AN ARTICULATED MODEL WITH A KALMAN FILTER FOR REAL TIME VISUAL TRACKING

Application to the Tracking of Pedestrians with a Monocular Camera

Youssef Rouchdy
CEREMADE, Univeristé Paris Dauphine, Place du Maréchal De Lattre De Tassigny, 75775 Paris Cedex 16, France

Keywords: Visual tracking, Pedestrian tracking, Articulated models, Kalman filter, Sequential filtering.

Abstract: This work presents a method for the visual tracking of articulated targets in image sequences in real time. Each part of the target object is considered as a region of interest and tracked by a parametric transformation. Prior geometric and dynamic informations about the target are introduced with a Kalman filter to guide the evolution of the tracking process of regions. An articulated model with two areas is proposed and applied to track pedestrians in the urban image sequences.

1 INTRODUCTION

The tracking approaches can be distinguished by several criteria, for example:

1. • 2D approach without an explicit shape model,
   • 2D approach with an explicit shape model,
   • 3D approaches.
2. • tracking of primitives
   • tracking of a region of interest (ROI)
3. • deterministic approach
   • probabilistic approach
   • classification approach

The choice of the tracking approach depends on the application:

- the target objects are rigid or not,
- the camera used is monocular, stereo, fixed, mobile,
- the precision and the computing time required for the application.

The aim of this work is to develop a real-time algorithm that allows the tracking of a deformable and articulated target in an image sequence acquired by a mobile mono-camera. The principal application is the tracking of pedestrians in an urban environment and to warn the driver should a pedestrian move into the security area (see figure 1) around the vehicle.

This application is difficult to achieve for several reasons: the camera is mounted on the vehicle (so a simple background subtraction does not apply), occlusions are frequent and real time computing is required. Also, the appearence and resolution of the target object -e.g. the pedestrian- change due to deformation of clothes, changes in the posture and the motion of the camera. Figure 2 gives an idea of the variation of the resolution. In the diagrams the width and height (in pixels) of the window that contains the pedestrian in the image are plotted against the distance of the camera.

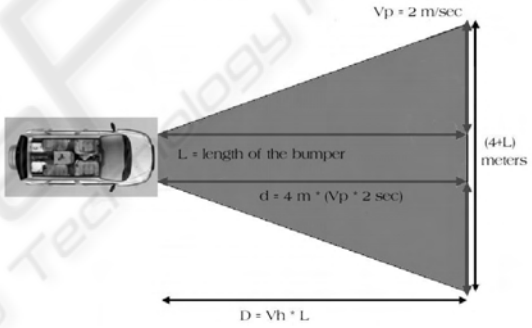

Figure 1: The security area is defined by the red (dark) region, where Vp is the pedestrian velocity and Vh is the vehicle velocity.

The most popular techniques for the estimation of motion are based on the parametric model (Bergen et al., 1992), which is adapted to real time tracking. These techniques model the motion of a ROI in an image for example by the affine (6 DOF) or the homographic (8 DOF) transformation. Problems occur when the motion model that is used does

Rouchdy Y. (2008).
AN ARTICULATED MODEL WITH A KALMAN FILTER FOR REAL TIME VISUAL TRACKING - Application to the Tracking of Pedestrians with a Monocular Camera.
In *Proceedings of the Third International Conference on Computer Vision Theory and Applications*, pages 686-693
DOI: 10.5220/0001089706860693

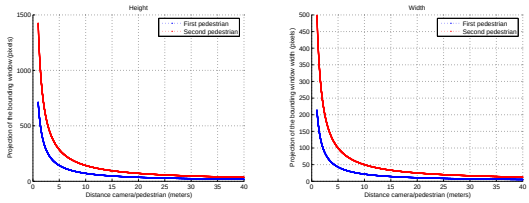

Figure 2: Dimensions of pedestrian in the image (in pixels) according to the variation of the distance between the camera and a pedestrian.

not describe well the motion of the ROI. In (Weiss and Adelson, 1996), the motion is segmented into independent multiple motion areas. This method is problematic when some of the areas do not contain a high enough number of pixels or when they contain noise. Others have introduced a constraint. The target segments are connected to each other by joints. In (Murray et al., 1994; Bregler et al., 2004), the twist and product exponential map are used to introduce constraints. In (Gavrila and Davis, 1995; Kakadiaris and Metaxas, 1996), the human body is modeled by rigid segments which are connected by joints. See (Gavrila, 1999) for a comprehensive bibliography of approaches used to analyse human motion, and see (Zhang et al., 2006) for an analysis of motion with an articulated model.

In this work, we model the target by rigid segments which are connected by a priori informations representing the joints. The constraints are introduced as a priori informations using a Kalman filter. This permits a higher reach connection between the multiple areas without increasing the DOF of the system. The contribution of each segment and of each joint is regulated by the covariance matrices of the Kalman Filter. So the Kalman filter proposed introduces *a priori* geometric (e.g. the connection between the multiple areas) and dynamic (constant velocity) informations about the target. Furthermore, the Kalman filter smooths the trajectory of the target. In the next section, a different approach to model the dynamic evolution in the context of visual tracking with sequential filtering is presented. Subsequently, we will focus on the Kalman filtering. In section 3, an articulated model based on a Kalman filtering is proposed. In section 4, an articulated model is proposed to track a pedestrian in image sequences and experimental results are given.

2 SEQUENTIAL FILTERING AND TRACKING IN IMAGE SEQUENCES

In this section, an introduction of the use of the sequential filtering in the context of the visual tracking is given. The measurements acquired up to frame t are denoted Y_t and X_t represents the configuration of the target objects at the time t. The process $\{X_t;\, t \in \mathbb{N}\}$ is modeled as a Markov process of initial distribution $p(X_0)$ and transition equation $p(X_t|X_{t-1})$. The observations $\{Y_t;\, t \in \mathbb{N}\}$ are assumed to be conditionally independent given the process $\{X_t;\, t \in \mathbb{N}\}$ and of marginal distribution $p(Y_t|X_t)$. The principle of sequential filtering is to apply Bayes's theorem at each time-step, obtaining a posteriori $p(X_t|Y_t)$ based on all available information:

$$p(X_t|Y_t) = \frac{p(Y_t|X_t)p(X_t|X_{t-1})}{p(Y_t)} \tag{1}$$

where we can write $p(Y_t|X_t)$ instead of $p(Y_t|X_t, Y_{t-1})$ due to the conditional independence assumption. According to custom in filtering theory, a model for the expected motion between time-steps is adopted. This takes the form of a conditional probability distribution $p(X_t|X_{t-1})$ termed the dynamics. Using the dynamics equation, (1) becomes

$$p(X_t|Y_t) = \frac{p(Y_t|X_t)\int p(X_t|X_{t-1})p(X_{t-1}|Y_{t-1})\mathrm{d}X_{t-1}}{p(Y_t)} \tag{2}$$

It is assumed that the predicted values of the states and the observations, X_t and Y_t, respectively, evolve in time according to:

$$X_t = f_t(X_{t-1}, V_t), \tag{3}$$
$$Y_t = g_t(X_t, W_t). \tag{4}$$

where f_t and g_t are the state and the observation functions, respectively, which are supposed to be known. The state noise V_t and the measurement noise W_t have known distributions.

In visual tracking the choice of the dynamical model $p(X_t|X_{t-1})$ depends on the type of the images, the *a priori* information available and the application. Typically, the elasticity model is used to track the elastic structure (Rouchdy et al., 2007) and the Navier-Stockes model is used to track a fluid structure (Cuzol et al., 2007). When a good *a priori* information is available the prediction can be introduced by learning (Blake et al., 1999). The most popular dynamical model used is autoregressive (Black and Fleet, 1999; Perez et al., 2002) and corresponds in the

first order to the model of constant velocity, which is the model adopted in this work.

The choice of the observations depends on the application and the image and can be subjective in some cases. The cues usually used are edge information (Blake and Isard, 1998) and color distributions (Perez et al., 2002). There exists also a model based on motion and appearance (Sidenbladh and Black, 2003). In (Sidenbladh and Black, 2003), several cues are combined to make the model robust to a change of appearance.

For a nonlinear system (e.g. the function f_t or g_t in equations (3) is nonlinear) the probability $p(X_t|Y_t)$ is approximated by a Monte Carlo (MC) method. Unfortunately, the classical sampling for the MC method is guaranteed to fail as time increases. To deal with this problem a step of selection is added -(Gordon, 1993) gives the first operationally effective method. Theoritical convergence results of this algorithm are given in (Del Moral, 1997). A good reference and coherent treatement of these techniques including convergence results and applications to visual tracking are presented in (Doucet et al., 2002).

When the observation density $p(Y_t|X_t)$ is assumed to be Gaussian and the dynamics are assumed linear with additive Gaussian noise the solution is obtained analytically and this method corresponds to the Kalman filter. In this case the dynamical system is written as:

$$X_t = A_t X_{t-1} + B_t V_t, \quad (5)$$

$$Y_t = C_t X_t + D_t W_t, \quad (6)$$

where A_t, B_t, C_t and D_t are matrices and V_{t-1}, W_{t-1} are vectors of i.i.d standard normal variants. The state noise V_t and the measurement noise W_t are supposed to be Gaussian and independent with the matrices covariances Q_t and R_t, respectively. In this case, $p(X_t|X_{t-1})$, $p(Y_t|X_t)$ and $p(X_t|Y_t)$ have a Gaussian distribution with the covariance matrices Q_t, R_t and Γ_t, respectively. Where the covariance matrix Γ_t and the estimation of the vector state X_t are computed recursively with the Kalman filter, the Kalman recursion is given in section 3.5 and documented in (Kalman and Bucy, 1961).

3 ARTICULATED MODEL BASED ON KALMAN FILTER

3.1 Motion Estimation

The image is a projection of 3D points of the space on an image plane. Let a rectangle that moves in the 3D space be such that the deformations in the image plane are described by a rigid transformation. Let I_1 be the image of this object at the time t_1 and let I_{t_n} be the image at t_n. In (Faugeras et al., 2001), it is shown that the points on the rectangle of the two frames are related by a *homographic* transformation and that they are defined by eight parameters. Subsequently, it is supposed that the deformation of a target is obtained by a homographic transformation. Otherwise, the object is approximated by a set of rigid links. We restrict ourselves to this type of motion to reduce the complexity.

3.2 Modelisation and Predictions

Let $\{R_r^l\}_{l=1}^N$ be a set of supposed rigid areas of an articulated target, let c_l be the coordinates of the barycenter of the region R_r^l, and let s_l be the surface of R_r^l. The elements of the set $\{R_r^l\}_{l=N}^1$ are correlated by their barycenters with the relations

$$\begin{array}{c} \psi_1(c_1,\cdots,c_N,s_1,\cdots,s_N) = d_1;\ \ldots\ ; \\ \psi_N(c_1,\cdots,c_N,s_1,\cdots,s_N) = d_m \end{array} \quad (7)$$

where m is the number of the constraint functions ψ. These constraints are introduced as a priori informations with the Kalman filter. An example of the tracking of a pedestrian in an image sequence using two correlated areas is given in section 4.

The constraintes (7) are supposed to be linear. If this is not the case, they can be linearized. The constraints are introduced into the dynamic system of the Kalman filter. The constraints (7) are introduced in the filter with a function $g_t = (g_t^1,\cdots,g_t^m)$. At the time t, the state vector X_t is defined by

$$X_t = \left(c_t^1,\cdots,c_t^n,v_1^t,\cdots,v_t^N,s_t^1,\cdots,s_t^N,g_t^1,\cdots,g_t^m\right),$$

and follows the state equations:

$$X_t = AX_{t-1} + BV_{t-1}, \quad (8)$$

$$Y_t = CX_t + DW_{t-1}, \quad (9)$$

where A, B, C and D are fixed matrices and V_{t-1}, W_{t-1} are vectors of i.i.d standard normal variants. The matrix A introduces dynamic, e.g. constant velocity, and geometric, e.g. the correlation between the areas, a priori informations about the target.

3.3 Measurement

For each set R_r^l, the estimation of the parameters of the transformation model is achieved by the minimization problem:

$$\lambda_t^l = \arg\min_{\lambda\in\mathbb{R}^\alpha} \sum_{p\in R_r^l} \left| I_t(\phi_t^l(\lambda,p)) - I_r(p) \right|^2 \qquad (10)$$

where I_r is the reference image, I_t the current image, ϕ_t^l are the parametric transformations determined by the parameters λ_t^l for each $l \in \{1,\cdots,N\}$ and where α is the parameters number. The measurement vector is computed from the transformations ϕ_t^l and defined by

$$Y_t^{\text{mes}} = \left(z_t^1,\cdots,z_t^n,v_1^t,\cdots,v_t^N,s_t^1,\cdots,s_t^N,g_t^1,\cdots,g_t^m\right),$$

where z_t^l and s_t^l are the barycenters and the surfaces, respectively, of the areas $\phi_t^l(\lambda_t^l,R_r^l)$. From z_t^l and s_t^l we compute the quantity g_t^i for each $(i,l) \in \{1,\cdots,m\} \times \{1,\cdots,N\}$. The image motion of the point $z_t^{i,l}$ from time $t-1$ to time t is :

$$v_t^l = \frac{z_t^l - z_{t-1}^l}{\Delta t}.$$

The minimization of the problem (10) is achieved by the ESM algorithm, see (Malis, 2004; Benhimane and Malis, 2004).

3.4 Initialization

Using the first reference image, the user segments the target manually into a set of areas. The surface of the areas, the distance between the barycenters of the areas and other geometric characteristics are computed from the initial set of areas to initialize the Kalman filter.

3.5 Filtering

This step allow us to introduce a goodness of fit criterion between a reference template and possible candidates in the current image. If the additional noise v_t is supposed to be Gaussian, then the observation density $p(Y_t|X_t)$, associated to the prediction and measurement described in the previous section, has a Gaussian distribution with a Covariance matrix error R. The density $p(X_t|Y_t)$ has also a Gaussian distribution with a covariance matrix error Q. So the estimation of the state is computed by a Kalman filter with the relations:

- initialization
  - X_0, P_0, R and Q are given
- prediction
  - $\bar{X_t} = AX_{t-1}$
  - $\bar{Y_t} = C\bar{X_t}$
  - $\bar{P_t} = AP_{t-1}A^* + Q$
- filtering
  - $K_t = \bar{P_t}C(C\bar{P_t}\cdot C + R)^{-1}$,
  - $X_t = \hat{X}_t + K_t(Y_t^{\text{mes}} - \bar{Y_t})$,
  - $P_t = (1 - K_tC)\,\bar{P_t}$.

K_t is called the gain. The difference $Y_t^{\text{mes}} - \bar{Y_t}$ is called the measurement innovation. The innovation reflects the discrepancy between the predicted measurement $C\bar{X_t}$ and the actual measurement Z_t^{mes}. Let

$$Q = \begin{pmatrix} Q^1 & 0 & 0 \\ 0 & Q^2 & 0 \\ 0 & 0 & Q^3 \end{pmatrix}, \quad R = \begin{pmatrix} R^1 & 0 & 0 \\ 0 & R^2 & 0 \\ 0 & 0 & R^3 \end{pmatrix}$$

The covariance matrix (Q^1,R^1), (Q^2,R^2) and (Q^3,R^3) are associated to the barycenters, the surface and the constraints, respectively.

4 APPLICATION TO TRACK A PERSON

4.1 Measurements and Predictions

To track a pedestrian in an image sequencee, we track the head and the torse which are supposed to be connected: the head stays close to the torse. Two areas F_1 and F_2 are used, one coresponding to the head and other to the torse.

4.1.1 Method 1

The vector state X_t is defined by the centers of the windows F_1 and F_2 and by their velocity v_t^1 and v_t^2, respectively:

$$X_t = (c_t^1, v_t^1, c_t^2, v_t^2)$$

The prediction matrices are defined by

$$A = \begin{pmatrix} 1 & 0 & \Delta t & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & \Delta t & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 \\ 1 & 0 & \Delta t & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 & 0 & \Delta t \\ 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 \end{pmatrix},$$

$$C = \begin{pmatrix} 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 \end{pmatrix}$$

These predicions introduce the *a priori* information that the first component of the window center c_t^1 is equal to the first component of c_t^2. Figure 3 gives the variations of the predicted value of the first component of one window when the first component of the other window is fixed to be zero. The amplitude variation of the predicted value of the first component depends on the variance of the considered gaussian noise.

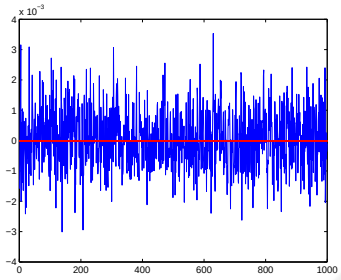

Figure 3: Variation of the predicted value of the first compenents of one window when the first component of the second window is fixed at zero.

4.1.2 Method 2

Here, the vector state is defined by only one window F_1 :

$$X_t = (c_t^1, v_t, d_t),$$

where d_t is a vecor of $\mathbb{R}^2$. The prediction matrices are defined by

$$A = \begin{pmatrix} 1 & 0 & \Delta t & 0 & 0 & 0 \\ 0 & 1 & 0 & \Delta t & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 \end{pmatrix},$$

$$C = \begin{pmatrix} 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 \\ 1 & 0 & 0 & 0 & 1 & 0 \\ 0 & 1 & 0 & 0 & 0 & 1 \end{pmatrix}$$

We remark that if

$$d_t = c_t^2 - c_t^1, \tag{11}$$

the observations Y_t correspond to the barycenters of the zones targeted. So the initial values of the vector d_t permit to introduce a constraint on the evolution of the two windows in the Kalman filter. The coordinates of the barycenter of the window F_2 are deduced from the estimation of d_t and c_t^1 by the relation

$$c_t^2 = d_t + c_t^1.$$

The initial value of the vector d_t is computed from the initial barycenters of the windows F_1 and F_2.

4.2 Initialization of the Optimization Algorithm

The user chooses the initial size of the windows F_1 and F_2 manually The window size is updated with the estimated values of d_t with the relations:

$$L_t^{l,x} = L_{t-1}^{l,x} \cdot \frac{||d_t||}{||d_{t-1}||}, \quad L_t^{l,y} = L_{t-1}^{l,y} \cdot \frac{||d_t||}{||d_{t-1}||}$$

where $L_t^{l,x}$ is the length of the window l and $L_t^{l,y}$ is the width of the window F_l at the time t. The size of the windows and their barycenters are used to initialize the values of the transformations for the next step of tracking. The initialization of the transformation for the minimization routine ESM is achieved by using the results of filtering to construct new homographic transformations defined by

$$H_t^{l,0} = \begin{pmatrix} 1 & 0 & \min\left(c_t^{l,x} - \frac{L_t^{l,x}}{2}, c_t^{l,x} + \frac{L_t^{l,x}}{2}\right) \\ 0 & 1 & \min\left(c_t^{l,y} + \frac{L_t^{l,y}}{2}, c_t^{l,y} - \frac{L_t^{l,y}}{2}\right) \\ 0 & 0 & 1 \end{pmatrix}$$

We note that, these homographies are computed from the estimation at the previous iteration of the window center. The update of the transformation by the transformation $H_t^{l,0}$ gives more stable results than when the initialization is performed with the transformations that we measured directely at the previous step. The necessity to update the transformation is due to the non-rigid motion of the target and changes of their appearence.

4.3 Change of Appearance and Resolution

Where the change of appearence and resolution becomes very large due to the deformation of the

clothes, changes in the posture of the pedestrian and the mobility of the camera, the update of the transformations proposed in the previous section is insufficient. It is necessary to update the reference template which is composed of the set of the targeted areas. In this work, we have used the estimation of d_t to update the reference template. Indeed, when the condition

$$\Big|\,||d_t|| - ||d_{t-1}||\,\Big| < Th_v$$

is not satisfyed, the set of areas is update from the set of areas obtained at the previous iteration. This criterion can be combined with another similarity measurement that compars the reference and current templates for example by using the sum-of-squared-differences (SSD). In (Arnaud et al., 2004), the eigenvalues of the covariance error in tke Kalman filter is used to define a threshold.

Intuitively, a more pertinent strategy to update the reference template consists to accumulate the reference templates (Morency et al., 2003). The problem with such an approach is to define a criterion of similarity that can be applied to the reference data and also the computing time can be expensive. In (Wu and Huang, 2001), a particle filter is used to evolve the reference template by a dynamical model. To deal with the change of apparence, Headvig et al. (Sidenbladh and Black, 2002) are use a learning approach based on the cues edge, ridge and motion. The cues are combined with a bayesian model.

4.4 Experiments

4.4.1 Data

We tested the proposed articulated models on data from an urban traffic environment. The two video sequences used were recorded by Renault in the context of the LOVe project[1] by a SMAL camera from CYPRESS company. The following table gives the main characteristics of the used camera. Since our model is adapted to a monocular camera, only one of the two sets of images obtained by the stereo camera were used.

Part Number	Features	Frame Rate	Optical Format	Pixel Size	Resolution	Supply Voltage
XAECK10005	Stereo	60 fps	1/3 inch	8 um x 8 um	640 x 480	110/220V

The first image sequence was aquired from an immobile vehicle. The trajectory of the pedestrian was perpendicular to the road. The second image sequence is acquired by a camera mounted on a moving vehicle and shows pedestrians crossing the road.

[1] http://www.love.univ-bpclermont.fr/

4.4.2 Results and Discussion

In the following section, some experimental results obtained with the urban image sequences presented in the last section are given to evaluate the articulated model 4.1.1. We compare it to the results obtained when only one area is trached with the ESM algorithm. Figure 4 shows the result obtained by the ESM algorithm: the tracker has lost the target at the second image due to a non-rigid motion and due to the background present in the initial area (pixels not belonging to the pedestrian). Figure 5 shows that the articulated model has succeeded in tracking the target. These results were obtained without updating the reference template.

Figure 6 shows the results of the tracking. The pedestrian was correctly tracked until he made a 90 degree turn in relation to the camera in image $t_8 = 2.0630\,s$. The update of the reference template using the distance correlation between the barycenters was insufficient to deal with the large change of appearance in frame $t_8 = 2.0630\,s$ when the pedestrian changed his posture completely. It will be interesting to combine the proposed update method to another update strategy based on the measurement of similarity, see section 4.3.

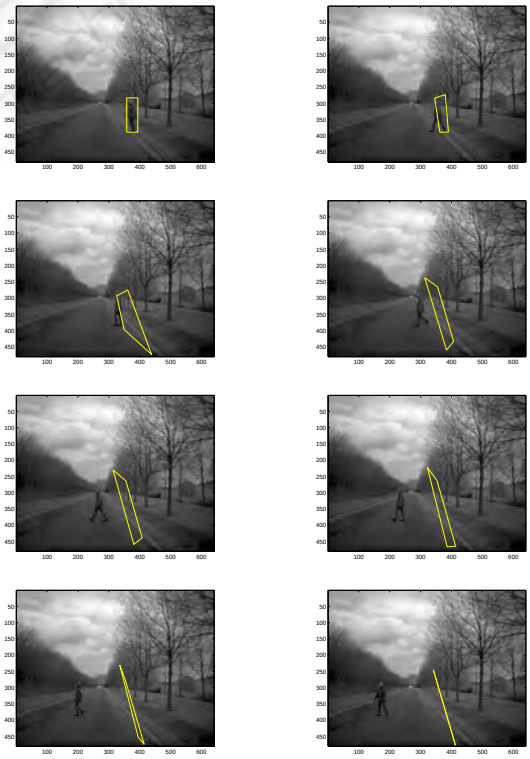

Figure 4: Results of the ESM tracking algorithm.

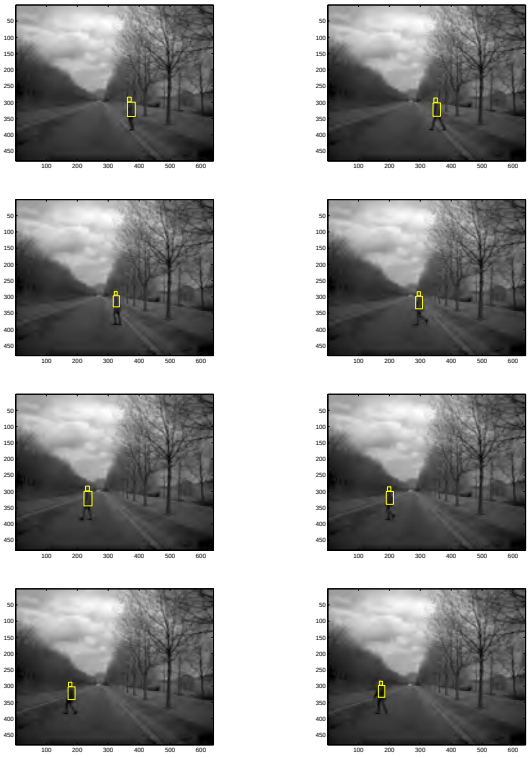

Figure 5: Results obtained with the first articulated model.

Figure 6: Tracking with the first articulated model applied to an an urban image sequence recorded with 30 images/seconde.

5 CONCLUSIONS

In this work, we propose a method for the tracking of an articulated target with a monocular camera using the Kalman filter. The advantages of this approach are: the joints are introduced as *a priori* information and not as constraints -e.g. the a priori informations can be left unsatisfied if the measurements (extracted from the images) are more pertinent-, a real time implementation is possible, it is easy to implement, it is stable and it smooths the trajectory of the target. For a large number of areas the algorithm can be easily parallelized: the measurement is computed separately for each area. However, this model is sensitive to large changes in appearence. Further work is necessary to tackle this problem, some directions are proposed in section 4.3.

ACKNOWLEDGEMENTS

We thank A. Doucet, E. Malis, D. M. Pierre and P. Rives for stimulating and useful discussions. We are very grateful to F. Solanet for providing the image sequences, the two video sequences used were recorded by Renault in the context of the LOVe project[2]

REFERENCES

Arnaud, E., Memin, E., and Cernuschi-Frias, B. (2004). Conditional filters for image sequence based tracking - application to point tracker. In *IEEE trans. On Im. Proc*.

Benhimane, S. and Malis, E. (2004). Real-time image-based tracking of planes using efficient second-order minimization. In *In IEEE/RSJ International Conference on Intelligent Robots Systems, Sendai, Japan, October 2004*.

Bergen, J. R., Anandan, P., Hanna, K. J., and Hingorani, R. (1992). Hierarchical model-based motion estimation. In *ECCV '92: Proceedings of the Second European Conference on Computer Vision*, pages 237–252, London, UK. (Springer-Verlag).

Black, M. J. and Fleet, D. J. (1999). Probabilistic detection and tracking of motion discontinuities. In *ICCV (1)*, pages 551–558.

[2]http://www.love.univ-bpclermont.fr/

Blake, A. and Isard, M. (1998). *Active Contours*. Springer, Berlin Heidelberg New York.

Blake, A., North, B., and Isard, M. (1999). Learning multi-class dynamics. *Advances in Neural Information Processing Systems*, 11:389–395.

Bregler, C., Malik, J., and Pullen, K. (2004). Twist based acquisition and tracking of animal and human kinematics. *Int. J. Comput. Vision*, 56(3):179–194.

Cuzol, A., Hellier, P., and Mmin, E. (2007). A low dimensional fluid motion estimator. *Int. Journ. on Computer Vision*.

Del Moral, P. (1997). Nonlinear filtering: interacting particle resolution. *C. R. Acad. Sci. Paris Sér. I Math.*, 325(6):653–658.

Doucet, A., de Freitas, N., and Gordon, N., editors (2002). *Sequential Monte Carlo Methods in Practice*. Statistics for Engineering and Information Science. Springer-Verlag, New York Berlin Heidelberg.

Faugeras, O., Luong, Q.-T., and Papadopoulou, T. (2001). *The Geometry of Multiple Images: The Laws That Govern The Formation of Images of A Scene and Some of Their Applications*. MIT Press, Cambridge, MA, USA.

Gavrila, D. M. (1999). The visual analysis of human movement: A survey. *Computer Vision and Image Understanding: CVIU*, 73(1):82–98.

Gavrila, D. M. and Davis, L. S. (1995). Towards 3d model-based tracking and recognition of human movement. In *Proc. of the IEEE International Workshop on Face and Gesture Recognition*, pages 272–277, Zurich, Switzerland.

Gordon, N. (1993). *Bayesian methods for tracking*. PhD thesis, University of London.

Kakadiaris, I. A. and Metaxas, D. (1996). Model-based estimation of 3D human motion with occlusion based on active multi-viewpoint selection. In *Proceedings of the 1996 Conference on Computer Vision and Pattern Recognition (CVPR '96)*, page 81, Washington, DC, USA.

Kalman, R. E. and Bucy, R. S. (1961). New results in linear filtering and prediction theory. *Trans. ASME Ser. D. J. Basic Engrg.*, 83:95–108.

Malis, E. (April 2004). Improving vision-based control using efficient secondorder minimization techniques. In *ICRA'04, New Orleans*.

Morency, L., Rahimi, A., and Darrell, T. (2003). Adaptive view-based appearance models. In *Proc. IEEE Conf. on Comp. Vision and Pattern Recogn.*, pages 803–810.

Murray, R. M., Sastry, S. S., and Zexiang, L. (1994). *A Mathematical Introduction to Robotic Manipulation*. CRC Press, Inc., Boca Raton, FL, USA.

Perez, P., Hue, C., Vermaak, J., and Gangnet, M. (2002). Color-based probabilistic tracking. In *ECCV*, number 2350 in LNCS, pages 661–675.

Rouchdy, Y., Pousin, J., Schaerer, J., and Clarysse, P. (2007). A nonlinear elastic deformable template for soft structure segmentation. Application to the heart segmentation in MRI. *Inverse Problems*, 23:1017–1035.

Sidenbladh, H. and Black, M. (2003). Learning the statistics of people in images and video. *Int. Journ. on Computer Vision*, 54(1-3):183–209.

Sidenbladh, H. and Black, M. J. (2002). Learning the statistics of people in images and video. *Int. Journal of Computer Vision*, 54.

Weiss, Y. and Adelson, E. H. (1996). A unified mixture framework for motion segmentation: Incorporating spatial coherence and estimating the number of models. In *Proceedings of the 1996 Conference on Computer Vision and Pattern Recognition (CVPR '96)*, page 321, Washington, DC, USA.

Wu, Y. and Huang, T. (2001). A co-inference approach to robust visual tracking. In *Proc. IEEE Conf. on Comp. Vision*, pages 26–33.

Zhang, X., Liu, Y., and Huang, T. S. (2006). Motion analysis of articulated objects from monocular images. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 28(4):625–636.