silhouette lookup to monocular 3D pose tracking. A knowledge base of silhouettes associated with known poses is populated. A silhouette extracted from an input frame identifies a set of silhouettes in the knowledge base. Subsequent Markov chaining exploits the temporal dependency of human motion to eliminate unlikely pose sequences. This solution is further smoothed and optimized. The idea of maximizing the per-frame match to the observations and the temporal similarity between successive frames simultaneously may also be applied to further enhance our solution. Adding a term to a cost function employed in our method rewarding a solution close to the one found in the preceding frame would only represent a minor alteration of our algorithm.

Our approach also relates to Collins [12]. They adopted object tracking approach based on color histogram appearance models. In addition, they consider both object and background information to better distinguish an object from its surroundings. Their mean-shift tracking system models object and background color distributions and tracks targets through partial occlusions and pose variations. Their work is restricted to tracking of rigid objects. As opposed to their approach we use mean-shift to fit an articulated model to image data. Our method is based on the motion segmentation computed using the adaptive background model.

Using the methodology introduced in Section 2, the model introduced in [6] could be adopted as a model with the least refined structure suitable for tracking persons at a low image resolution. Our model could be viewed as a more refined model, half-way to a model representing all body parts. More complicated models accounting for arm motion and aspiring to higher accuracy seem to be impractical in our context as the computational demand appears to be too high.

Substantial part of this work was done during diploma thesis project of the first author defended in February 2006 at the Czech Technical University in Prague.

4 Extended Six-link Biped Model

A human body may project to the image in a variety of forms. The articulated structure of a person makes a vision-based tracking difficult even if a constrained type of motion as walking or running is assumed. The idea behind the model-based approach is exploiting explicit a priori knowledge about the human body appearance in 2D.

A desired human model for surveillance should be simple for computational speed reasons and should enable capturing appearance of a variety of individuals. On the contrary, the model should possess enough structure allowing the expression of distinguishing appearance of a human.

4.1 Body model

We adopted the 2D five-link biped articulated model of [5] and extended it into a six-link model. Our model has added an articulated head and it can also cope with a frontal view of a person. The biped human model consists of six rectangles representing individual body parts. The rectangles are connected by the joints. The model is shown in Figure 1a.

Arms are not included in the model as a reliable recovery of exact arm positions of distant pedestrians is often difficult. The biped model is also intentionally kept simple having in mind that low computational demand is desired.

The model is parameterized by eight parameters. Finding a posture T using the biped model M_b means determining its eight parameters, $T = \{C = \{x, y\}, \Theta = \{\theta_1, \ldots, \theta_6\}\}$, where C is the model center and Θ is the inclination vector consisting of angles between the axes of the body part and the vertical axis y as shown in Figure 1a.

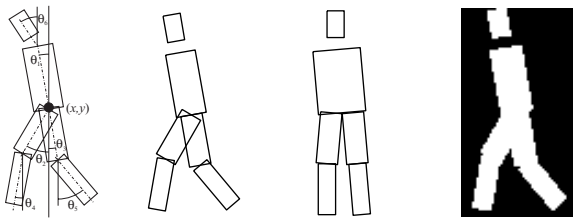


Fig. 1. Six-link biped human model (a). Lateral (b) and frontal (c) model view. Silhouette mask (d) used for evaluation of the extended model.

The biped human model is designed to cope with changing scale which is common in surveillance sequences due to perspective projection from 3D $\rightarrow$ 2D. The humans further from the observer look smaller. A scale parameter of the model deals with the foreshortening of a projected human.

Each rectangle of the model is defined by its height l and length of its base b. Each body part is described by $\{\alpha, l\}$, where $\alpha = b/l$ is a base-to-height ratio. Body part heights are normalized with respect to the height of the torso l_1 and the shape model is thus scale-invariant. A shape S of a human is parameterized as follows: $S = \{K, R\}$, where $K = \{\alpha_1, \ldots, \alpha_6\}$ is the base-to-height ratio vector, $R = \{r_1, \ldots, r_6\}$ the relative height vector, $r_i = l_i/l_1$, $i = 1, \ldots, 6$. The six-link biped model is composed as $M_b = \{T, S\}$. The frontal and lateral view is coped with this model. The difference is in parameter α_1 (base-to-height ratio of the torso) and positions of joints connecting the thighs and the torso. The proportions can be seen in Figure 1b,c.

4.2 Extended body model

We extended the biped model to consider both the region explained by the model and the region in the neighborhood of the model which is not represented by the model. A similar idea has been previously applied to a selection of a configuration of models explaining observed clutter of humans [6].
The extended model M is formed by adding a rectangular neighborhood to the biped model M_b. This neighborhood has the height of the biped model M_b. The width of the rectangular neighborhood used is chosen so that it is always possible that the whole walking body fits in the region. The width is calculated as $0.53 \cdot$ height. The rectangular neighborhood and the biped model M_b share a common center.

4.3 Model evaluation

The motion image I is the result of a pixel-based motion detector and constitutes the input to the human detection/tracking system. I is formed as a binary image the values of which tell if the pixel moves/is stationary with respect to the local background, see Figure 8a,b,c for examples of the outcome of motion segmentation. The commonly used motion detection algorithm is based on a finite mixture of Gaussians, see Stauffer [13]. This approach can deal with a slowly changing background. The implementation of the motion detection algorithm by Fexa [14] is used.
To evaluate the model in current frame, a silhouette mask is created according to current model view and posture, see Figure 1d. This mask is used to compute the amount of missing measurements in the region RB explained by the biped model (white region) and the unexplained observations in the model neighborhood RB (black region). Region RE represents a complement of the region RB within the area of the extended model.
The cost of the current posture of the model M is calculated from the motion image I as

$$C(I \mid M) = \exp\left(-a\left[1 - \frac{1}{A_{RB}\sum_{x,y\in RB}} I(x,y)\right] - b\left[\frac{1}{A_{RE} \setminus A_{RB}} \sum_{x,y\in RE\setminus RB} I(x,y)\right]\right) \tag{1.1}$$

where A_{RB} and A_{RE} are the areas of the regions RB and RM, respectively. Scaling parameters a and b were determined experimentally.

5 Matching Extended Model against Data

A model M^* of a human which fits the motion image I best is sought. Using the cost function in Equation (1.1), the problem can be formulated as

$$M^* = \underset{M}{\operatorname{argmax}}\, C(I \mid M) \tag{1.2}$$

5.1 Mean shift optimization

Finding globally optimal parameters of the model M^* in Equation (1.2) is computationally complex. We engage the mean shift algorithm, an iterative procedure that shifts a data point to the average of data points in its neighborhood, to find locally optimal solution.

The mean shift idea was introduced by Fukunaga [15]. It was shown that the mean shift vector pointing to a sample mean of local samples generally points in the direction of higher density and thus provides a gradient estimate. The mean shift algorithm was proposed in Fukunaga [16]. The method was then generalized and presented as a mode-seeking process by Cheng [17]. At last, the convergence of the iterated mean shift procedure on discrete data was proved in Comaniciu [18].

We start from some initial estimate of the current stance $\mathbf{v} = (v_1, \ldots, v_8)^T = (x, y, \theta_1, \ldots, \theta_6)^T$. The local mean shift vector represents an offset to a posture vector $\mathbf{v}'$, which is a translation towards the nearest mode according to the cost function in Equation (1.1). The local mean shift vector is computed within the local neighborhood $\{\mathbf{v}_i\}_{i=1\ldots n}$ of the point $\mathbf{v}$. The point $\mathbf{v}_i$ is assigned a weight $C\left(I \mid M(\mathbf{v}_i)\right)$ according to Equation (1.1), where $M(\mathbf{v}_i) = \{T(\mathbf{v}_i), S\} = \{\{\{v_1, v_2\}, \{v_3, \ldots, v_8\}\}, S\}$.

The new posture vector $\mathbf{v}'$ is computed employing the uniform kernel as

$$\mathbf{v}' = \frac{\sum_{i=1}^{n} \mathbf{v}_i C\left(I \mid M(\mathbf{v}_i)\right)}{\sum_{i=1}^{n} C\left(I \mid M(\mathbf{v}_i)\right)} \tag{1.3}$$

Starting from the initial estimate of the current stance, the new posture vector is repeatedly computed until it converges to the sought stance.

5.2 Decreasing computational complexity

The model M has eight degrees of freedom and so we are faced with a multidimensional optimization problem. An iterative optimization procedure was suggested in Section 5.1 to solve the task. However, seeking for an optimal posture in eight dimensional space, i.e., computing an eight dimensional mean shift would still be too time consuming. We decreased the computational demand by both reducing the size and dimension of the considered model parameter space.

The size of the mean shift search space may be reduced in several ways. First, impossible orientations of the individual body parts are considered. Second, we exploit the fact that an observation of walking/running humans is assumed. We thus expect the torso to stay upright and further constrain the ranges of possible orientations of the limbs according to the type of motion assumed. Last, location constraints are considered when initializing the model. This last simplification will be further addressed in Section 6.2.

The complexity of the problem is further decreased by reducing the dimension of the considered search space. This is achieved by not optimizing all the parameters at once, but always only a subset of those. Satisfactory results were achieved by splitting the problem into several 2D and one 1D optimization problems instead of considering all eight dimensions. This point will be discussed in more detail in Section 6.2.
If a subset of parameters is optimized then only body parts involved are used to evaluate the model, i.e., to create a model silhouette mask. All body parts are used only if the current body posture is evaluated to decide whether the model provides sufficient explanation of seen data and hence if tracking can be started or continued.

6 Detection and Tracking

In our algorithm, we first use simple means to find a human candidate, i.e. we focus attention to a ROI. Subsequently, the model is validated at the candidate location. At last, the tracker is started to follow the target in the subsequent frames.

6.1 Human candidate detection

The attention is focused to the ROI in the motion image I where human candidates could be as illustrated in Figure 8a. This step significantly reduces the amount of processed data and contributes to computational efficacy. ROI is detected by correlation of a uniform rectangular mask with the motion image I. The local maxima of the correlation point to moving entities of human size in the image. These maxima are used as centers of ROI.
The local maxima of the correlation refer to moving entities of human size in the motion image. At present, only the strongest correlation maximum is used as the current implementation tracks only a single person.
The height and width of the correlation mask has to be set. We work with expected scaling of humans at a given vertical image location. This information is obtained after a simple calibration performed manually for a given image capturing setup. It is assumed that the scale of a human is linearly dependent on the vertical image location. Such consideration is based on the assumption that a camera provides an oblique view of a planar scene. The more distant humans appear smaller and located higher in the 2D projection.
This step is inspired by [6]. We added the least-squares estimate of the scale function which is based on the height of multiple humans present in the scene at different vertical image positions.
In present implementation, it is also assumed the humans are roughly of the same size. Even though minor height variations should be tolerated by the model at considered scale, taking toddlers into account would require a scale adjustment in the initialization phase.

The width of the mask used in correlation is chosen so that it is always possible that the whole walking body fits in the region. The width is calculated as $0.53 \cdot$ height.

6.2 Model validation

Having detected human candidate, we try to validate our model, i.e., to see whether the model explains data satisfactorily. We first initialize our model at the candidate location. The initialized posture is then further refined in attempt to achieve better fit of the model to data. The resulting pose is evaluated according to Equation (1.1).

Model initialization: After a ROI is detected, a 2D 6-link biped human model is initialized at the candidate location in the motion image I, see Figure 8b for illustration. This means that the position of the model and orientation of individual body parts in the image are determined. The pose explaining the data optimally is searched using the mean shift algorithm introduced in Section 5.1.
It is convenient to initialize the model incrementally. The torso is initialized first, followed by limbs and head. Finding the initial body posture as a whole would be computationally demanding. Starting from torso leads to insensitivity to unimportant phenomena in the image such as shadows (to be described later).
A subregion corresponding to the torso within the ROI is sought first. This subregion has the width of the ROI. The height and the position of the subregion correspond to the trunk of a person at this scale. The trunk region is initialized in the middle of the selected subregion. The optimal horizontal position and orientation of the trunk is found according to Equation (1.3) within this subregion. The horizontal position of the model is updated.
Detected ROI yields a possible body center of a human. We expect, however, this information to be only a rough estimate of the true position. We try to improve our estimate by handling the data on a finer resolution. For this reason, we do not start by setting the model center to the middle point of the detected human silhouette, as opposed to [5]. We assume that some parts of the segmented region may not belong to the human body. A shadow underneath the body is very common. Torso tends to be less affected by such shadows. The body center update based on torso initialization usually does not take such spurious subregions into account when the body posture is estimated.
After positioning the torso, we proceed by placing the limbs. Orientation of both thighs is determined at once, followed by positioning of both calves. Treating thighs/calves simultaneously yields a better fit. Finally, the head orientation is found.
Initializing the body parts incrementally yields three 2D and one 1D optimization problems instead of one huge 8D problem. The remaining one degree of freedom (vertical position of the whole model) was already initialized when ROI was set. This reduces significantly the problem complexity.
Both frontal and lateral model views are initialized in this step. The model view better explaining the motion image according to Equation (1.1) is chosen.

Posture refinement: The model initialized in the previous step, keeps the vertical position of the previously detected ROI. Its horizontal position is allowed only within the ROI. The posture refinement does not take this limitation into account any longer. We aim to optimize the previously initialized human stance. The possible wrong vertical position of the model estimated in the previous step is refined in this phase.

Individual model parts are again handled separately with the goal of decreasing computational demands. There is a tradeoff between the model/image data fit and the degree of model separation in optimizations.

Satisfactory results are achieved when following iterations are run until convergence. First, the horizontal position of the model is optimized. Second, the vertical position of the model together with the trunk orientation is calculated. Third, both thighs are fit simultaneously. Fourth, both calves are adjusted to the data at the same time. Finally, the orientation of the head is estimated. See Figure 8c,d for outcome illustration. The found model is superimposed on both the original data (Figure 8d) and the motion image (Figure 8c).

The model fails to explain data satisfactorily if the cost, recall Equation (1.1), of the current model posture does not exceed an experimentally estimated threshold. In this case, the ROI location is not considered any more in the current frame.

6.3 Tracking

When the initialization of the model is successful then the mean shift tracking is launched in the next frame.

The pose of the model is available from the previous frame. By considering dynamics of human motion, we take the advantage of temporal constraints provided by walking and running and modify the pose. Dynamics are implemented as half a dozen hand coded rules which take periodic motion into account. For instance if the limit position of limbs is reached then the movement in the opposite direction is anticipated. Another rule is meant to anticipate the swing of the calf when both legs are occluded. The speed of this body part is greatest in this phase of a gait cycle and the tracker looses the body part in case the number of frames per second is insufficient. This step aims at anticipating the human pose in the current frame based on the model pose found in the preceding frame. Introducing dynamics helps to overcome the local character of the optimization procedure.

Tracking uses the same iterative procedure described in Section 6.2 which starts from the anticipated pose and is repeated until convergence.

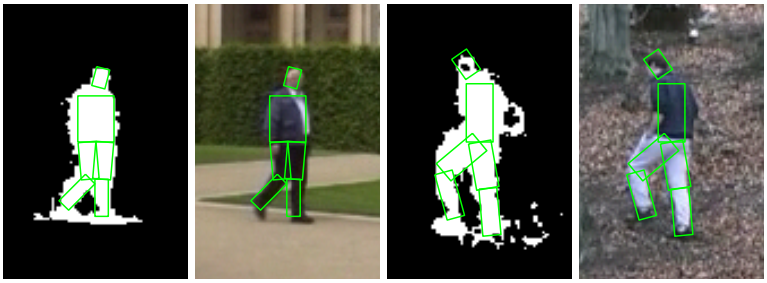
It is assumed that the frame rate is sufficiently high so that the person tracked remains in the scope of the mean shift kernel. In case the tracked person is lost the algorithm is restarted.

7 Experiments and Lessons Learned from Them

The method was tested on several real sequences, see the CMP Demo Page [19]. The motion detection algorithm [13] provides the sequence of binary images telling which pixels are in motion. This sequence is the input to our implementation. Our model-based tracking is implemented in MATLAB.

7.1 Body model

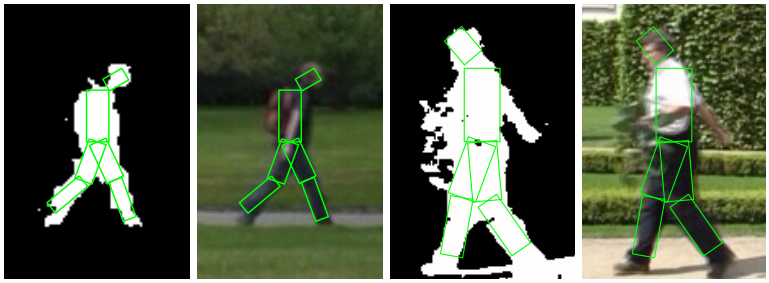
We choose an articulated biped model without arms. The choice of the employed model can be justified by observing the found model postures superimposed on the silhouette images of tracked subjects.
The extension of the model by the head improves matching the model against image data. The head together with shoulders is a significant characteristics for a frontal human appearance especially if limbs cannot be distinguished in the outcome of the motion segmentation. This feature helps to position the model horizontally, see Figure 6b. The improvement of employing the articulated head is noticeable when the person is seen laterally and at the same time the head is not coaxial with the torso. A better fit in vertical direction is found in this case, see Figure 3b.
It can be seen that there is very little information in the motion images at considered scale for recovering the position of arms. This task would require using additional features other than segmented motion and would lead to slowing the system down. The position of arms is not crucial in our context. Hence, we do not consider arms in our model. The legs and the head, on the contrary, appear to be distinguishing features. For this reason, we include these parts in the employed model. By observing the silhouette images, it may be verified that the used model corresponds to the level of detail provided in data at the considered scale. See the images in Figure 8c for examples.
The frontal model view and the lateral model view are used to fit a person seen from a corresponding perspective. The cost function in Equation (1.1) provides a criterion for choosing a view when facing the task of fitting the model to data on individual frame level. Figure 8a,b,c shows examples of initializing the model to motion data. These images depict several poses and both the frontal and the lateral view. In Figure 8d, the resulting posture and the chosen model view are superimposed on the original image data. Figure 2 illustrates that both the frontal and the lateral model view may be employed to fit a person observed from a range of diagonal views. In Figure 2a,b fitting of the frontal model view to a person viewed diagonally may be observed. Figure 2c,d depicts how the lateral model view is matched against a human seen diagonally. The motion images are displayed on the left of the original data.
We tested the robustness of the model by fitting it to images of persons walking under different conditions changing slightly their appearance in the motion data. Figure 3 illustrates outcome of the experiment. The tested conditions included walking with a backpack (Figure 3a,b) and walking with a plant in hand (Figure 3c,d). Again, the motion images are displayed on the left of the original data



(a) (b) (c) (d)

Fig. 2. Fitting the frontal model view (a,b) and the lateral model view (c,d) to a person seen diagonally. The detected model is superimposed on the motion data (a,c) and the original images (b,d). The images of the segmented motion appear on the left of the corresponding original image.

in Figure 3. The model was matched properly in both cases. It appears that slight variations in the appearance of the silhouette image are handled properly by the model.



(a) (b) (c) (d)

Fig. 3. Matching the model against silhouette images of persons walking under different conditions: walking with a backpack (a,b) and walking with a plant in hand (c,d). The detected model is superimposed on both the motion data (a,c) and the original images (b,d). The images of the segmented motion appear on the left of the corresponding original image.

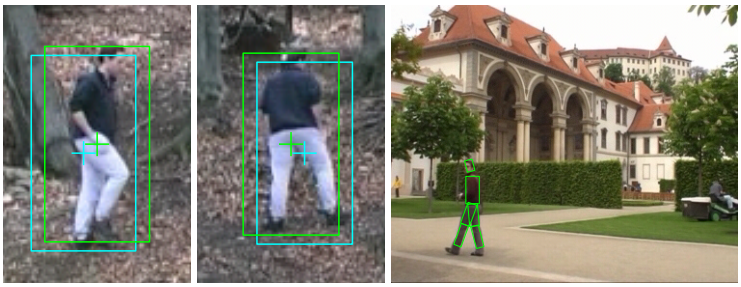
Our experiments also illustrate that a model capable of accounting for a human articulated structure allows not only to detect a desired object but also provides better accuracy as opposed to one that is less refined. Model center found by the simple rectangular model employed to find the ROI is typically improved by employing articulated model. Figure 4a,b shows two detailed views from a sequence with a person walking in a garden (WalkingPersonKinskyGarden.avi, [19]). The detected ROIs and the corresponding model centers are drawn in cyan. Final lo-

cations of the models are drawn in green. Frame in Figure 4a presents a person seen laterally. The position was improved in the direction of both x (6 pixels) and y (4 pixels) image axes. The height of the model at the ROI location and at the final location was 89 and 88 pixels respectively. Figure 4b displays a person seen frontally. The center location was improved in both x and y image axis by 4 and 6 pixels respectively. The height of the model at the ROI location and at the final location was 83 and 82 pixels respectively. Variable scale of the human present in the scene is caused by the perspective foreshortening.

All individuals in our experiments were matched using the same model. It is only the scaling factor that varies. Experimental results illustrate in the presented images that the chosen model may be fit to a variety of individuals at different scales.

7.2 Distinguishing human appearance from non-humans

The ability to discriminate humans from other moving objects present in the scene is illustrated in a sequence which contains a walking person and a small agricultural tractor moving in the background (MovingPersonAndTractorWallensteinGarden.avi, [19]). A frame where both the figure and the moving object are far apart is shown in Figure 4c. Both moving objects represent two separate regions in corresponding motion image. Current version of the proposed algorithm tracks one individual. However, all detected ROIs were processed by our algorithm in this particular frame. Human appearance has been correctly detected by the model and the moving tractor remained unnoticed. However, it is only the small agricultural tractor that was present in the sequence. This experiment provided thus a relatively easy case to test. Hence, the ability to distinguish humans from non-humans needs to be further tested on subjects closer to the appearance of persons.



(a) (b) (c)

Fig. 4. Improving accuracy by employing articulated model (a,b). A simple rectangular model employed to find ROI is drawn in cyan. Our articulated human model is drawn in green. Discriminating human appearance from a moving non-human object (c).

Figure 5 shows two different detailed views from the sequence mentioned above. Both the motion segmentation (Figure 5a,c) and the original data (Figure 5b,d) are provided. The walking person occludes the moving object in the first view (Figure 5b). As a result, regions in the motion image (Figure 5a) corresponding to the person and the moving object merge into a single blob of foreground pixels. A ROI was found at this location and both model views were initialized. The frontal model view was classified as a view better explaining the observed data. The detected human yields a relatively good fit as far as the region explained by the biped model is concerned. However, there are too many foreground pixels in the biped model neighborhood in this case. As a result, the data is classified as unexplained and the model is rejected. In the second view (Figure 5d), the person still partially occludes the moving object. Again, a relatively good fit is found with regard to the region explained by the biped model. However, there are less unexplained observations in the region outside the biped model in this example (Figure 5c). The data is thus classified as sufficiently explained and the human appearance is validated. It can be observed that a correct view has been identified in this case.

(a) (b) (c) (d)

Fig. 5. Detecting a walking person occluding other non-human moving object. A rejected model is superimposed on the motion data (a) and on the original image (b). A successfully validated model is superimposed on the motion data (c) and the original image (d). The images of the segmented motion appear on the left of the corresponding image of the original data.

The pose found by the proposed algorithm may yield a correct pose even in the case the data is classified as unexplained at last. Regardless of the incorrectly chosen view, the pose found in Figure 5b may be described as a correct one.
A rejected model may represent a correct pose provided characteristic features are present in the input data. It is the head and the legs that help the algorithm to place the model in this particular case. However, the uncertainty that a human appearance was found in this particular case was considered too high and the detected pose was rejected.
A found posture is rejected whenever the characteristic appearance of the human silhouette is lost in a large foreground blob. Motion data do not allow to

discriminate the desired object any longer in this case and the process needs to be supported using other information. An appearance model [10] would provide additional cues on the individual frame level. These two approaches have complementary strengths, and may support each other. See Elgammal **??** for more details on learning representations for the shape and the appearance of moving objects. Introducing a model of human dynamics would help to tackle the problem using temporal information. The dynamical model could further be extended by integrating prior knowledge via an a priori pose distribution. Gall **??** discusses how such pose distribution may be learned from training samples. In the current work, however, only tracking employing the proposed extended human model is implemented. A multi-cue tracking is considered in the future work.
The motion detection algorithm is capable of handling slowly moving background. However, background clutter is still very common in the outdoor motion images. Vegetation such as grass, trees or bushes swaying in the wind is often a cause of such background clutter. It is thus crucial to distinguish the desired object from such spurious artifacts. The ROI detector uses a simple rectangular model to find human-like rectangular regions. A threshold is responsible for accepting the ROI. Such region possesses a portion of foreground pixels which may constitute a human silhouette. A good fit of the model is then possible provided a human is really present. The rectangular model considers the data in the region as a whole. The extended model, on the contrary, provides means to account for the structure of the data. The configuration of the foreground pixels has to be such that most of the pixels appear in the subregion explained by the biped model and minimum elsewhere in the considered region. A human like silhouette is formed in this case. Our articulated model together with basic kinematic constraints allows to validate variety of human silhouettes. Our experiments suggest that the model is sufficiently robust in the case of background clutter.

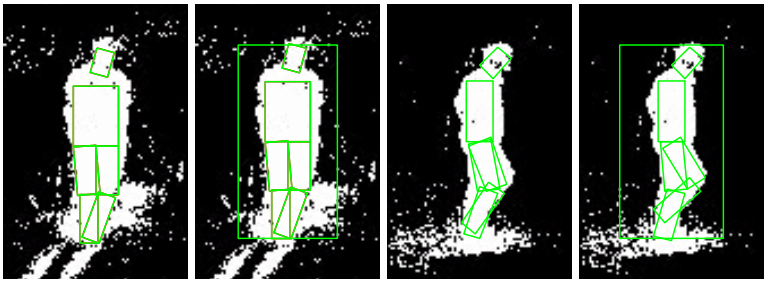
7.3 Employing extended model to fit data

The principal idea of fitting the proposed extended model to a motion image is to minimize the amount of missing measurement in the region explained by the model and at the same time the amount of unexplained observations in the model neighborhood. To illustrate how the extended model improves matching the model against data as compared to the biped model without the extension we performed two types of experiments. The model was initialized in a region with a human silhouette. First, we initialized the biped model without the extension. This is achieved by simply neglecting the second term in the cost function in Equation (1.1), i.e, by setting the parameter $b = 0$. The unexplained observations in the model neighborhood have no influence on the value of the cost function in this case. In the second experiment, we initialized the proposed extended model. These two experiments show that employing the extended model brings two main benefits.
First benefit, the model placement is improved. A shadow underneath the human body is common and often causes a significant artifact in the motion image.

The head on the other hand is usually less pronounced. A strong shadow and a less pronounced head sometimes lead to a false model placement. This happens in the case when we only consider the region explained by the model. It would result in fitting the model to the strongly pronounced but unwanted region. The second term in Equation (1.1) is responsible for taking the unexplained observation into account. This helps to tackle the problem by lowering the cost of the posture when a shadow is present, helping thus to drive the model away and identifying the head at the same time. Figure 6a,b shows a detailed view from the sequence with a person walking in a garden (WalkingPersonKinskyGarden.avi, [19]). Only motion segmentation data are displayed. The images present a rather noisy silhouette of a person seen frontally. A well pronounced shadow region is present underneath the body in the motion image. The model superimposed on the motion data in Figure 6a was initialized without the extension. On the contrary, the model displayed in Figure 6b was initialized employing the proposed model extension. The position of the extended model was improved by two pixels in vertical direction in this particular case.

Second benefit, the extension of the model improves limbs placement. A fitting procedure is desired which yields a proper position when (a) one leg occludes the other and also (b) it rewards placing of limbs astride if possible, as illustrated in Figure 6c,d. A procedure minimizing only the missing observation in the region explained by the model would lead to a proper position in the case (a), however, it would often result in placing occluded legs also in the case (b). The first term in Equation (1.1) forces the body parts to cover the motion region while minimizing the missing measurement in the region explained by the model and solving thus the case (a). The second term in Equation (1.1) should be insignificant in the case (a) when little unexplained observation is present in the model neighborhood. In the case (b), however, the cost function should force the body parts to cover the region maximally thus lowering the value of the second term. Figure 6c,d shows a detailed view from the sequence mentioned in the preceding paragraph. Again, motion segmentation data are displayed. This time, a person seen frontally is present in the images. In this experiment, we aim to illustrate the criterion presented in the Equation (1.1). However, the outcome of both experiments may be difficult to evaluate if the optimization employed in our algorithm is used. If two local optima are found in both experiments, it is difficult to judge what leads to such outcome. The resulting dissimilarity of the results may be caused by the criterion. However, it may also be caused by the local optimization which detected two local optima which are close to each other or far apart. Hence, the model position was fixed and the pose was found by the exhaustive search in this case. This enables to illustrate the fitting criterion and to neglect the influence of the optimization procedure. Figure 6c illustrates how the model was initialized without the extension. The model displayed in the Figure 6d was initialized with the use of the proposed model extension. Improved placement of the legs may be observed in the case of the proposed extended model. Figure 6c,d illustrates that the cost function in Equation (1.1) rewards placing the limbs in a way that covers the foreground region maximally.

This helps to find the correct pose on the individual frame level and leads to a better fit in case (b) mentioned above.

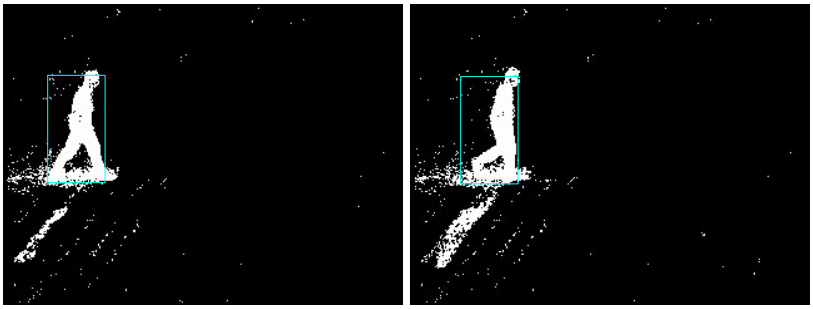


(a) (b) (c) (d)

Fig. 6. A slightly improved model placement and limbs localization. Matching model against motion data without (a,c) and with extension (b,d).

7.4 Detection and tracking

Detecting human candidates: The first column of images in Figure 8 illustrate the outcome of the ROI detector. It may be observed that a human silhouette is found approximately. This step is clearly influenced by the shadow underneath the body which shifts the detected model center under its true location. Hence, the ROI detector fails to include the head in these cases. In addition, these results illustrate that the horizontal position also needs to be refined. The third row of images in Figure 8 has two parallel traces present in the bottom part of the view. These were caused by a sudden change of illumination. Resulting shape of these artifacts was caused by surrounding trees. Such artifacts further influence the task of ROI detection.
Figure 7 shows the outcome of the motion segmentation of two frames from the sequence with a person walking in a garden (WalkingPersonKinskyGarden.avi, [19]). The outcome of ROI detection is superimposed on the motion data. The case presented in Figure 7a shows a person with legs being apart and the body appearing approximately symmetrical with regard to the vertical image axis. A relatively good estimate of the model center is yielded by the ROI detector in this case. In addition, the shadow underneath the body does not seem to influence the detection in this particular case and most of the region representing the head is included in the ROI. In the second frame displayed in Figure 7b, however, the person does not appear symmetrical with respect to the vertical image axis any more. The estimated model center is clearly shifted to the left regarding the true position. Additionally, due to the pronounced shadow underneath the body, part of the head is excluded from the detected region, as shown in Figure 7b.
This step provides only a rough estimate of the true position of human candidate.



(a) (b)

Fig. 7. Detecting the region of interest (ROI). The found ROI is superimposed on the motion data.

Model initialization: The ROI detector yields a location where a human could be. The extended human model is then initialized in the detected region. The model initialization step is crucial for several reasons. First, it has to be verified whether a person is present at the candidate location. Second, provided a human is really present in the ROI, initializing the model to data should result in a fit that explains observed data sufficiently. If the above conditions are fulfilled then the tracker is started. Third, the correctly initialized view and posture lead to proper tracking. Initialization procedure that would lead to frequent erroneous rejection of a human candidate would cause the high false negative rate. That is undesirable in the considered application area. For these reasons, it is vital to initialize the model correctly for successive tracking. We tested the functionality of the proposed initialization procedure in three experiments. The outcomes were evaluated by observation.

In the first two experiments we considered models detected in every frame. We looked at several issues. These issues included: portion of frames with a detected person (regardless of the fact whether the model was successfully initialized or rejected), portion of frames with a successfully initialized model, portion of frames with a properly identified model view, portion of frames which have a model with all the parts positioned correctly and at last, portion of frames which contain a model with maximally one misplaced part. A person is considered as detected in case the region representing the human appears mostly in the region of the extended model. A part is classified as correctly matched provided that a greater part of the region representing the part is explained by the corresponding model part.

The process is influenced by a number of aspects with regard to the considered scene and camera setup. Hence, for the first and the second experiment we chose two sequences from our data-set with relatively steady conditions that allowed evaluation of the chosen approach. On the other hand, both experiments presented persons carrying an object. We applied the ROI detection and subsequent

Test sequence	Frames	Detected	Initialized	Correct view	All parts	Max. 1 part
Sequence 1	108	100%	100%	99%	96%	100%
Sequence 2	60	100%	100%	98%	87%	100%

Table 1. Human candidate detection and model initialization. Summary of the experiment where ROI detection and subsequent model initialization are applied to every frame separately. Table shows number of frames, portion of frames where ROI has been detected, portion of frames where model has also been successfully initialized, portion of frames with correctly identified model view, portion of frames where model with all its parts is positioned correctly and at last, portion of frames which contain a model with maximally one misplaced part. A person is present in all considered frames.

model initialization to every frame of the test sequence. No information from the preceding frames was propagated to subsequent ones.

Sequence 1 had 137 frames of 360×288 pixels each, 25 frames per second and presented a relatively fast walking pedestrian carrying a backpack (WalkingPersonWithBackpackCharlesSquare.avi, [19]). This made the model view choice more challenging. The person walked from the left to the right and was fully visible in 108 frames. The person was observed laterally at a constant scale. The height of the person in the images was 75 pixels. A ROI is initialized when an object is believed to appear fully in the image. For this reason, we only used these frames for a test assessment, i.e., 108 frames. All the detected models were considered in the evaluation including the rejected ones. See Figure 3b for detailed view of the tracked person and Table 7.4 for results.

Sequence 2 had 116 frames of 360×288 pixels each, 25 frames per second and showed a walking gardener with a plant in hand (WalkingPersonWithPlantWallensteinGarden.avi, [19]). Again, this should make the model view choice and the fit itself more challenging. First ROI was detected in a frame number 26. Subsequently, ROI was provided in the next 60 frames. The person walked from the right to the left. The person was observed laterally at a constant scale. The height of the person in the image was 150 pixels. Again, only those frames for a test assessment were used in which a ROI was available, here 60 frames. All the detected models were used in evaluation. Figure 3d shows detailed view of the tracked person and Table 7.4 results of the experiment.

The purpose of the third experiment was to learn how the individual steps of our algorithm preceding the tracking itself contribute to a resulting initial fit. Another goal was to assess the overall functionality of the initialization step under more varying conditions. The sequence which has a person walking in a garden (WalkingPersonKinskyGarden.avi, [19]) was found to be appropriate for the task. The tested conditions included variable scale, changing viewpoint, present shadows, background clutter and noisy data yielded by the motion detector. The scale of the person in these experiments varied from 82 to 88 pixels. An overview of the experiment is given in Figure 8.

The first column of images in Figure 8 shows the outcome of the ROI detector. These views illustrate how the detection was influenced by shadows, background

clutter and noise in the data. Results of ROI detection shown in the first column of images in Figure 8 suggest that the model center needs to be refined in the direction of both image axes.

Views in the second column in Figure 8 depict how the model was initialized in the ROI. Typically, the horizontal position of the model is improved and the first rough fit of the model is found in this step. However, the model vertical position is fixed in this phase and keeps the values yielded by the ROI detector. The model view was correctly chosen in these examples.

As previously explained, the vertical position yielded by the ROI detector needs to be corrected in the presented examples. The third column of images in Figure 8 shows the final fit of the model after the refinement step. The final model posture superimposed on the original data may be seen in the last column. Comparison of these images with the ones illustrating the rough initialization reveals that it is the vertical location in the first place that was refined in this step. Further inspection shows that orientations of the remaining body parts have also been slightly improved. See the second and the third image in the first row in Figure 8 for example. The vertical location was clearly improved in this case. This instance further illustrates the role of the model extended by the head when dealing with unwanted phenomena such as shadows. The model was driven away from a well pronounced region to a region less significant in the image. However, a location was found which better corresponds to the employed model and the head was thus identified. These two views also illustrate how, for instance, the orientation of the front calf was improved in the refinement step.

Tracking: Our data-set contains videos of six individuals, who are walking in the outdoor environment under varying conditions: walking with a backpack, walking with a plant in hand, walking with a jacket over the shoulder and walking with a small agricultural tractor moving in the background. The types of motion tested included slow walk, fast walk, running, standing still shortly and their transitions. Both tracking at the constant and slightly variable scale was tested. Our tracking algorithm was presented with persons seen in continuum of views ranging from frontal to lateral. In addition, view transitions were tested. At last, the experiments illustrate how partial and full occlusions are handled. The sequences were processed by the proposed tracking algorithm, see CMP Demo Page [19], all using the same parameter setting (the view calibration was different for each scene, of course).

Let us further demonstrate the results on two sequences in more detail. It has been already said that a model part is classified as correctly matched provided that a greater part of the region representing the projected body part is explained by the corresponding model part. Results were again evaluated by observation. Only those frames were used for a test assessment, where a ROI was available.

In the first experiment, the proposed tracking algorithm was tested on the sequence which has a relatively fast walking pedestrian carrying a backpack (WalkingPersonWithBackpackCharlesSquare.avi, [19]). This sequence was previously introduced as Sequence 1.

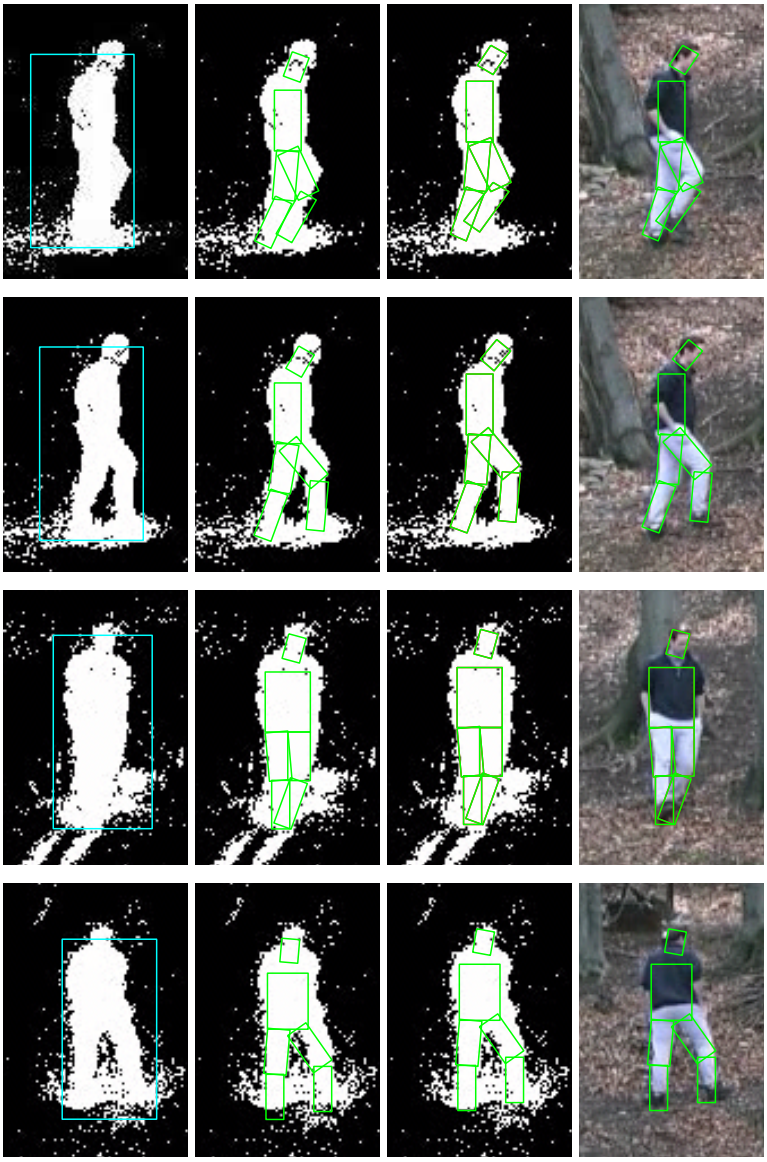


Fig. 8. ROI detection (a), model initialization (b), posture refinement (c) and outcome (d). Images of segmented motion are displayed as binary black and white images. Original data is displayed in the column (d). The found model is superimposed on both types of data. Courtesy A. Fexa for the sequence.

The person has been correctly detected and the model initialized in the first considered frame. The algorithm tracked the person properly throughout the se-

Test sequence	Frames	# of re-initializations	All parts	Max 1 Part
Sequence 1	108	0	98%	100%
Sequence 2	60	2	91%	100%

Table 2. Tracking. Summary of the experiment where ROI detection, model initialization and subsequent tracking are applied. Table shows number of frames, number of times tracking has been re-initialized, portion of frames where model with all its parts is positioned correctly and at last, portion of frames which contain a model with maximally one misplaced part. A person is present in all considered frames.

quence. No re-initialization of the tracking procedure was needed. All the model parts were matched correctly in 98% of the frames. In the remaining 2% of the frames, only one body part was classified as being not positioned properly. However, these cases were still found to be close to the correct position.
In the second experiment, the tracker was tested on the sequence which showed a walking gardener with a plant in hand (WalkingPersonWithPlantWallenstein-Garden.avi, [19]). This sequence was previously introduced as Sequence 2.
The person has been correctly detected and the model initialized in the first frame. The algorithm tracked the person properly in the next 53 frames. The tracker re-initialized twice at the end of the sequence. The first re-initialization happened due to the wrong position of a leg. The second re-initialization was due to many unexplained observations in the motion data caused by the carried object. The tracking process continued afterwards. All the model parts were matched correctly in 91% of the frames. Only one body part was classified being positioned improperly in the remaining 9% of the frames. It appears that an incorrect scale of the person determined during the calibration step is the cause of the misplaced parts. It was mostly the head that appeared above its true position. Smaller model may yield a better fit in these cases. Results of the the two experiments are summarized in Table 7.4. Processed sequences can be viewed at the CMP Demo Page [19].
The sequence which has a person walking in a garden (WalkingPersonKinskyGarden.avi, [19]) was found appropriate for testing slightly variable scale, changing viewpoint while walking or turning, present shadows, background clutter and noisy data yielded by the motion detector. The scale of the person in the experiment varied from 82 to 88 pixels. Variety of conditions mentioned above were satisfactorily handled by the proposed algorithm. Results may be observed at the CMP Demo Page [19].

8 Conclusion

This work contributes to the 2D model-based whole body tracking in motion images. The extended biped model of humans was proposed. It has been shown that the model is general enough to capture the appearance of a variety of individuals. Our experiments also suggest that the model possesses enough structure to express the distinguishing characteristics of humans observed in the motion

images. Model parametrization is scale independent and allows tracking of a person at a variable scale. The frontal and the lateral views are treated in a single framework. As a result, a person tracked by our method may be seen from continuum of viewpoints, as opposed to [3–5].

The model to data fit criterion considers both motion image data in the region explained by the biped model and in its neighborhood. Both the number of missing measurements in the foreground and the number of unexplained observations in the background are minimized simultaneously. This provides an inherent mechanism which allows it to cope with limbs placement in the case of self-occlusion. As opposed to [12], where the use of background is also made, we do not restrict ourselves to tracking of rigid objects. When tracking, both the foreground and the local background need to be explained at the same time. This provides means for tracking assessment and dropping the tracking in case the target is occluded by a large still or moving object. Such fit criterion also copes with spurious artifacts such as shadows.

Besides tracking, our solution addresses both target detection and tracking initialization. We employ a computationaly unexpensive method first to focus attention to a region of interest before more refined model is used to validate human appearance. This allows to formulate tracking as a local mean shift optimization and decrease computational complexity of articulated tracking. As opposed to [10], our approach is inherently model-based and does not require training sequences and offline training.

Our algorithm aims at tracking of humans in low resolution videos. Processed sequences, see [19], illustrate the capability to track humans under realistic outdoor scene conditions. A person is tracked under variable illumination and with shadows present. Oclusions, slightly variable scale are handled, basic discrimination between a human and a non-human object has been illustrated. As opposed to [6, 7], we do not assume humans to be the only moving objects in the scene and provide means for discrimination between moving human and non-human. A person walking, stopping, turning around is successfully tracked. Last, our experiments illustrate that the approach is robust to typical conditions changing slightly human appearance.

As the future work, we propose to test the approach more extensively, improve the dynamics by treating both abrupt changes in motion as well as smooth transitions. Biped model pose initialization can be improved by taking coordination between limbs [9] into account. This could suppress the ambiguity caused by projection of 3D world into 2D images. Blake [22] suggests how a dynamical model can be learned.

Basic multi-person tracking may already be achieved within current framework. This would mean using our algorithm to initialize multiple ROIs and tracking these models in the consecutive frames as described in the text. The tracker does not try to fit the model to a large foreground blob that can not be explained using single model. As a result, tracking would automatically be dropped in case of occlusions and started after the tracked person appears unoccluded again. This is a desired behavior when using the motion information only as the segmentation

does not provide means to distinguish human appearance in this case. Occlusions would thus be handled in a naive yet sensible way.
The tracking through occlusions can be supported by adopting an appearance model described in [10, 12] and by utilizing the temporal information according to [11]. Initializing multiple models to a foreground blob representing a clutter of persons is another possible extension of the multi-person tracking. For instance, the approach to detection of humans within groups and verification of their hypothesized configuration described in [6] could be extended using articulated models.
In future work, we also intend to test the approach more extensively on data-sets that allow comparison with different tracking algorithms. Hence, we consider testing the proposed method on both the USF Gait Challenge data-set [20] and the CMU MoBo data-set [21].
The time performance has not been tested in the current work. Further experiments are needed to evaluate the potential of the chosen approach for real time applications. Our current implementation is in MATLAB. We consider re-implementation in C++ and time performance testing. We would like to learn in what degree our more complicated model as compared to [6] slows the system down and, on the other hand, brings better accuracy and robustness.

Acknowledgment: We would like to thank to Aleš Fexa for motion detection code and one of the image sequences. The first author is grateful to Dr. Csaba Beleznai from Advanced Computer Vision, GmbH., Vienna, Austria who supervised his summer project in person detection in summer 2005. Both authors were supported by the EC projects FP6-IST-004176 COSPAL, INTAS 04-77-7347 PRINCESS, MRTN-CT-2004-005439 VISIONTRAIN and by the Czech Ministry of Education under project 1M0567. Any opinions expressed in this paper do not necessarily reflect the views of the European Community. The Community is not liable for any use that may be made of the information contained herein.

References

1. Gavrila, D. M.: The Visual Analysis of Human Movement: A Survey. In *Computer Vision and Image Understanding*, **73**:82–98, 1999.
2. Hogg, D. C.: Model-Based vision: A Program to See a Walking Person. In *Image and Vision Computing*, **1**:5–20, 1983.
3. Rohr, K.: Incremental Recognition of Pedestrians from Image Sequences. In Proc. *IEEE Conference on Computer Vision and Pattern Recognition*, 8–13, 1993.
4. Ju, S. X., M. J. Black and Y. Yacoob: Cardboard People: A Parameterized Model of Articulated Image Motion. In Proc. *IEEE International Conference on Automatic Face and Gesture Recognition*, 38–44, 1996.
5. Zhang, R., C. Vogler and D. Metaxas: Human Gait Recognition. In Proc. *IEEE Conference on Computer Vision and Pattern Recognition Workshop*, **1**: 18, 2004.
6. Beleznai, C., B. Frühstück and H. Bischof: Human detection in groups using a fast mean shift procedure. In Proc. *IEEE International Conference on Image Processing*, 349–352, 2004.

7. Beleznai, C., B. Frühstück and H. Bischof: Tracking multiple humans using fast mean shift mode seeking. In Proc. *IEEE International Workshop on Visual Surveillance and Performance Evaluation of Tracking and Surveillance*, 25–32, 2005.
8. Lan, X. and D. P. Huttenlocher: A unified spatio-temporal articulated model for tracking. In Proc. *IEEE Conference on Computer Vision and Pattern Recognition*, **1**:722–729, 2004.
9. Lan, X. and D. P. Huttenlocher: Beyond Trees: Common-Factor Models for 2D Human Pose Recovery. In Proc. *IEEE International Conference on Computer Vision*, **1**:470–477, 2005.
10. Lim, J. and D. Kriegman: Tracking humans using prior and learned representations of shape and appearance. In Proc. *IEEE International Conference on Automatic Face and Gesture Recognition*, 869–874, 2004.
11. Howe, N. R: Silhouette Lookup for Automatic Pose Tracking. In Proc. *IEEE Conference on Computer Vision and Pattern Recognition Workshop*, **1**:15–22, 2004.
12. Collins, R., Y. Liu and M. Leordeanu: On-Line Selection of Discriminative Tracking Features. In Proc. *IEEE Transaction on Pattern Analysis and Machine Intelligence*, **27**:1631–1643, 2005.
13. Stauffer, C. and E. Grimson: Adaptive background mixture models for real-time tracking. In Proc. *IEEE Conference on Computer Vision and Pattern Recognition*, **2**:246–252, 1999.
14. Fexa, A.: Separation of individual persons in a crowd from a videosequence. Diploma Thesis, Faculty of Mathematics and Physics, Charles University in Prague, July 2004.
15. Fukunaga, K.: *Introduction to Statistical Pattern Recognition.* Academic Press, New York, 1972.
16. Fukunaga, K. and L. Hostetler: The estimation of the gradient of a density function, with applications in pattern recognition. In Proc. *IEEE Transactions on Information Theory*, **21**:21–40, 1975.
17. Cheng, Y.: Mean Shift, Mode Seeking, and Clustering. In Proc. *IEEE Transaction on Pattern Analysis and Machine Intelligence*, **17**:790–799, 1995.
18. Comaniciu, D., V. Ramesh and P. Meer: Real-Time Tracking of Non-Rigid Objects using Mean Shift. In Proc. *IEEE Conference on Computer Vision and Pattern Recognition*, **2**:142–149, 2000.
19. CMP Demo Page: http://cmp.felk.cvut.cz/demos/Tracking/TrackHumansKorc/. Center for Machine Perception, Czech Technical University in Prague, 2006.
20. Phillips, P. J., S. Sarkar, I. Robledo, P. Grother, K. W. Bowyer: The Gait Identification Challenge Problem: Data Sets and Baseline Algorithm. In Proc. *IEEE Conference on Computer Vision and Pattern Recognition*, **1**:385-388, 2002.
21. Gross, R. and J. Shi: The CMU Motion of Body (MoBo) database. Technical Report CMU-RI-TR-01-18, Robotics Institute, Carnegie Mellon University, June 2001.
22. Blake, A., B. North and M. Isard: Learning multi-class dynamics. In Proc. *Conference on Advances in Neural Information Processing Systems*, 389–395, 1999.