MULTI-OBJECT TRACKING BASED ON SOFT ASSIGNMENT OF DETECTION RESPONSES

Sami Huttunen and Janne Heikkilä
Machine Vision Group, Department of Electrical and Information Engineering, University of Oulu
P.O. Box 4500, FIN-90014, Oulu, Finland

Keywords: Multi-object tracking, Kalman filter, Data association.

Abstract: We introduce a new detection-based method that is able to track multiple objects from a single camera. The method is built upon an approach that combines Kalman filtering and the Expectation Maximization (EM) algorithm. The benefit of this approach is that soft assignment of the detections to corresponding objects can be performed automatically using their a posteriori probabilities. This is a general approach for detection-based multi-object tracking, and there are various ways to detect the objects. In this paper, we demonstrate the applicability of the approach for tracking multiple pedestrians and faces using a basic cascade detector.

1 INTRODUCTION

Tracking is an essential part of current computer vision applications. Especially many modern visual surveillance and human computer interaction systems rely on reliable multi-object tracking. In principle tracking over time involves matching objects in consecutive frames using features such as points, lines or blobs. One difficulty in the application of multi-object tracking involves the problem of getting reliable observations and associating them with the appropriate objects. The association process would be simple if there were only one measurement for each object, but in order to get reliable information about the position, the number of observations has to be larger.

One approach to address the limitation of getting reliable measurements is to combine tracking with detection. Previously detection has been used to initialize new objects or detectors are applied only to selected frames. Between the detections tracking has been carried out, for example, by a method based on color or texture features. Due to increased computational power, it is nowadays possible to use detectors for every frame instead of just some frames. Therefore the approaches using detector responses directly as observations for tracking are gaining more and more attention.

Since a typical object detector (Viola and Jones, 2001) is insensitive to small changes in translation and scale, multiple detection responses will usually occur around each object in a scanned image, and typically it often makes sense to return one final detection per object. To obtain this kind of result, it is therefore useful to post process the detected sub-windows in order to combine overlapping detections into a single detection. However, it is unclear how fusing multiple overlapping detections to yield the final object detections should be performed. Unfortunately it is also difficult to distinguish the false positives using the post processing, and in some cases a detector can also output inaccurate responses. The output of the detector can be thought of as a series of noisy measurements, and therefore our approach uses the original detector responses as a set of measurements and assigns them to the objects currently tracked. In that way we can leave out the problematic post processing entirely and at the same time get a number of measurements for tracking.

We present a novel method, which utilizes soft assignment to associate the detection responses to the objects tracked. Due to soft assignment it is able to cope with inaccurate responses and inter-object occlusions. The method includes a component which combines the Kalman filtering algorithm (Kalman, 1960) and the expectation maximization (EM) algorithm (Dempster et al., 1977) to estimate the parameters of the objects tracked and to assign the measurements softly. One of the benefits of this approach is also that neither iterations nor long measurement history are needed. The basic idea of the Kalman-EM algorithm was originally presented by Hannuksela et al. (2007) and later it was used for multi-object tracking

Huttunen S. and Heikkilä J. (2010).
MULTI-OBJECT TRACKING BASED ON SOFT ASSIGNMENT OF DETECTION RESPONSES.
In *Proceedings of the International Conference on Computer Vision Theory and Applications*, pages 296-301
DOI: 10.5220/0002820602960301

for the first time (Huttunen and Heikkilä, 2008). Here we extend the previous work (Huttunen and Heikkilä, 2008), and propose a new method for tracking multiple objects from image sequences using detector responses as measurements. This makes it possible to initiate new objects, as well as terminate tracks that are no longer valid. In addition, the proposed method is able to adjust the scale of the objects tracked. The method presented by Huttunen and Heikkilä (2008) does not provide any of these capabilities.

One of the benefits of detector based tracking is that it enables to track only the objects of interesting category, for instance, humans or their faces. Another advantage is that we are not required to use a static camera as is the case with multi-object tracking methods relying on background subtraction.

Related Work. The Kalman filter (Kalman, 1960) is widely used in the context of tracking with noisy measurements and data association. In multiple hypothesis tracking (MHT) algorithm (Reid, 1979), target states are estimated from data-association hypotheses using the Kalman filter. For each measurement, probabilities are calculated for hypotheses that the measurement came either from previously known targets or from a new target. The MHT algorithm is computationally exponential both in memory and time. Later Joo and Chellappa (2007) have proposed an improved algorithm based on MHT. Another classical approach for data association is Joint Probabilistic Data Association Filter (JPDAF) (Fortmann et al., 1983), in which joint posterior association probabilities are computed for multiple targets or multiple discrete interfering sources in Poisson clutter. The major limitation of the JPDAF algorithm is its inability to initialize new objects entering the scene and to deal with objects exiting the scene.

There exists a wide variety of detection methods. At one end of the spectrum are part-based detectors (Mikolajczyk et al., 2004; Wu and Nevatia, 2005), which represent an object as an assembly of distinct parts. At the other end are detection methods (Gavrila, 2000; Viola and Jones, 2001) that try to find a specific object as a whole.

One way of integrating detection and tracking is to link detection responses in consecutive frames. Huang et al. (2008) present a detection-based three-level hierarchical association approach. The more recent work by Singh et al. (2008) introduces a two-stage multi-object tracking approach using a pedestrian detector and association of track segments. Leibe et al. (2007) have introduced an approach which considers object detection and space-time trajectory estimation as a coupled optimization problem.

When comparing the aforementioned methods (Huang et al., 2008; Leibe et al., 2007; Singh et al., 2008) with our work, the biggest difference is that we associate the detector responses to objects without utilizing trajectory history. This means we do not need iterations or long measurement history in order to track the objects. In addition, the tracking algorithm presented in this paper is not dependent on a specific object detector or object category. Later in Section 3 we demonstrate the applicability of the approach for tracking multiple pedestrians and faces using a basic cascade detector.

The rest of the paper is organized as follows. Section 2 describes the tracking algorithm in detail, and experimental results are reported in Section 3. Finally, Section 4 concludes the paper.

2 TRACKING ALGORITHM

Our multi-object tracking method is based on soft assignment of detector responses. For every frame, first an object detector is applied and the resulting output is passed to the actual tracking algorithm. If there are responses that are not assigned to any object, possibly new objects are initialized. On the other hand, if an object does not get any measurements it might be necessary to end tracking it.

2.1 Object Detection

The object detector used in this work is built on the cascade system proposed by Viola and Jones (2001) and improved by Lienhart and Maydt (2002). Like any other detectors based on a binary object/non-object classifier, the detector scans the image with a detection window at all positions and scales, running the classifier in each window and yielding multiple overlapping detections.

It is worth noting that the tracking algorithm presented in this paper is not bound up with any specific detector. The only requirement the object detector used has to meet is that it must be able to output the bounding boxes of the objects in a single frame.

2.2 Kalman-EM Algorithm

In order to track multiple objects we are using a method, which utilizes soft assignment to associate the detection responses to the corresponding objects. The method embeds the EM algorithm into the Kalman filtering algorithm to estimate the parameters of the objects tracked.

System model. We assume that each object $j = 1, \dots, M$ is represented by a vector $\mathbf{x}_j = [s_j, u_j, t_j, v_j]^T$ of four state variables which contain information about the object's position (s_j, t_j) and velocity (u_j, v_j) in the X and Y directions respectively. It should be noted that the dimensions of the object are not included in the state model and are therefore updated separately as shown later. Since the object tracked is represented as a point, then only a translational model can be used, and the state-space model of the object j can therefore be formulated as

$$\mathbf{x}_j(k) = \mathbf{\Phi}\mathbf{x}_j(k-1) + \mathbf{\Gamma}\varepsilon_j(k-1), \qquad (1)$$

where $\mathbf{x}_j(k)$ denotes the state of the object j at time step k, and $\mathbf{\Phi} = \left[\begin{smallmatrix} 1 & 1 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 1 \\ 0 & 0 & 0 & 1 \end{smallmatrix}\right]$, $\mathbf{\Gamma} = \left[\begin{smallmatrix} 0.5 & 0 \\ 1 & 0 \\ 0 & 0.5 \\ 0 & 1 \end{smallmatrix}\right]$ are the the state transition and disturbance matrices respectively. Finally, $\varepsilon_j(k)$ is the process noise term, which is assumed to be zero-mean white Gaussian noise with a 2×2 covariance matrix $\mathbf{Q}_j = \sigma_j^2\mathbf{I}$.

Measurement Model. Let us denote a detection response by $\mathbf{r_i} = [r_i^x, r_i^y, r_i^w, r_i^h]^T$, in which (r_i^x, r_i^y) are the pixel coordinates of the center of the bounding box, and (r_i^w, r_i^h) are the dimensions. The observation i of the position $\mathbf{l}_i = [r_i^x, r_i^y]^T$ is assumed to follow the measurement model

$$\mathbf{l}_i(k) = \mathbf{H}\sum\nolimits_{j=1}^{M} \lambda_{i,j}\mathbf{x}_j(k) + \eta_i(k), \qquad (2)$$

where M is the number of objects being tracked, $\mathbf{H} = \left[\begin{smallmatrix} 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 \end{smallmatrix}\right]$ is the measurement matrix, and $\lambda_{i,j}$ is a hidden binary assignment variable which indicates the object that generated the measurement. $\eta_i(k)$ is the observation noise, which is assumed to obey zero-mean Gaussian distribution with a covariance matrix $\mathbf{R}$. In addition, the process noise $\varepsilon_j(k)$ and the observation noise $\eta_i(k)$ are assumed to be independent of each other.

Soft Assignment. It follows from equations (1) and (2) that the observations $\{\mathbf{l}_i\}_{i=1}^N$ form a set of 2-D points that follow a dynamically evolving Gaussian mixture model where the mean values of the components change during the course of time. Were the values of the binary assignment variables $\lambda_{i,j}$ known beforehand, it would be possible to accomplish state estimation simply by using M ordinary Kalman filters independently. Since this information is not available in general, we are compelled to estimate the assignments as well. For this purpose, we use an algorithm (Hannuksela et al., 2007; Huttunen and Heikkilä, 2008), which efficiently combines Kalman filtering and the EM algorithm. It has been shown experimentally that the algorithm converges to the mean values of the mixture components. A detailed description of the above-mentioned algorithm is given in Algorithm 1.

When applying the algorithm, the basic assumption is that there are M distributions corresponding to different objects, and the location measurements $\{\mathbf{l}_i\}_{i=1}^N$ are originating from them. Having the previous estimate of distribution parameters, we can evaluate a posteriori probabilities of the measurements to obtain the soft assignments $w_{i,j} \in [0, 1]$. In this work, the predicted estimates of $\mathbf{x}_j(k)$ and $\mathbf{P}_j(k)$ in conjunction with a priori probabilities $\pi_j(k)$ of associating observations to the objects are used to compute soft assignments $w_{i,j}$ using the Bayesian formulation. It can be seen that this part corresponds to the "E step" of the EM algorithm.

Soft assignments are then used in computation of the Kalman gains $\mathbf{K}_j(k)$ which are needed to get the filtered estimates of $\mathbf{x}_j(k)$. Traditionally one way of thinking about the weighting by $\mathbf{K}_j(k)$ is that as the measurement error covariance $\mathbf{R}$ approaches zero, the actual measurement is trusted more, while the predicted measurement is trusted less. On the other hand, as the a priori estimate error covariance $\mathbf{P}_j^-(k)$ approaches zero the actual measurement will have smaller effect on the estimate, while the predicted measurement is given more weight. However, in our approach we have multiple measurements instead of a single measurement and therefore we have to estimate the uncertainties for each of them. The principle is that the covariance matrix $\mathbf{R}_{i,j}$, which represents the uncertainty of one measurement, is inversely proportional to $w_{i,j}$, and directly proportional to $\mathbf{R}$. The small constant δ has the effect that the measurements not belonging to any objects, in other words, outliers, get very small weights and large uncertainty values and are therefore discarded. This part corresponds to the "M step" of the EM algorithm.

Implementation Details. In the actual implementation of the filter, the measurement noise covariance is usually measured prior to operation of the filter. Estimating the measurement error covariance is possible because we should usually be able to take some offline sample measurements in order to determine the variance of the measurement noise.

When using detector responses as measurements the system noise characteristics are not exactly known. If too much emphasis were given to the dynamical model, the estimation would ignore the information from new measurements. It is even possible that this can lead to filtering instability and divergence. This can happen, for example, when an object is at the same position for a very long time. In order

Algorithm 1: The combined Kalman filter and EM algorithm to estimate state using the given model.

Step 1. Predict estimate $\hat{\mathbf{x}}_j^-(k)$ by applying dynamics (1)

$$\hat{\mathbf{x}}_j^-(k) = \mathbf{\Phi}\hat{\mathbf{x}}_j^+(k-1) \tag{3}$$

and predict error covariance $\mathbf{P}_j^-(k)$

$$\mathbf{P}_j^-(k) = \mathbf{\Phi}\mathbf{P}_j^+(k-1)\mathbf{\Phi}^T f + \mathbf{\Gamma}\mathbf{Q}_j\mathbf{\Gamma}^T, f > 1. \tag{4}$$

Step 2. Compute the weights $w_{i,j}$ for each position estimate $\mathbf{l}_i$ using a Bayesian formulation. Let $\pi_j(k) > 0$ be the *a priori* probability of associating a measurement with the object j $\left(\sum_j \pi_j(k) = 1\right)$. The weight $w_{i,j}$ is the *a posteriori* probability given by $\left(\sum_j w_{i,j} = 1\right)$

$$w_{i,j} \propto p\big(\,\mathbf{l}_i \mid \mu_j(k), \mathbf{C}_j(k)\big)\;\pi_j(k-1), \tag{5}$$

where the likelihood function $p(\cdot)$ is a Gaussian pdf, with mean

$$\mu_j(k) = \mathbf{H}\hat{\mathbf{x}}_j^-(k), \tag{6}$$

and covariance

$$\mathbf{C}_j(k) = \mathbf{H}\mathbf{P}_j^-(k)\mathbf{H}^T + \mathbf{R}. \tag{7}$$

Step 3. Use the weights $w_{i,j}$ to set the observation noise covariance matrices in (2) according to

$$\mathbf{R}_{i,j} = \mathbf{R}(w_{i,j} + \delta)^{-1}, \tag{8}$$

where δ is a small positive constant to prevent a division by zero. Compute the Kalman gain

$$\mathbf{K}_j(k) = \mathbf{P}_j^-(k)\mathbf{O}^T\left(\mathbf{O}\mathbf{P}_j^-(k)\mathbf{O}^T + \mathbf{R}_j\right)^{-1}, \tag{9}$$

where $\mathbf{R}_j$ is a block diagonal matrix composed of $\mathbf{R}_{i,j}$, and $\mathbf{O} = [\mathbf{H}\mathbf{H}\cdots\mathbf{H}]^T$ is the corresponding $2N \times 4$ observation matrix. Note that if $w_{i,j}$ has a small value, the corresponding measurement is effectively discarded by this formulation.

Step 4. Compute filtered estimates of the state

$$\hat{\mathbf{x}}_j^+(k) = \hat{\mathbf{x}}_j^-(k) + \mathbf{K}_j(k)\left(\mathbf{z}(k) - \mathbf{O}\mathbf{x}_j^-(k)\right) \tag{10}$$

and compute the associated error covariance matrix

$$\mathbf{P}_j^+(k) = \left(\mathbf{I} - \mathbf{K}_j(k)\mathbf{O}\right)\mathbf{P}_j^-(k), \tag{11}$$

where $\mathbf{z}(k) = [\mathbf{l}_1(k)^T, \mathbf{l}_2(k)^T, \ldots, \mathbf{l}_N(k)^T]^T$.

Step 5. Update *a priori* probabilities for assignments with a recursive filter

$$\pi_j(k) = a\pi_j(k-1) + (1-a)\tfrac{1}{N}\sum\nolimits_{i=1}^{N} w_{i,j}, \tag{12}$$

where $a < 1$ is a learning rate constant.

to alleviate the aforementioned problem it is wise to introduce a constant fading factor f into the proposed filtering solution (4) to keep it stable.

2.3 Multi-Object Tracking

Using detector responses as measurements makes it possible to initiate new objects, as well as terminate tracks that are no longer valid. In addition, the proposed method is able to adjust the scale of the objects tracked unlike the previous method (Huttunen and Heikkilä, 2008).

Object Initialization. When there exist a number of detections which are not assigned to any objects tracked, it is very likely that there is at least one new object in the image. Detector responses that are left unassigned and are overlapping with each other form the bounding box of a new object candidate. In the current implementation detections are combined in a very simple fashion. The set of unassigned detections is first partitioned into disjoint subsets. Two detections are in the same subset if their bounding regions overlap. Each partition yields a single final detection. A new object is created only if number of detections in the subset is greater than a predetermined threshold. Dimensions of the new object are selected as the average of each of the corners of all detections in the overlapping set. A new object cannot be entirely initiated from a single measurement since it does not provide velocity information, and also false detector responses may cause problems. Therefore we are using several frames in order to initialize an entirely new object.

Object Scale. Since the original method (Huttunen and Heikkilä, 2008) is based on color features, it does not provide any means to update the dimensions of the objects tracked. In this work, we are using detector responses and are able to update the dimensions $\left\{d_j^w, d_j^h\right\}$ using the formula

$$d_j^{\{w,h\}}(k) = a \cdot d_j^{\{w,h\}}(k-1) + (1-a)\sum_{i=1}^{N} w_{i,j} \cdot r_i^{\{w,h\}} \tag{13}$$

where the weights $w_{i,j}$ are given by (5), and a is the learning rate used in (12).

Track Termination. When an object under tracking goes out of the camera view, the tracking algorithm must have a way to detect it and remove the object. In our method, the criteria for terminating a trajectory is as follows. If no measurements are assigned to an object j within a certain time, i.e.$\forall w_{i,j} = 0$, the object is considered to be lost and is therefore deleted.

Occlusion Handling. There are two kinds of occlusions that can take place. The first case is occlusion due to a static obstacle, and the second alternative is that two or more tracked objects occlude each other. The proposed algorithm can handle both of the cases. The latter case is taken care of straightforwardly, since the measurements are assigned softly to the occluded objects. In other words, the same measurements are shared between several objects. When the objects finally split, all of the objects are assigned different measurements. When an object goes, for ex-

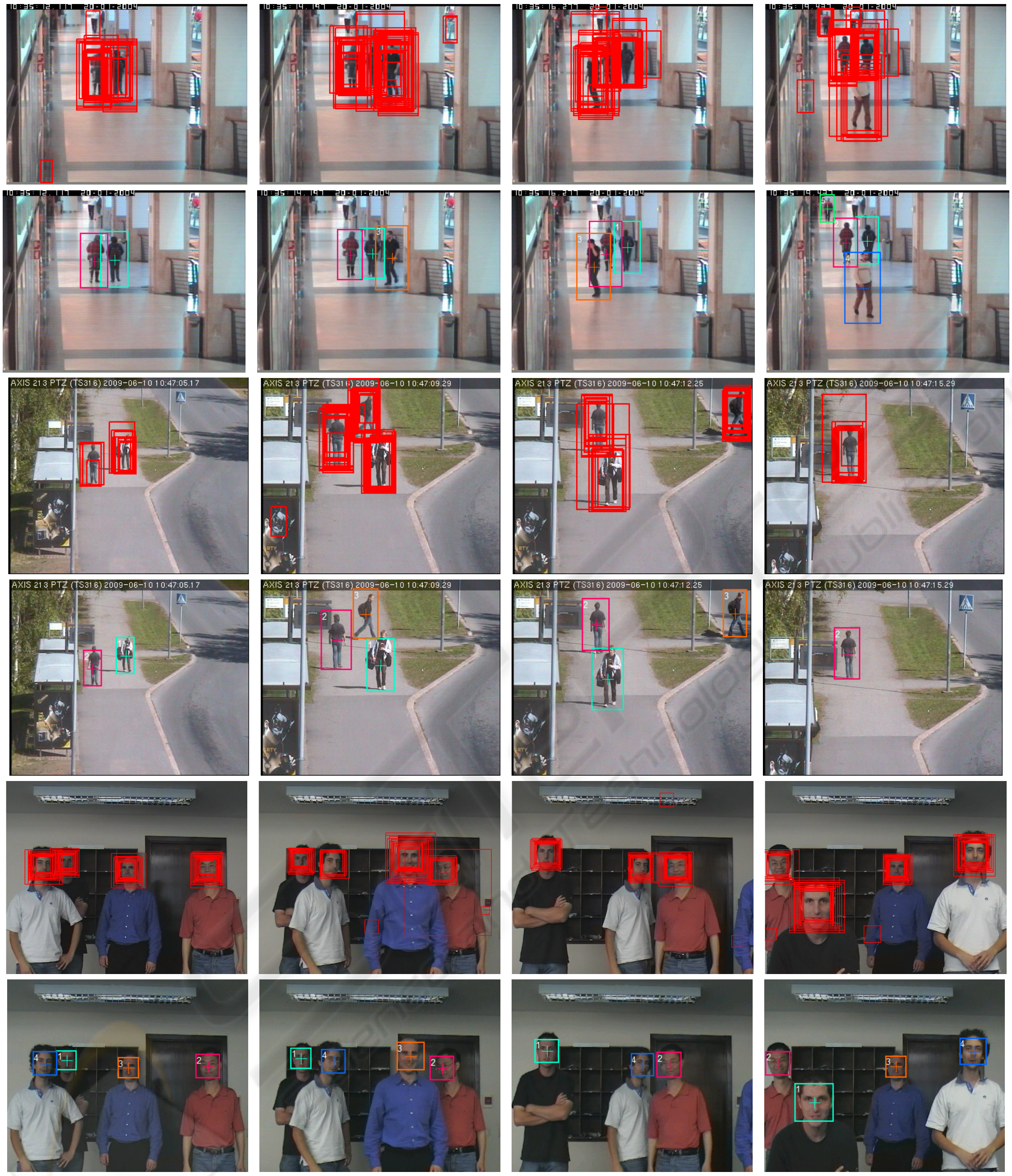

Figure 1: The tracking results of the proposed system for the *ExitEnterCrossingPaths1cor* (rows 1-2), *Axis_Busstop* (rows 3-4), and *motinas_multi_face_frontal* (rows 5-6) sequences. Detector responses (top), and the final tracking results (bottom).

ample, behind a static obstacle, there will not be any detections which could be used to update the position of the occluded object. It is therefore necessary to update the state of the object based on the dynamic model (1). When the target reappears from behind the obstacle, there are going to be measurements available again, and tracking can continue normally.

3 EXPERIMENTAL RESULTS

The object detectors used in the experiment are built on the cascade system proposed by Viola and Jones (2001) and improved by Lienhart and Maydt (2002). The actual implementation of the detector is based on the software found in the Intel OpenCV Library (`http://sourceforge.net/projects/`

`opencvlibrary/`).

Human Tracking. For the experiments on human tracking, the training samples needed for the cascade detector are taken from the DaimlerChrysler Pedestrian Classification Benchmark Dataset (Munder and Gavrila, 2006). The proposed algorithm has been tested on several sequences from the CAVIAR database (`http://homepages.inf.ed.ac.uk/rbf/CAVIAR/`). To demonstrate the feasibility of the concept described in this paper, some results of the test sequence *ExitEnterCrossingPaths1cor* are shown in Fig. 1. The proposed method can successfully track two objects even if they are occluded by a third person. Our own *Axis_Busstop* sequence has been captured using a PTZ camera that pans and zooms in to a person when he is walking by a bus stop. During the course of sequence the size of the objects changes significantly and turning the camera causes occasionally a large number of false measurements. Since there are also several objects that are tracked, the results indicate that the method is also able to track objects with a moving camera (Fig. 1).

Face tracking. To evaluate the usefulness of our method for tracking multiple faces, the method was tested in conjunction with a basic face detector. For face detection we used the face detector included in the OpenCV library directly. The face sequence *motinas_multi_face_frontal* used for testing is part of the AVSS2007 dataset (`http://www.elec.qmul.ac.uk/staffinfo/andrea/avss2007_d.html`). The sequence in question includes many situations where four targets repeatedly occlude each other while appearing and disappearing from the field of view of the camera. The results show that the method is able to track the objects after a total occlusion (Fig. 1).

All the tests were run on a regular Pentium 4 2.8GHz desktop PC using MATLAB. Based on the studies on all test sets, the most computationally intensive part of the method is usually detection. Also computation of the Kalman gain (9) may take some time depending on the number of measurements. However, based on the performance study of the current MATLAB implementation we are confident that the method is feasible for different applications when implemented in C/C++.

4 CONCLUSIONS

We have presented a new algorithm for tracking multiple objects based on detector responses. The method utilizes the Kalman filter and Expectation Maximization (EM) algorithms in order to update the state of the objects and assign detector responses to them. Current implementation uses a well-known cascade classifier to detect the objects of interest. Preliminary experiments conducted clearly indicate the usefulness of the approach proposed in this paper.

REFERENCES

Dempster, A., Laird, N., and Rubin, D. (1977). Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society*, 39(1):1–38.

Fortmann, T., Bar-Shalom, Y., and Scheffe, M. (1983). Sonar tracking of multiple targets using joint probabilistic data association. *IEEE-JOE*, 8(3):173–184.

Gavrila, D. (2000). Pedestrian detection from a moving vehicle. In *Proc. ECCV*, volume 1843 of *LNCS*, pages 37–49.

Hannuksela, J., Huttunen, S., Sangi, P., and Heikkilä, J. (2007). Motion-based finger tracking for user interaction with mobile devices. In *Proc. CVMP*.

Huang, C., Wu, B., and Nevatia, R. (2008). Robust object tracking by hierarchical association of detection responses. In *Proc. ECCV*, volume 5303 of *LNCS*, pages 788–801.

Huttunen, S. and Heikkilä, J. (2008). Multi-object tracking using binary masks. In *Proc. ICIP*, pages 2640–2643.

Joo, S.-W. and Chellappa, R. (2007). A multiple-hypothesis approach for multiobject visual tracking. *IEEE-TIP*, 16:2849–2854.

Kalman, R. E. (1960). A new approach to linear filtering and prediction problems. *Trans. of the ASME-Journal of Basic Engineering*, 82:35–45.

Leibe, B., Schindler, K., and Van Gool, L. (2007). Coupled detection and trajectory estimation for multi-object tracking. In *Proc. IEEE ICCV*, pages 1–8.

Lienhart, R. and Maydt, J. (2002). An extended set of haar-like features for rapid object detection. In *Proc. ICIP*, volume 1, pages 900–903.

Mikolajczyk, K., Schmid, C., and Zisserman, A. (2004). Human detection based on a probabilistic assembly of robust part detectors. In *Proc. ECCV*, volume 3021 of *LNCS*, pages 69–82.

Munder, S. and Gavrila, D. (2006). An experimental study on pedestrian classification. *IEEE-TPAMI*, 28(11):1863–1868.

Reid, D. (1979). An algorithm for tracking multiple targets. *IEEE-TAC*, 24(6):843–854.

Singh, V. K., Wu, B., and Nevatia, R. (2008). Pedestrian tracking by associating tracklets using detection residuals. In *Proc. IEEE WMVC*, pages 1–8.

Viola, P. and Jones, M. (2001). Rapid object detection using a boosted cascade of simple features. In *Proc. IEEE CVPR*, volume 1, pages 511–518.

Wu, B. and Nevatia, R. (2005). Detection of multiple, partially occluded humans in a single image by bayesian combination of edgelet part detectors. In *Proc. IEEE ICCV*, volume 1, pages 90–97.