Combining multiple motion estimates for vehicle tracking

Gil Milanese, Sylvia; Milanese, Ruggero; Pun, Thierry

How to cite

GIL MILANESE, Sylvia, MILANESE, Ruggero, PUN, Thierry. Combining multiple motion estimates for vehicle tracking. In: 4th European Conference Computer Vision - ECCV 96. B. Buxton and R. Cipolla (Ed.). Cambridge (UK). [s.l.] : Springer-Verlag, 1996. p. 307–320. (Lecture Notes in Computer Science) doi: 10.1007/3-540-61123-1_149

This publication URL: https://archive-ouverte.unige.ch/unige:47756
Publication DOI: 10.1007/3-540-61123-1_149

Tracking (2)

Combining Multiple Motion Estimates for Vehicle Tracking

Sylvia Gil, Ruggero Milanese and Thierry Pun

University of Geneva
Computer Science Department
24, rue Général Dufour,
1211 Genêve-4, Switzerland
E-mail: {*gil, milanese, pun*}*@cui.unige.ch*

Abstract. In this paper, the problem of combining estimates provided by multiple models is considered, with application to vehicle tracking. Two tracking systems, based on the bounding-box and on the 2-D pattern of the targets, provide individual motion parameters estimates to the combining method, which in turn produces a global estimate. Two methods are proposed to combine the estimates of these tracking systems: one is based on their covariance matrix, while the other one employs a Kalman filter model. Results are provided on three image sequences taken under different viewpoints, weather conditions and varying vehicle/road contrasts. Two evaluations are made. First, the performances of individual and global estimates are compared. Second, the two global estimates are compared and the superiority of the second method is assessed over the first one.

1. Introduction

In order to model a physical process, multiple models can often be designed, each of which is specialized on a particular aspect of the problem. This gives rise to the issue of combining multiple estimates, which is often done in two ways: by selecting the model providing the best estimate on a particular instance of the process (also called *hard-switch* method), or by constructing a combined estimate which weights the results of the individual ones and adaptively allows smooth transitions between them (also called *soft-switch* method [1]). The availability of multiple models can be useful for a robust vehicle tracking system which has to process data acquired under a large variety of conditions. For instance, weather and light conditions can produce drastic changes in the contrast between the road and the vehicles, introducing road reflectance and irregularities (rain), shadows (sunny), or the presence of vehicle lights at night. Let us assume that several tracking techniques are available. Any given input sequence $I(x, y, t)$ can be partitioned along the *time* dimension into various classes, representing characteristic illumination/weather conditions. Each estimator will generally perform best on some of these partitions. Thus, for each data partition or class, several estimators can provide meaningful information for the tracking task. Since it is not feasible to automatically estimate at each time the "best" estimator, an unsupervised combination of the tracking systems is needed.

In this work, we first describe two individual tracking systems (section 2), each of which independently estimates object motion parameters based on different visual features. The first feature used to represent the target is contour-based, consisting of the coordinates of the bounding rectangle obtained from the convex hull approximation of a moving-object's profile. The second feature is region-based, consisting of the 2-D pattern of the object. Then, two unsupervised methods for combining individual estimates of the two independent estimators are presented in section 3. Both combining methods take into account the instantaneous performance of the individual estimators. The first one computes a set of instantaneous weighting coefficients, used to combine the individual outputs into a global estimate. The second one provides in addition to the global estimate a confidence measure, which in turn is used to gate its own update. The performances (section 4) of the global estimates are first compared with the individual ones, then, a comparison is made between the two combining methods, in order to favor one. Results have been obtained on three image sequences with different time/weather/illumination conditions, acquired from different cameras. Finally, conclusions are reported in section 5.

2. Two Independent Tracking Systems

Traffic surveillance on urban and highway scenes has been widely studied in the past five years. One of the most popular methods, called model-based tracking, uses a 3-D model of a vehicle and is structured in two steps: (i) computation of scale, position and 3-D orientation of the vehicle (pose recovery), and (ii) tracking of the vehicle by fitting the model in subsequent frames by means of maximum-a-posteriori techniques [2] or Kalman filters [3] [4]. The vehicle model being quite detailed (3-D model), model-based tracking provides an accurate estimate of the vehicles 3-D position, which might not be needed for most applications, especially for highway surveillance. A simplified model of the vehicle is proposed in [5] by means of a polygon with fixed number of vertices, enclosing the convex hull of some vehicle's features. This approach dramatically reduces the model complexity. In [5] Kalman filters are used in order to estimate a vehicle's position and its motion using an affine model, which allows for translation and rotation. Although the method has shown good results, the fixed number of polygon vertices allows little variations on the objects shape. Some improvements on this point are proposed in [6] through the use of B-cubic splines, instead of polygons. In this case, a Kalman filter is used in order to track the curve in subsequent frames with a search strategy guided by the local contrast of the target in the image, i.e. with no use of motion information. In the context of traffic scenes, especially in the case of highways, vehicle's motion may be a powerful cue to direct the search for the target's position in subsequent frames. Another system that combines active contour models with Kalman filtering has been presented in [7]. In this case, the use of separate filters for the vehicle position and other motion parameters (affine model: translation and scale), has been shown to provide better results.

2.1 The Features Choice

Several choices are possible for the target's features, such as its color, its contour or a pattern defining its spatial layout. The major tendencies in the existing literature correspond to representations of the target's contour and to its description as a region. Both representations have advantages and drawbacks. Contour-based approaches [8] are fast, since they are based on the (efficient) detection of spatio-temporal gradients. Their major drawback is that, they may not have a physical meaning. Indeed, contour extraction depends on the local intensity variation between an object and the background, so that changes in their relative intensity may cause the appearance/disappearance of a contour. This type of features is thus reliable only when the contrast between the target and the background is sufficiently high, and constant in time. On the other hand, region-based approaches [9] represent the target through a 2D-pattern; they are quite accurate and do not depend on the background. Their drawbacks are the high computing time required for their manipulation (such as pattern matching [10]) and the sensitivity of pattern matching techniques to changes in scale and rotation. Contour- and region-based approaches thus appear to be complementary.

In this paper, both types of representations have been implemented. The contour-based feature is based on the bounding rectangle of the convex polygon and is represented through its center of gravity computed through its two characteristic corners (upper-left and lower-right). The region-based feature is the spatial pattern of the target, which is stored in a rectangular window. For both types of features, the tracked position remains the same: the center of the bounding box which is also the center of the 2D-pattern of the vehicle. An affine motion model (translation and scale changes) is used and ruled by Kalman filters.

2.2 The Kalman Filters

The two tracking systems are based on similar Kalman filter equations described in [3][7]. For both systems, the Kalman filter is used to track each visual feature of the moving target. For each feature we use a state vector, $\underline{x}_k$ to represent its position x_k, y_k and instantaneous motion parameters u_k, v_k, s_k. An affine motion model is used, which takes into account the translations along the x and y axes: u_k, v_k, as well as the scaling factor s_k representing the shrinkage/magnification of the target as it moves away or gets closer from/to the camera: $\hat{\underline{x}}_k = (x_k, y_k, u_k, v_k, s_k)^T$.

The measurements $\underline{z}_k$ are the features positions, as computed from the k-th image frame. Therefore, the correspondence between the measurements and the position state vector is given by Equation (1), where $\underline{w}_k$ is the measurement error and where the *observation matrix* H_k for a given feature at a given time is defined in Equation (2).

$$\underline{z}_k = H\hat{\underline{x}}_k(-) + \underline{w}_k \,, \tag{1}$$

$$H_k = \begin{bmatrix} 1 & 0 & 1 & 0 & \hat{x}_k(-)-x_{ck} \\ 0 & 1 & 0 & 1 & \hat{y}_k(-)-y_{ck} \end{bmatrix}. \tag{2}$$

The last column of the matrix H_k represents the vector joining the center of gravity of the target $(x_{ck}, y_{ck})^T$ to one of the corners of the bounding rectangle. A change in this vector indicates a change in the target's scale, and allows the estimation of the scale factor s_k. The tracking system is decoupled into two inter-dependent subsystems composed of the position coordinates of the target on the one hand, and of the velocity parameters on the other hand. This decomposition allows for a dimensionality reduction of the system which is advantageous for low rank matrix inversion and also for the association of a covariance matrix to each one of the sub-system state variables.

2.3 The Features Measurement

The first feature to be considered is the bounding rectangle of the moving vehicles. Its computation is based on the convex hull approximation of the targets profile [7]. The convexity assumption is not restrictive in the case of vehicles because in most projections their profiles are pretty compact. It also considerably simplifies the matching step required by the tracking procedure, since it allows to by-pass problems such as contour regularization [11] [13]. Furthermore, an extensive literature is available describing efficient methods for convex hull computation [14]. The measurement of the convex hull can be summarized as follows. At each step k, a search window is obtained by translating the previous bounding rectangle, according to the predicted motion and to a tolerance margin for safety. The spatial and the temporal gradients are computed inside the search window, and points where the gradient exceeds two fixed thresholds respectively, are used for the convex hull computation[10] (cf. Figure 1). It is then straightforward to derive its bounding rectangle, whose center is the tracked feature. This feature leads to a considerable information compression and avoids the problem of tracking vectors of varying size (variable number of vertices).

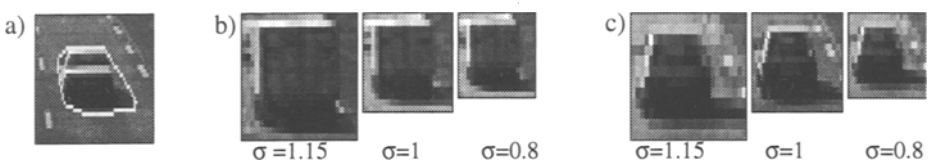

Figure 1: The tracked features: (a) the convex hull approximation of the targets' contours, shown in white; (b) and (c) representations of the target 2D patterns, with different scaling factors.

During the computation of the convex hull, some errors may occur. For instance, those introduced by an incorrect estimation of the predicted motion. In this case, the computed convex hull may exclude part of the target. Another source of errors degrading the measurement step is the fact of processing odd and even image fields (image parity change) resulting from an error in the image video sampling. This produces artificial temporal gradients at locations of a high spatial gradients, causing a deterioration in the shape of the convex hull.

The second feature representing the tracked objects is its 2-D pattern, stored as a gray-level mask. Given the target mask at a given frame, two scaled versions are computed by bilinear interpolation: $M_\sigma(x, y)$. These two masks provide an approximation of the object's appearance, caused by the objects's approaching or getting further away

from the camera (positive or negative scale parameter). For each target, three scaled patterns are thus available for matching, corresponding to the scaling parameters $\sigma = \{0.8, 1, 1.15\}$ (cf. Figure 1). The measurement is performed by the correlation of the masks $M_\sigma(x, y)$ and a window of interest in the next image frame. For each scaled mask $M_\sigma(x, y)$, a recognition rate R_σ is computed based on the correlation peak value, and the autocorrelation of the scaled masks:

$$R_\sigma = \max_{u,v} \left(\frac{\sum_u \sum_v (I(x+u, y+v, t+1) - \overline{I(x+u, y+v, t+1)}) \cdot (M_\sigma(u, v) - \overline{M_\sigma(u, v)})}{\sum_m \sum_n (M_\sigma(m, n) - \overline{M_\sigma(m, n)})^2} \right) \quad (3)$$

where $I(x, y, t+1)$ is the frame used for the correlation, and the sign ' $\bar{}$ ' denotes the average value. The value of R_σ quantifies the similarity between each scaled pattern and the target in the next frame. The locations of the highest peaks of the correlation surfaces (obtained with the masks $M_\sigma(x, y)$ yielding the highest value of R_σ) are retained as measurement candidates. When a scaled mask $M_\sigma(x, y)$, $\sigma \neq 1$ yielding the highest recognition rate has been selected for a sufficient number of consecutive frames, the 2-D pattern of the target is updated, its new scaled versions are computed, and the scale parameter of the motion model is updated. The mask update is achieved by copying into the new 2-D pattern window (with a new size), the values of image $I(x, y, t+1)$ around the selected correlation peak. As the size of the targets 2D pattern gets smaller, its correlation generally becomes unstable, giving rise to wrong measurements. In order to prevent these problems, a minimum size for the target is fixed, below which shrinkage of the 2D pattern is prevented. One of the advantages of this region-based tracking system is its immunity to incorrect temporal sampling (image parity change). The dynamic equations defining the tracking process require an initialization step, performed by a motion detection system described in [15].

3. Combining Estimates

Independently from motion tracking applications, estimates combination techniques are widely used in domains such as forecasting (see [16] [12] for a survey), statistics and neural network for problems such as regression, classification and time-series prediction, which require robustness to noise and the capability to cope with missing features. The most popular combination techniques are probably the "winner take all" and the averaging schemes. Despite its simplicity, averaging can provide interesting results [17]. For instance, the averaged output of a set of n unbiased and uncorrelated estimators, perturbed with uncorrelated noise, yields a mean squared error which is n times smaller than the mean squared error produced by individual estimators. In next section an overview of other more sophisticated combining techniques is presented.

3.1 Related Work

Although estimate combination is an intuitively attractive idea, some considerations must be made in order to quantify its actual efficiency. The work described in [18] con-

siders the consequences of combining a set of individual estimates into a global one, in terms of global bias and variance for regression. Here, the computation of weighting factors multiplying the estimators' outputs is not addressed, the focus being on the performances that can be reached by combining estimates. Some general conclusions can be stated. The bias depends exclusively on single-estimator properties. The variance is composed of two additive terms: the first one depends exclusively on each estimator and thus grows as the number of estimators increases. The second term depends on the covariance of the different estimators. Thus, uncorrelated estimators should be used in order to decrease the variance.

An interesting method for combining estimates, called *generalized ensemble method* (GEM) was originally introduced for regression problems [17]. It consists of a linear combination of a population of n individual estimators weighted by normalized factors α_i. It is shown that weakly correlated estimators provide large weighting factors. The performance of the GEM saturates as the number of individual regression estimators increases, that is, when individual estimators start violating the uncorrelation assumption. Therefore a small enough number of independent estimates generally provides the best performances of the GEM. Another formulation of the problem leads to similar conclusions [19]. In combining methods, it is useful to introduce the notion of *ambiguity*, which quantifies the disagreement of an ensemble of estimators, i.e. how a single estimator's output differs from the averaged output of all estimators [20]. It has been shown that the largest the ambiguity, the smaller the quadratic ensemble error. As it has already been pointed out, it is important to choose estimators which do not agree (i.e. are not correlated) in order to increase the ambiguity term and so decrease the ensemble error.

The fusion of data issued from individual sensors through a formalism called *hierarchical estimation* is presented in [21]. Its goal is to merge n local estimates produced by n different sensors (or "local agents"), into a single global estimate. The assumption underlying this formulation is the linearity of the modeled dynamic process. The proposed architecture is shown in Figure 2.

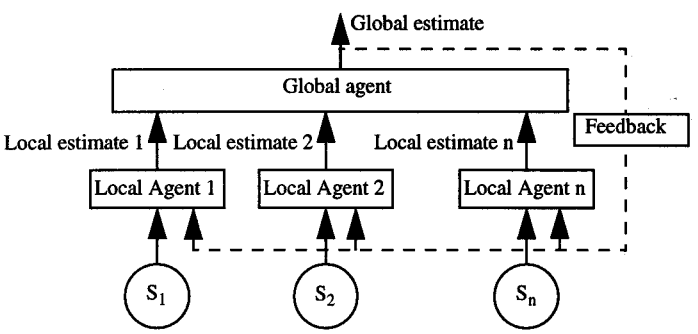

Figure 2: The fusion system is composed of n sensors, n local agents and one global agent. The feedback switch can be set *on* or *off*.

Global and individual (local) estimators are all ruled by Kalman filters. Local equations update is performed according to the customary Kalman equations, whereas the global equations update is achieved by the integration of local information. Therefore,

the global equations are updated according to the performances of individual estimators, that is inversely proportional to their respective measurement error. This method for computing the global estimate has two advantages: it builds a global covariance matrix which provides a confidence measure, and it prevents from updating the estimate when individual measurement errors are too large. Several methods for computing normalized *weighting functions* depending on the input data have been presented in [1]. Here we only report the *variance-based* method. The main idea is to use the estimators which are the most *certain* of their estimation. Under the assumption that individual estimators are uncorrelated and unbiased, it can be shown that the combined estimator is also unbiased and has the smallest variance, if the weighting functions are inversely proportional to the variance of the individual estimator.

3.2 Our Contribution

In consideration of these previous works, the following points should be emphasized: (i) it is important to combine only uncorrelated estimators, and (ii) a small number of estimators must be selected, in order not to increase the global bias and variance. Thus, a reduced number of independent estimators is considered. In particular, the two independent tracking systems based on different visual features (cf. section 2) are used as individual estimators. Two combining methods based on the literature are tested.

Method 1: The Co-variance Method. The first combination method applied to our problem is derived from the method introduced in [17], where the i-th constant weighting factors α_i is obtained from the covariance matrix C of the misfit functions m_i:

$$\alpha_i = \sum_j C_{ij}^{-1} / \sum_k \sum_j C_{kj}^{-1} \tag{4}$$

$$C_{ij} = E[m_i(x) \cdot m_j(x)] \tag{5}$$

$$m_i(x) = f(x) - f_i(x) \tag{6}$$

where $f(x)$ is the target function and $f_i(x)$ is the i-th estimation. However, in the context of object tracking, target functions are not available and therefore the misfit functions cannot be computed. An alternative error function able to quantify the adequacy of the estimate to the target function is thus required. The measurement error, i.e. the error between the estimate prediction and its measurement, is proposed as the alternative error function:

$$\underline{w}_k = \underline{z}_k - \hat{\underline{x}}_k(+) \; . \tag{7}$$

This approximation is valid as long as the measurement is close to the target function and it becomes biased as soon as the measurement is not exact. In this case, our misfit functions will be worst-case estimates of the original misfit functions and it will be best-case when both, the measurement and the prediction, are equally biased. The advantage is that the error function of the individual estimates is computed at each time k, yielding an instantaneous covariance matrix of the misfit functions, and providing the instantaneous weights.

Method 2: The Hierarchical Estimation Method. The second combination method applied to our tracking problem is derived from [21]. Local agents provide measurements at each time t, modeled by the following equation:

$$z_i(t) = H_i \cdot x(t) + v_i(t) \qquad i = 1, ..., n \tag{8}$$

where $z_i(t)$ is the measurement vector of agent i, $x(t)$ is the state vector to be estimated, $v_i(t)$ is the zero-mean Gaussian measurement noise, with covariance matrix $R_i(t)$, and H_i is a known measurement matrix. The fusion agent, collecting all individual measurement equations yields a global observation equation:

$$z(t) = H \cdot x(t) + v(t) \tag{9}$$

where:

$$z(t) = [z_1^T(t), ..., z_n^T(t)]^T \tag{10}$$

$$H = [H_1^T, ..., H_n^T]^T \tag{11}$$

$$v(t) = [v_1^T(t), ..., v_n^T(t)]^T \tag{12}$$

The individual estimate and covariance matrix are *updated* according to the customary Kalman equations [3], whereas the global estimate $\hat{x}$ and the global covariance matrix P are updated by integrations of local agents information:

$$P^{-1}(t|t) = P^{-1}(t|t-1) + \sum_{i=1}^{n} [P_i^{-1}(t|t) - P_i^{-1}(t|t-1)] \tag{13}$$

$$P^{-1}(t|t) \cdot \hat{x}(t|t) = P^{-1}(t|t-1) \cdot \hat{x}(t|t-1) + \sum_{i=1}^{n} [P_i^{-1}(t|t) \cdot \hat{x}_i(t|t) - P_i^{-1}(t|t-1) \cdot \hat{x}_i(t|t-1)]$$

4. Experiments

In this section the tracking performances of the individual and global estimators are analyzed. The two methods described in section 3 are used to combine estimates. In order to accurately quantify the position error of individual and global estimates, the "true" position of the vehicles at each frame has been manually extracted for each vehicle for all image sequences. In some cases, the decision of assigning a true position to a vehicle is not obvious, especially when vehicle shadows appear or disappear from one frame to the next. Two comparisons are made. The first one is between the position error of the individual methods vs. the global one, in order to quantify the gain introduced by using global estimates. The second comparison is between the two combining methods, in order to make a selection between them.

In order to compute the global position estimate, two individual estimates issued from the tracking systems are available: the bounding rectangle, and the 2-D pattern of the target. For the combining method 2, the covariance matrix of the position needs to reflect the confidence of the bounding rectangle center-of-gravity position. A worst-case approach is used by choosing the position covariance matrix of the corner presenting the largest trace (highest uncertainty). For the motion parameters, three individual

estimates are available: two provided by the corners of the bounding rectangle, and one originated by the correlation peak.

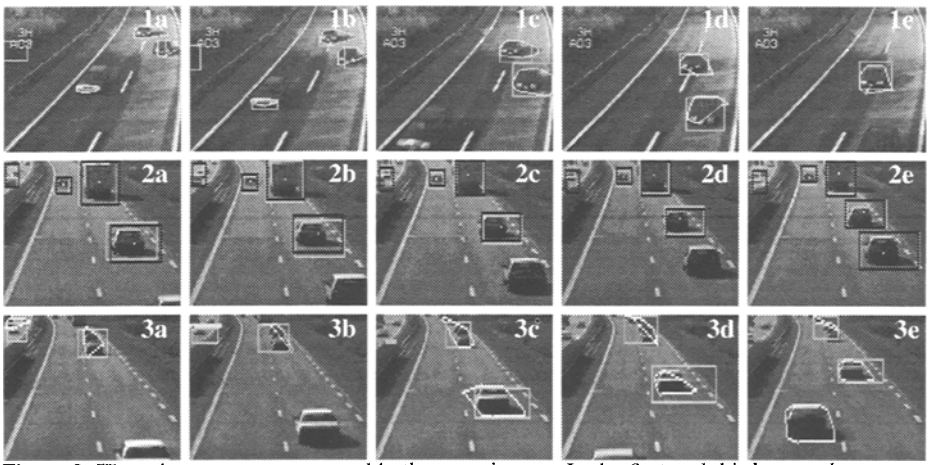

Figure 3: Three image sequences used in the experiments. In the first and third rows, the convex hull measurement is presented in white, while in the second row the correlation peak appears as a single white dot. The global prediction is represented (only for the second row) by a black rectangle. For all sequences, position predictions are reported by gray rectangles.

4.1 Input Image Sequences

The results obtained from three different image sequences are presented. They have been recorded at different locations, with different cameras, each of them with its own relative position to the road (cf. Figure 3). Of the three sequences, the latter two have been taken on a sunny day, which produces strong vehicles shadows. The first sequence has also been recorded with sunshine, but immediately after a rainstorm. The road thus appears partially wet and shadows are noticeable only where the road has dried. Moreover, shadows due to neighboring trees darken parts of the road. It should be noticed that the contrast between the cars and the road changes significantly between these three sequences.

In the third sequence (frame 3c), a wrong parity image occurs. This produces a noticeable deformation of the convex hull due to the artificially introduced strong spatio-temporal gradients on the white lane marks. Objects moving far away from the camera present weak gradients which produce an unstable convex hull shape. This effect can be observed in the third row of Figure 3, especially for the group of cars close to the top of the image. Finally, two consecutive frames of the third sequence are missing. Although this is not noticeable in Figure 3, the effect of this loss appears as a high error peak in the results (cf. Figure 4, i) and j), frame #18).

4.2 Comparing Global and Individual Estimates

The comparison of the error of the individual vs. global estimates are shown through diagrams where three error plots are superposed: those of the two individual methods

$e_k^1 = \hat{\underline{x}}_k^1(+) - \underline{x}_k$, $e_t^2 = \hat{\underline{x}}_k^2(+) - \underline{x}_k$ and the combined one $e_k^{comb} = \hat{\underline{x}}_k^{comb}(+) - \underline{x}_k$, where $\underline{x}_k$ are the vector of the hand-entered "true" vehicle position at time k . The first plot, shown with circles, represents the instantaneous largest individual estimate position error: $e_k^{max} = \max(e_k^1, e_k^2)$, while the second curve (crosses) shows the instantaneous lowest position error: $e_k^{min} = \min(e_k^1, e_k^2)$. Finally, the last curve (diamonds, solid line) represents the global estimate position error. For each image sequence, two such diagrams will be shown: the left one for method 1 and the right one for method 2.

The first target to be analyzed is the car in the 1st image sequence (cf. Figure 3, 1st row), located at the top the image. In this sequence the difficulty of dealing with vehicle shadows is manifest. During the initialization step, performed on a darker part of the road, shadows were not visible and so were excluded from the vehicle's mask. Shadows then appear in frames 12 and 13 producing a significant increase of the error (cf. Figure 4, (a) and (b)). In this case, the hand-entered "true" positions did not include car shadows. The gradient-based tracking system is highly sensitive to these morphological changes of the target, and produces an increase in the estimated scale factor and an enlargement of the prediction of the target width and height. This strongly affects the tracking process and points out the need for two independent scaling factors. On these critical frames, the convex hull measurement (including shadow) is considerably different from its prediction (no shadow, cf. Figure 4 (a) and (b)). For this reason, the measurement error of this tracked feature is high and its associated weights remain small, compared to those of the correlation-based method. It can be seen that both combining methods yield a global estimate that is close to the best individual estimate, which is the correlation-based one.

Let's now analyze the second image sequence (Figure 3, 2nd row) by focusing on the small car, the one closer to the truck. Individual estimators provide robust tracking performance, despite extremely small frame-to-frame displacements, thanks to the high vehicle/road contrast. However, its individual estimators present large errors, relative to the small target size ($\cong$ 10 pixels). The estimate of the combining method 1 is often between the two individual ones, indicating that individual measurement errors are of the same order (cf. Figure 4 (c) and (d)). The estimate of the combining method 2 (cf. Figure 4, (d)), does not follow any of the individual estimates, indicating that the individual measurement errors are large. Its error is almost always smaller than even the best individual one. Let's now consider the truck. Individual errors are very small, relative to the targets size ($\cong$ 27 pixels). Since the 2-D pattern is so large, the correlation-based method is very reliable. Some errors are introduced by the image crop when the truck slowly exits from the upper border of the image (cf. Figure 4 (e) and (f)). It can be seen that both global predictors perform well.

Let's analyze the performances on two vehicles of the third image sequence (Figure 3, 3rd row). The first object to be analyzed (Figure 4 (g) and (h)) is the group of cars at the top of the image. For the correlation-based tracking system, errors are due to a wrong mask initialization which includes part of the road. For the other tracking system, the cause of the large errors are the weak spatio-temporal gradients. Similarly to the previous experiments, when both individual measurement errors are equally high, the estimate of the combining method 1 falls in between the individual ones. For the combining method 2, these large measurement errors prevent from updating the global estimate, which thus provides almost always smaller error than both individual ones.

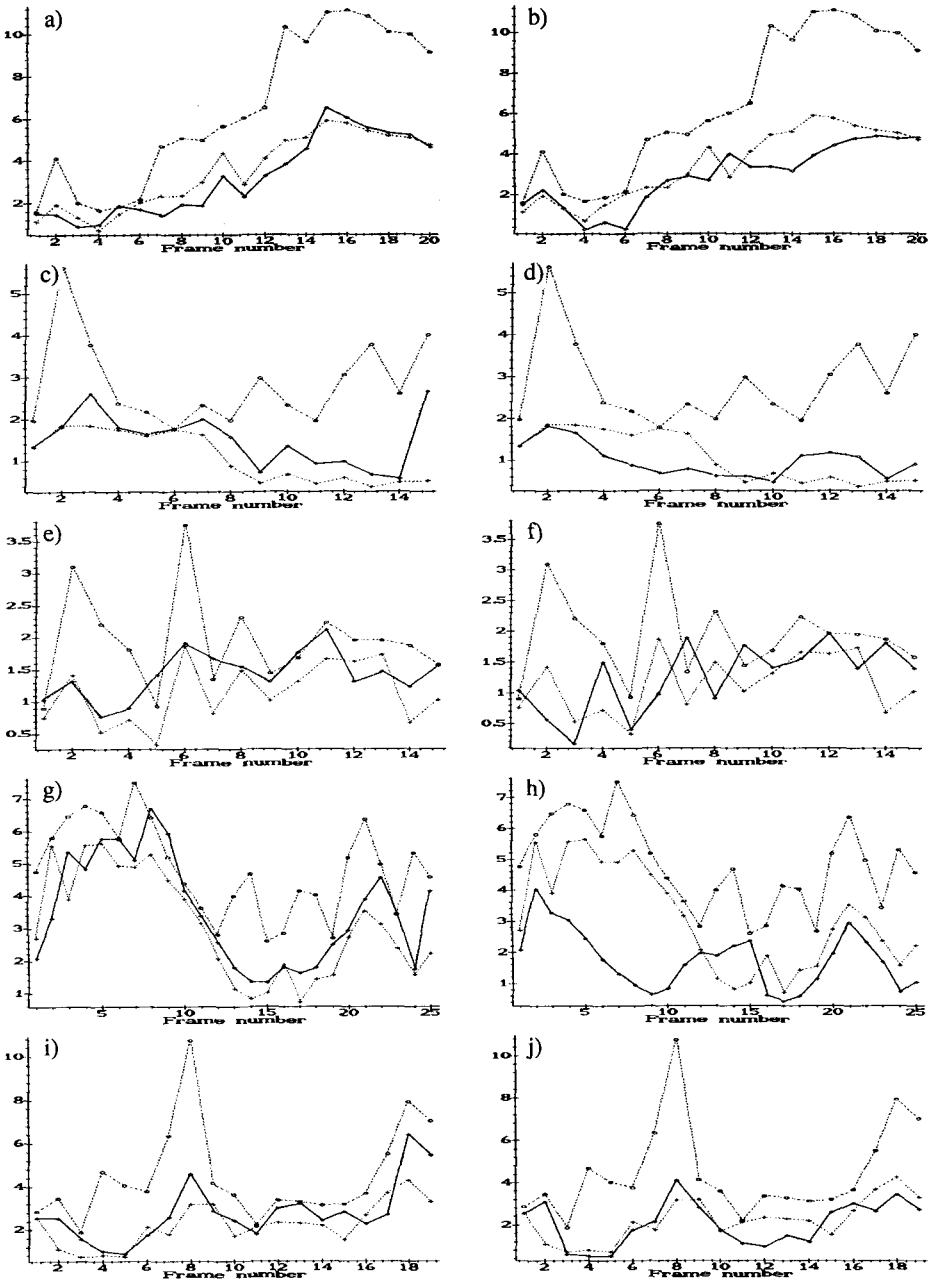

Figure 4: Position error diagrams: largest individual (⊖), lowest individual (+), and global (◇) with solid line. First sequence: (a), (b) car at the top of the image. Second sequence: (c), (d) small car; (e), (f) truck. Third sequence: (g), (h) group of small cars; (i), (j) single car. Left: combining method 1; right: combining method 2.

The vehicle described in Figure 4 (i) and (j) is the large clear car on the right lane of Figure 3, 3rd row. The effect of the wrong temporal sampling is clear in frames nos. 7 and 8, as well as the effects of the missing frames at frame no. 18. These two peaks in the error are mostly limited to the gradient-based tracking system, correlation being rather insensitive to these accidents. Besides these critical frames, errors are generally small relative to the target size ($\cong$ 50 pixels). The two combining methods are approximately equivalent, and both are generally better than the best individual one.

TABLE 1. Performance of the combining estimates method #1.

Tracked vehicle	% $e_k^{comb} \le e_k^{min}$	% $e_k^{comb} \ge e_k^{max}$	% better than average	approx. target size (pixels)	avrg. error of best individual	avrg. error combined estimate
Fig. 4a)	16 %	8 %	64 %	20	3.0	3.3
Fig. 4c)	31 %	0 %	63 %	50	2.1	2.6
Fig. 4e)	20 %	0 %	80 %	10	1.1	1.5
Fig. 3.2 large car	60 %	6.66 %	80 %	36	1.5	1.3
Fig. 4g)	40 %	20 %	53 %	27	1.1	1.2
Fig. 4i)	60 %	0 %	90 %	20	3.5	3.3

TABLE 2. Performance of the combining estimates method #2.

Tracked vehicle	% $e_k^{comb} \le e_k^{min}$	% $e_k^{comb} \ge e_k^{max}$	% better than average	approx. target size (pixels)	avrg. error of best individual	avrg. error combined estimate
Fig. 4b)	88 %	0 %	96 %	20	3.0	1.8
Fig. 4d)	73 %	0 %	89 %	50	2.1	2.0
Fig. 4f)	60 %	0 %	100 %	10	1.1	1.0
Fig. 3.2 large car	66 %	6.7 %	93 %	36	1.5	1.3
Fig. 4h)	13 %	26 %	53 %	27	1.1	1.3
Fig. 4j)	70 %	0 %	95 %	20	3.5	3.0

A more quantitative comparison of the performances of individual vs. global estimators is given in tables 1 and 2, which respectively concern combining methods 1 and 2. The comparison between the errors of different estimators is given in terms of several table entries. One is the percentage of occurrences where the global estimate performs better than the best individual estimate or, in error terms, when $e_k^{comb} \le e_k^{min}$. Another index counts the percentage of occurrences in which the global estimates is worse than the worst individual estimates ($e_k^{comb} \ge e_k^{max}$). Finally, we compare the global estimates with what could be the estimates of a trivial integration method, i.e. the simple average of the individual estimates ($e_k^{comb} \le (e_k^1 + e_k^2)/2$). The two last columns show the average of the position error, over multiple frames. In order to have an indication of rela-

tive importance of these errors, the vehicle size (in pixels) is reported with each set of measures.

In terms of average errors, by looking at table 1, it appears that the estimate of the 1st combining method, although it remains within reasonable bounds, only outperforms the best individual estimate for one vehicle. The second combining method, however, has much better performance, consistently better than the lowest individual error. In the only case where this is not true, the individual error is already small (1.1 pixels), for a target size of approximately 27 pixels. It appears that the global estimate provided by the combining method 1 performs well only when the features measurement does not present major difficulties for neither individual tracking systems, as predicted in section 3. Apart from this specific weakness, global estimates of both methods are in general sensitive to the following sources of errors: (i) wrong mask initialization; (ii) inaccurate mask update; and (iii) consistently wrong feature measurements in both individual tracking methods (only for the combining method 1).

5. Conclusions

In this paper, we introduced the problem of combining multiple models of a dynamic process, and presented an application to target tracking using multiple motion estimators. Two independent tracking systems are used: one is based on the bounding-box of the moving object, the other one uses the object's 2-D pattern. Both individual tracking systems provide motion parameters estimates which are then combined into a global one. Several classes of combining methods presented in the literature are reviewed. Tracking performances are evaluated on 6 vehicles, from three different outdoors image sequences. To precisely compute each individual and global estimate error, the "real" vehicle positions at each frames have been hand entered.

The first method is based on a linear combination of the individual estimates, whose weights are inversely proportional to the covariance matrix coefficients. When the two individual methods both provide good estimates, then the results of this combining method represent an improvement, yielding smaller position error than each individual one. When neither individual method performs correctly, however, then combining method does not introduce any improvements. This is due to the practical need to replace the error function, usually computed through a training set, by the measurement error, which is a pessimistic estimate when the measurement is not exact, and optimistic when both measurement and prediction are equally wrong. Overall, this combining method has been shown to be superior to the averaging technique. However, only in some cases does this method outperform the best individual estimate.

The second combining method integrates the estimates of the individual methods using a Kalman filtering approach. In this case, the combined estimates clearly outperform both the averaging and the best individual estimate. In five cases out of six, the error of the combined method was significantly reduced, while in the sixth case, the individual estimates were already very good. The performances of this combining method can be further improved by avoiding errors due to wrong mask initialization, and improper mask updates. A comparison between the two proposed combining methods clearly shows the superiority of the second, Kalman filter-based one, both in

terms of average position error, and in the number of frames where the results outperform the best instantaneous individual method. These result are encouraging.

Acknowledgments We would like to thank researchers from the connectionnist and vision mailing list, Prof. J. Malik (U. Berkeley), Dr. D. Koller, IRISA and IRBSA for their help, encouragement and discussions, and for providing image sequences.

References

1. V. Tresp, M. Taniguchi, "Combining Estimators Using Non-Constant Weighting Functions", to appear in Proc. of the Neural Information Processing Systems, 1995.
2. D. Koller, K. Daniilidis, H.H. Nagel, Model-Based Object Tracking in Monocular Image Sequences of Road Traffic Scenes, *Int. Jour. of Computer Vision*, Vol. 3, pp. 257-281, 1993.
3. A. Gelb, Applied Optimal Estimation, The MIT Press, MA, and London, UK, 1974.
4. K. Baker, G. D. Sullivan, Performance Assessment of Model-based Tracking, *Proc. of the IEEE Workshop on Applications of Computer Vision*, pp. 28-35, Palm Springs, CA, 1992.
5. F. Meyer, P. Bouthémy, Region-Based Tracking in an Image Sequence, *Proceedings of the European Conference on Computer Vision*, pp. 476-484, S. Margarita-Ligure, Italy, 1992.
6. A. Blake, R. Curwen, A. Zisserman, A Framework for Spatiotemporal Control in the Tracking of Visual Contours, *Int. Journal of Computer Vision*, Vol. 11, pp. 127-147, 1993.
7. D. Koller, J. Weber, J. Malik, Robust Multiple Car Tracking with Occlusion Reasoning, *Proceedings of the Third European Conference on Computer Vision*, Vol. 1, pp. 189-199, Stockholm, Sweden, 1994.
8. O. Faugeras, Three-Dimensional Computer Vision: A Geometric Viewpoint, The MIT Press, London, UK, 1993.
9. J. Shi, C. Tomasi, Good Features to Track, *IEEE Conf. on Computer Vision and Pattern Recognition*, pp. 593-600, Seattle, USA, 1994.
10. S. Gil, R. Milanese, T. Pun, "Comparing Features for Target Tracking in Traffic Scenes", to appear in *Pattern Recognition*.
11. D. Geiger, J.A. Vlontzos, Matching Elastic contours, *IEEE Conf. on Computer Vision and Pattern Recognition*, pp. 602-604, New York, June 1993.
12. International Journal of Forecasting, special issue on combining forecasts, Vol. 4(4), 1989.
13. F. Leymarie, D. Levine, Tracking Deformable Objects in the Plane Using an Active Contour Model, *IEEE Trans. on Pattern Analysis and Machine Intelligence*, Vol. 15, 1993.
14. F. P. Preparata, M. I. Shamos, Computational Geometry, Springer-Verlag, 1985.
15. S. Gil, T. Pun, Non-linear Multiresolution Relaxation for Alerting, *Proceedings of the European Conference on Circuit Theory and Design*, pp. 1639-1644, Davos, CH, 1993.
16. C.V. Granger, "Combining forecasts - twenty years later", Journal of Forecasting, Vol. 8, pp. 167-173, 1989.
17. M.P. Perrone and L.N. Cooper, "When networks disagree: Ensemble methods for hybrid Neural Networks", in "Neural Networks for Speech and Processing", Editor R.J. Mammone, Chapman-Hall, 1993.
18. R. Meir, "Bias, variance and the combination of estimators; the case of linear least squares", Preprint, Technion, Heifa, Israel, 1994.
19. S. Hashem, "Optimal Linear Combinations of Neural Networks", Ph.D. Thesis, Purdue University, December 1993.
20. A. Krogh, J. Vedelsby, "Neural Network Ensemble, Cross Validation and Active Learning", to appear in Proc. of the Neural Information Processing Systems, 1995.
21. Y. Bar-Shalom, "Multitarget Multisensor Tracking: Advanced Applications", Artech, 1990.