Article:
Loza, A., Mihaylova, L., Bull, D. et al. (2009) Structural similarity-based object tracking in multimodality surveillance videos. Machine Vision and Applications, 20 (2). 71 - 83. ISSN: 0932-8092

https://doi.org/10.1007/s00138-007-0107-x

Machine Vision and Applications manuscript No.
(will be inserted by the editor)

Artur Łoza · Lyudmila Mihaylova · David Bull · Nishan Canagarajah

Structural Similarity-Based Object Tracking in Multimodality Surveillance Videos

Received: date / Accepted: date

Abstract This paper addresses the problem of object tracking in video sequences for surveillance applications by using a recently proposed structural similarity-based image distance measure. Multimodality surveillance videos pose specific challenges to tracking algorithms, due to, for example, low or variable light conditions and the presence of spurious or camouflaged objects. These factors often cause undesired luminance and contrast variations in videos produced by infrared sensors (due to varying thermal conditions) and visible sensors (e.g. the object entering shadowy areas). Commonly used colour and edge histogram-based trackers often fail in such conditions. In contrast, the structural similarity measure reflects the distance between two video frames by jointly comparing their luminance, contrast and spatial characteristics and is sensitive to relative rather than absolute changes in the video frame. In this work, we show that the performance of a particle filter tracker is improved significantly when the structural similarity-based distance is applied instead of the conventional Bhattacharyya histogram-based distance. Extensive evaluation of the proposed algorithm is presented together with comparisons with colour, edge and mean-shift trackers using real-world surveillance video sequences from multimodal (infrared and visible) cameras.

Keywords structural similarity measure · object tracking · video sequences · particle filtering · colour and edge cues · multimodal data

Artur Łoza · David Bull · Nishan Canagarajah
Department of Electrical and Electronic Engineering
University of Bristol
Bristol BS8 1UB, UK,
E-mail: Artur.Loza@bristol.ac.uk

Lyudmila Mihaylova
Department of Communication Systems
Lancaster University
Lancaster, LA1 4WA, UK,
E-mail: Mila.Mihaylova@lancaster.ac.uk

1 Introduction

Recently there has been an increasing interest in target tracking in video sequences that is of particular importance to military and civilian surveillance applications [10,11,9,27, 20,12]. Various methods for object tracking in video sequences were proposed in the literature. Bayesian methods are the most often used where the problem reduces to the reconstruction of the probability density function of the object states given the measurements and the prior knowledge. Different image features such as colour, motion, edges, shape and texture can be used to track the moving object [27,24]. This problem presents many challenges, including how to model the moving object, the choice of measurement model and the function characterising the similarity between two images or video frames. One specific issue with object tracking in video sequences (compared to, for example, tracking with radar data) is that no measurement model exists in explicit form.

Multiple feature-based tracking provides more information about the object and hence increases the robustness of the algorithms to occlusions and clutter. An optimisation approach for different features selection is proposed in [33] which is embedded in a layered hierarchical vision-based tracking algorithm. When the target is lost, the layers cooperate to perform a rapid search for the target and continue tracking. Features can be implemented concurrently in the democratic integration approach [34] and contribute simultaneously to the overall result. Some other methods rely on applying tensor theories to visual surveillance, e.g. for gait recognition [31,32].

The performance of the tracking algorithm, however, also depends on the measure characterising the similarity between the two subsequent video frames. Frequently used functions include the Bhattacharyya distance [1,6] and the non-metric Kullback-Leibler measure.

In this paper we propose a particle filter (PF) based on a structural similarity measure for object tracking in video sequences. The PF framework has been proven to be a scalable and powerful approach, able to cope with non-linearities and

uncertainty present in video sequences (see, for example, [27,25,3,4]). In this work, we combine these strengths of the particle filtering with the properties of the structural similarity measure. The similarity measure proposed in [37] captures the spatial characteristics of an image and has been shown to be robust to illumination and contrast changes. The way this measure is formulated reflects the fact that the Human Visual System (HVS) is highly adapted to extract structural information from the visual scene [38]. It has been originally used for the purposes of quality assessment of distorted and fused images [37,28,18]. In this paper, we show how this measure can be applied to the video tracking problem. It replaces histograms used for calculation of the measurement likelihood function within a particle filter. We demonstrate that it is a good and efficient alternative to histogram-based tracking. This work extends the early results reported in [19] with more detailed investigation including the consideration of tracking in multimodality and fused videos. The performance of the proposed method is compared with results obtained with a PF using the Bhattacharyya distance (some of which have been reported in [21]) and mean-shift tracker [5], both applied to multimodality and fused videos.

The remaining part of the paper is organised as follows. Section 2 presents the structural similarity-based distance for tracking. Section 3 presents a PF with the proposed similarity measure, describes the likelihood and motion model of the object of interest. Section 4 contains results over real-world video sequences. Finally, Section 5 summarises the results and discusses open issues for future research.

2 Distance measure

2.1 Structural similarity measure

The proposed method uses a similarity measure computed directly in the image spatial domain. This approach differs significantly from other particle filtering algorithms, that compare image distributions represented by their sample histograms [25,27,30].

Although many simple image similarity measures exist (for example, mean square error, mean absolute error or peak signal-to-noise ratio), most of these have failed so far to capture the perceptual similarity of images/video frames under the conditions of varied luminance, contrast, compression or noise [37]. Recently, based on the premise that the HVS is highly tuned to extracting structural information, a new image metric has been developed, called the Structural SIMilarity (SSIM) index [37]. The SSIM index, between two images, **a** and **b** is defined as follows:

$$S(\mathbf{a},\mathbf{b}) = \left(\frac{2\mu_a\mu_b}{\mu_a^2+\mu_b^2}\right)\left(\frac{2\sigma_a\sigma_b}{\sigma_a^2+\sigma_b^2}\right)\left(\frac{\sigma_{ab}}{\sigma_a\sigma_b}\right) = l(\mathbf{a},\mathbf{b})\, c(\mathbf{a},\mathbf{b})\, s(\mathbf{a},\mathbf{b}), \tag{1}$$

where μ denotes the sample mean

$$\mu_a = \frac{1}{L}\sum_{j=1}^{L} a_j, \tag{2}$$

σ denotes the sample standard deviation

$$\sigma_a = \sqrt{\frac{1}{L-1}\sum_{j=1}^{L}(a_j-\mu_a)^2} \tag{3}$$

and

$$\sigma_{ab} = \frac{1}{L-1}\sum_{j=1}^{L}(a_j-\mu_a)(b_j-\mu_b) \tag{4}$$

corresponds to the sample covariance. The estimators are defined identically for images **a** and **b**, each having L pixels. The image statistics are computed in the way proposed in [37], i.e. locally, within a 11×11 normalised circular-symmetric Gaussian window.

2.2 Selected properties of the SSIM

The three components of (1), l, c and s, measure the luminance, contrast and structural similarity of the two images, respectively. Such a combination of image properties can be seen as a fusion of three independent image cues. The relative independence assumption is based on a claim that a moderate luminance and/or contrast variation does not affect structures of the image objects [38].

In the context of the multimodal data used in our investigation, an important feature of the SSIM index is (approximate) invariance to certain image distortions. It has been shown [37,38], that the normalised luminance measurement, l, is sensitive to the relative rather than absolute luminance change, thus following the masking feature of the HVS.

Similarly, the contrast comparison function, c, is less sensitive to contrast changes occurring in images with high base contrast. Finally, the structure comparison, s, is performed on contrast-normalised signal with mean luminance extracted, making it immune to other (non-structural) distortions.

These particular invariance properties of the SSIM index make it suitable for the use with multimodal and surveillance video sequences. The similarity measure is less sensitive to the type of global luminance and contrast changes produced by infrared sensors (results of varied thermal conditions or exposure of the object) and visible sensors (for example, the object entering dark or shadowy areas or operating in variable lighting conditions). Moreover, the structure comparison is expected to be more reliable in scenarios when spurious objects appear in the scene or when there is not enough discriminative colour information available. The latter may be the result of the tracked object being set against background of similar colour or when background-like camouflage is deliberately being used.

It should be noted that some desirable properties of the structural similarity measure, resulting from the use of the covariance, make the proposed method likely to fail when the target undergoes significant structure changes, for example due to the rotation or change of the viewing angle. In Section 4.5 some of the issues involved in SSIM-based tracking under such conditions are discussed and possible solutions are suggested.

2.3 Image dissimilarity

Below, we present a method of evaluating the likelihood function $\mathscr{L}$ (see Section 3.3), based on the structural similarity between two greyscale images. We begin by noting that the measure defined in (1) is symmetric, i.e.

$$S(\mathbf{a},\mathbf{b}) = S(\mathbf{b},\mathbf{a}) \tag{5}$$

and has a unique upper bound

$$S(\mathbf{a},\mathbf{b}) \leq 1,\ S(\mathbf{a},\mathbf{b}) = 1 \text{ iff } \mathbf{a} = \mathbf{b}. \tag{6}$$

A natural way of converting such a similarity $S(\mathbf{a},\mathbf{b})$ into dissimilarity $D(\mathbf{a},\mathbf{b})$ is to take $D(\mathbf{a},\mathbf{b}) = 1 - S(\mathbf{a},\mathbf{b})$. An alternative way [39],

$$D(\mathbf{a},\mathbf{b}) = \frac{1}{|S(\mathbf{a},\mathbf{b})|} - 1 \tag{7}$$

is preferred, however, as it maps an interval [-1, 1] into [0, ∞] (0 when the images are identical) and as a result is more sensitive to very dissimilar vectors. It can be easily shown that the measure (7) satisfies non-negativity, reflexivity and symmetry conditions. Although sufficient for our purposes, this dissimilarity measure is not a metric, as it does not satisfy the triangle condition.

3 A particle filter for object tracking

3.1 Particle filtering

Particle filtering [2,8,13,14,29,27] is a method relying on sample-based reconstruction of probability density functions. The aim of sequential particle filtering is to evaluate the *posterior* probability density function (pdf) $p(\mathbf{x}_k|\mathbf{Z}_k)$ of the state vector $\mathbf{x}_k \in \mathbb{R}^{n_x}$, with dimension n_x, given a set $\mathbf{Z}_k = \{\mathbf{z}_1,\ldots,\mathbf{z}_k\}$ of sensor measurements up to time k. The Monte Carlo approach relies on a sample-based construction to represent the state pdf. Multiple particles (samples) of the state are generated, each one associated with a weight $W_k^{(\ell)}, \ell = 1,2,\ldots,N$, which characterises the quality of a specific particle ℓ.

An estimate of the variable of interest is obtained by the weighted sum of particles. Two major stages can be distinguished in the particle filtering method: *prediction* and *update*. During prediction, each particle is modified according to the state model of the region of interest in the video frame, including the addition of white noise in order to simulate the effect of the random walk.

Assuming that the posterior pdf at time $k-1$ (the initial pdf) is available, the prior pdf of the state at time k is obtained in prediction stage via Chapman-Kolmogorov equation:

$$p(\mathbf{x}_k|\mathbf{Z}_{k-1}) = \int_{\mathbb{R}^{n_x}} p(\mathbf{x}_k|\mathbf{x}_{k-1}) p(\mathbf{x}_{k-1}|\mathbf{Z}_{k-1}) d\mathbf{x}_{k-1}. \tag{8}$$

Once a measurement $\mathbf{z}_k$ is available, $p(\mathbf{x}_k|\mathbf{Z}_k)$ is recursively obtained in the update step

$$p(\mathbf{x}_k|\mathbf{Z}_k) = \frac{p(\mathbf{z}_k|\mathbf{x}_k) p(\mathbf{x}_k|\mathbf{Z}_{k-1})}{p(\mathbf{z}_k|\mathbf{Z}_{k-1})}, \tag{9}$$

where $p(\mathbf{z}_k|\mathbf{Z}_{k-1})$ is a normalising constant. Thus, the recursive update of $p(\mathbf{x}_k|\mathbf{Z}_k)$ is proportional to

$$p(\mathbf{x}_k|\mathbf{Z}_k) \propto p(\mathbf{z}_k|\mathbf{x}_k) p(\mathbf{x}_k|\mathbf{Z}_{k-1}). \tag{10}$$

The posterior density $p(\mathbf{x}_k|\mathbf{Z}_k)$ is approximated as

$$p(\mathbf{x}_k|\mathbf{Z}_k) \approx \sum_{\ell=1}^{N} \widehat{W}_k^{(\ell)} \delta(\mathbf{x}_k - \mathbf{x}_k^{(\ell)}) \tag{11}$$

based on the likelihood $\mathscr{L}(\mathbf{z}_k|\mathbf{x}_k^{(\ell)})$ (18) of the measurement and particle weights. Here, $\delta(.)$ is the Kronecker delta function. The details of PF implementation, particle weights and likelihood calculation are given in Table 1 and Section 3.3.

An inherent problem with particle filters is degeneracy (the case when only one particle has a significant weight). A *resampling* procedure helps to avoid this by eliminating particles with small weights and replicating the particles with larger weights. Various approaches for resampling have been proposed (see [8], for example). In this work, the residual resampling method [17,35] was used.

3.2 Motion model

The initial (reference) region surrounding the object of interest, denoted as $\mathbf{t}_{\text{ref}}$, is chosen manually. In our case this is a rectangular region, and we are tracking its centre (see Section 3.4 for more detailed description of the target region model).

The motion of the moving object is modelled by the random walk model,

$$\mathbf{x}_k = \mathbf{F}\mathbf{x}_{k-1} + \mathbf{v}_{k-1}, \tag{16}$$

with a state vector $\mathbf{x}_k = (x_k, y_k, s_k)^T$ comprising the pixel coordinates of the centre of the region surrounding the object and the region scale s_k. $\mathbf{F}$ is the transition matrix ($\mathbf{F} = \mathbf{I}$ in the random walk model) and $\mathbf{v}_k$ is the process noise assumed to be white, Gaussian, with a covariance matrix

$$\mathbf{Q} = \text{diag}(\sigma_x^2, \sigma_y^2, \sigma_s^2). \tag{17}$$

Table 1 The particle filter algorithm

Initialisation

1. Generate samples $\{\mathbf{x}_0^{(\ell)}\}, \ell = 1,2,\ldots,N$, from the initial distribution $p(\mathbf{x}_0)$. Initialise weights $W_0^{(\ell)} = 1/N$

For $k = 1,2,\ldots,$

Prediction Step

2. For $\ell = 1,2,\ldots,N$, sample $\mathbf{x}_k^{(\ell)} \sim p(\mathbf{x}_k|\mathbf{x}_{k-1}^{(\ell)})$ from the motion model for the object region.

Measurement Update: evaluate the importance weights

3. The cue is used as a 'measurement'. Compute the weights

$$W_k^{(\ell)} \propto W_{k-1}^{(\ell)} \mathscr{L}(\mathbf{z}_k|\mathbf{x}_k^{(\ell)}). \tag{12}$$

based on the likelihood $\mathscr{L}(\mathbf{z}_k|\mathbf{x}_k^{(\ell)})$ (18) of the cue.

4. Normalise the weights,

$$\widehat{W}_k^{(\ell)} = W_k^{(\ell)} / \sum_{\ell=1}^{N} W_k^{(\ell)}. \tag{13}$$

Output

5. The posterior mean state estimate $\mathbf{x}_k$ is computed using the collection of samples (particles)

$$\hat{\mathbf{x}}_k = \sum_{\ell=1}^{N} \widehat{W}_k^{(\ell)} \hat{\mathbf{x}}_k^{(\ell)}. \tag{14}$$

Selection step (resampling)

6. Estimate the effective number of particles N_{eff} [17]

$$N_{\text{eff}} = \frac{1}{\sum_{\ell=1}^{N} (\widehat{W}_k^{(\ell)})^2}. \tag{15}$$

If $N_{\text{eff}} \leq N_{\text{thr}}$ (N_{thr} is a given threshold) then perform resampling: multiply/suppress samples $\mathbf{x}_k^{(\ell)}$ with high/low importance weights $\widehat{W}_k^{(\ell)}$, in order to introduce variety and obtain N new random samples. Set $W_k^{(\ell)} = \widehat{W}_k^{(\ell)} = 1/N$.

The estimation of the scale permits adjustment of the region size of the moving objects, e.g., when it goes away from the camera, when it gets closer to it, or when the camera zoom varies.

3.3 Likelihood model

The normalised distance between the two regions $\mathbf{t}_{\text{ref}}$ (reference region) and $\mathbf{t}_k$ (current region), for particle ℓ is substituted into the likelihood function, modelled as an exponential:

$$\mathscr{L}(\mathbf{z}_k|\mathbf{x}_k^{(\ell)}) \propto \exp\left(-D^2(\mathbf{t}_{\text{ref}}, \mathbf{t}_k)/D_{\text{min}}^2\right), \tag{18}$$

where $D_{\text{min}} = \min_{\mathbf{x}}\{D(\mathbf{t}_{\text{ref}}, \mathbf{t}_k)\}$. This smooth likelihood function, although chosen empirically by the authors of [27], has been in widespread use for a variety of cues ever since. The structural properties of the region are extracted through SSIM (1) and are used directly to calculate the distance D in (18) as shown in (7). The likelihood function is then used to evaluate the importance weights of the particle filter, to update the particles and to obtain the overall estimate of the centre of the current region $\mathbf{t}_k$, as shown in (12)–(14), Table 1.

3.4 Target model

The tracked objects are defined as image regions within a rectangle specified by the state vector. The reference image regions associated with the object of interest are shown in Figure 1. In the particle filtering framework as specified in Table 1, a rectangle corresponding to each particle, centred at location (x,y) and resized according to the scale parameter of the state, is computed. The extracted region is then compared to the target region using the distance measure (7).

4 Performance evaluation

4.1 Video sequences

The performance of our method is demonstrated over six multimodal video sequences, in which we aim to track a pre-selected moving person. The sequence *cross* (5 sec duration), taken from our multimodal database [15], contains three people walking rapidly in front of a stationary camera. The main difficulties posed by this sequence are: the colour similarity between the tracked object and the background or other passing people, and a temporal near-complete occlusion of the tracked person by a passer-by.

The sequence *man* (40 sec long), has been obtained from [26]. It is a long recording showing a person walking along a car park. Apart from object's similarity to the nearby cars, and the shadowed areas, the video contains numerous instabilities. These result from a shaking camera (changes in the camera pan and tilt), fast zoom-ins and zoom-outs, and a altered view angle towards the end of the sequence.

The sequence *doorway_ir* (10 sec), taken from [15], contains an infrared recording of two people walking towards a stationary camera. The two persons look quite similar and the tracked object is often partially occluded by nearby objects.

The three multimodal sequences *bushes* [15], contain simultaneous registered infrared (*ir*), visual (*vi*) and complex wavelet transform fused (*cwt*, see [16] for details) recordings of two camouflaged people walking in front of a stationary camera (10 sec). The tracked individual looks very similar to the background. The video contains changes in the illumination (the object entering shadowy areas) together with nonstationary surroundings (bushes moved by strong wind).

Fig. 1 Reference frames from the test videos

4.2 Trackers used in the comparison

In order to assess the performance of our proposed tracking algorithm, we compare it with a PF tracker based on independent and combined colour and edge cues [4] (referred to as 'colour', 'edges' and 'colour&edges' in the text). Unlike in the structural similarity-based 'structure' tracker, the distance between the reference and the current frame in (18) is calculated by the Bhattacharyya distance, as first proposed in [24,6]. In the PF based on combined cues, 'colour&edges', the likelihood is calculated as a product of the likelihoods of the separate 'colour' and 'edges' cues. No motion-based trackers have been used in the comparison, as preliminary experiments have shown that the multiple moving objects, moving background and low light videos pose a great challenge to such a tracker.

The covariance matrix $\mathbf{Q}$ of the motion model (17) is chosen as follows: $\mathrm{diag}(2.5^2, 10^2, 0.01^2)$ for *cross* video sequence, $\mathrm{diag}(2.5^2, 2.5^2, 0.05^2)$ for *man*, $\mathrm{diag}(2^2, 2.5^2, 0.02^2)$ for *doorway_ir*, and $\mathrm{diag}(5^2, 5^2, 0.005^2)$ for all three *bushes* sequences. A relatively low number ($N = 100$) of particles has been used for all videos.

An additional benchmark technique, the mean-shift tracker, based on a different principle than the particle filtering, has been used in the comparison. The mean-shift algorithm is a non-parametric technique relying on finding the modes of the underlying target–current frame pdf in the feature space (colour space in this case). It is an adaptive gradient ascent method, whose estimation resolution is controlled by a kernel. The video trackers based on a recursive mean-shift have been very popular mostly due to their simplicity of implementation and usage (only kernel and iteration limit required from the user) and minimal computational complexity. The implementation used here is based on [5,6] and employs Epanechnikov kernel and maximum of 20 iterations. Additionally, in order to preliminarily ascertain the potential of using SSIM with this technique, a combined structural similarity–mean-shift version of the procedure has been developed. The mean-shift tracker has been modified by replacing the Bhattacharyya distance with the SSIM measure when determining the scale of the target state.

4.3 Error plots

The Root Mean Squared Error (RMSE) in the state space has been used to evaluate the performance of the developed technique. The RMSE can be formulated as follows

$$\mathrm{RMSE}(k) = \sqrt{\frac{1}{M}\sum_{m=1}^{M}(x_k - \hat{x}_{k,m})^2 + (y_k - \hat{y}_{k,m})^2} \qquad (19)$$

where $(\hat{x}_{k,m}, \hat{y}_{k,m})$ stand for the upper-left corner coordinates of the tracking box determined by the central position and scale, corresponding to the state estimated by the PF in the frame k, in m^{th} independent Monte Carlo realisation ($M = 50$ in our simulations). The ground truth states (x_k, y_k) correspond to the true positions of the object and have been gen-

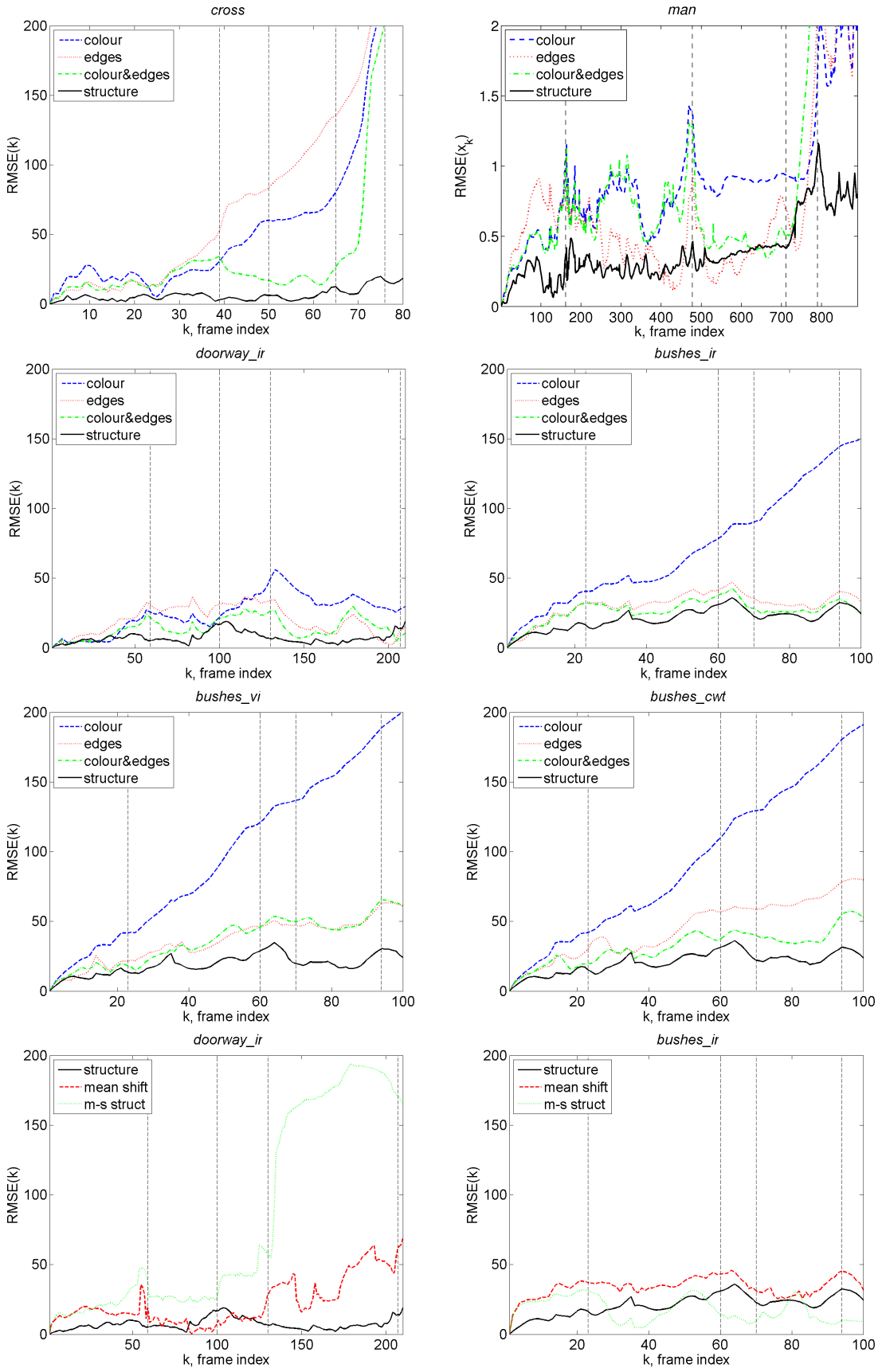

Fig. 2 RMSE plots for the tested trackers. The video frames marked by the vertical lines are shown in Figures 3–6

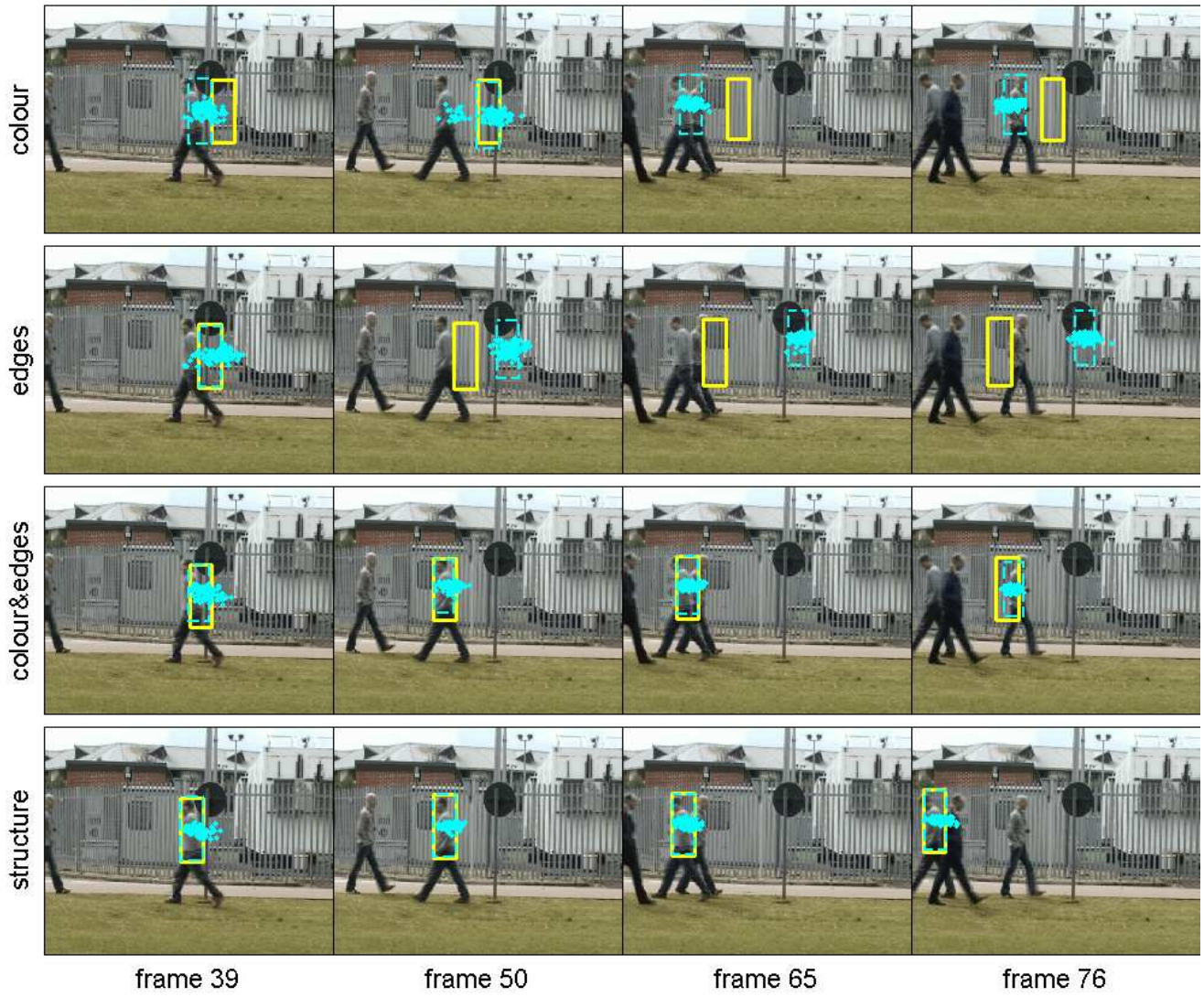

Fig. 3 Example video frames with average output of the tracker (solid line rectangle), a single trial output (dashed line rectangle and particles) superimposed, sequence *cross*

erated by manually creating the tracking box surrounding the object in the test videos.

The error estimates (19) are shown in Figure 2. It can clearly be seen that the proposed method never looses the object and outperforms the other methods in nearly all instances. The colour-based PF tracker is the most prone to fail or give imprecise estimates of the object's state. Combining edge and colour cues is usually beneficial, however in some cases (videos *man* and *cross*) the errors of the colour-based tracker propagate through the performance of the tracker, making it less precise than the tracker based on edges alone. Another observation is that the 'structure' tracker has been least affected by the modality of *bushes* and the fusion process, which demonstrates the robustness of the proposed method to luminance and contrast alterations.

Bottom of Figure 2 shows the performance of the mean-shift tracker as compared with the proposed method. For brevity, only error plots of two videos are shown, for which the mean-shift tracker achieved the best performance. In two videos, *man* and *cross*, it has failed to track the object throughout the sequence. Since the mean shift algorithm is a memoryless colour-based tracker, its poor performance in these two sequences is due to the object's fast motion and its similarity to the surrounding background. In other sequences, it was found that the mean-shift tracker is less precise than the PF tracker with the proposed structural similarity measure, and performs comparably either to 'colour&edges' or 'edges' tracker. However, as the plots in Figure 2 illustrate, the mean-shift tracker occasionally matched the performance of the similarity measure tracker, half-way through the *doorway_ir* and *bushes* sequences.

The use of the combined structural similarity–mean-shift algorithm resulted in improved tracking performance in most of the videos, compared with the aforementioned perfor-

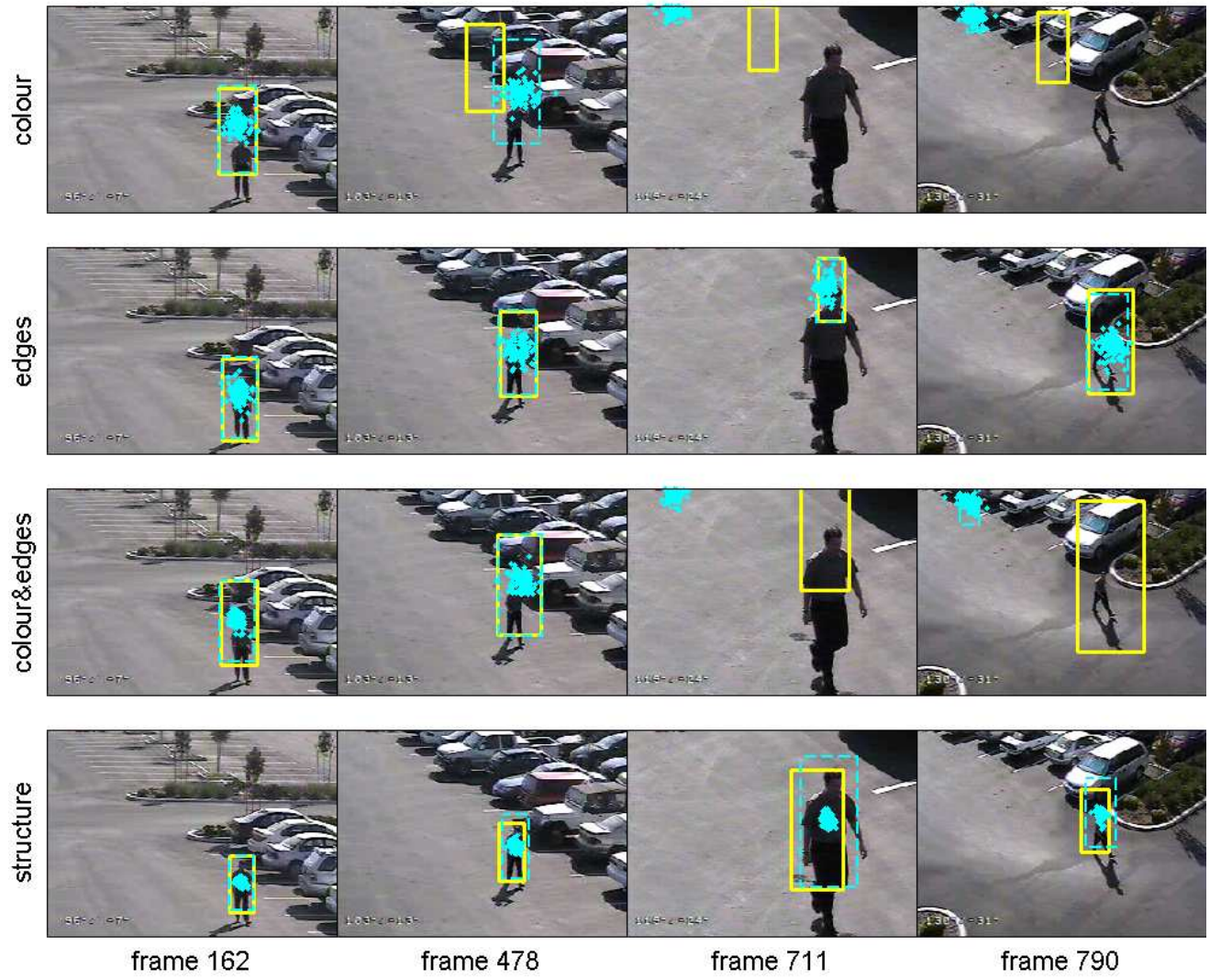

Fig. 4 Example video frames with average output of the tracker (solid line rectangle), a single trial output (dashed line rectangle and particles) superimposed, sequence *man*

mance of the conventional mean-shift tracker. The combined tracker has not been able to track the object in *cross* sequence. It has succeeded, however, in tracking the object accurately during the most of *doorway_ir* and *man* sequence (not shown here). Moreover, the tracking accuracy in all three *bushes* videos improved significantly (see Figure 2). This crude combination of the two techniques therefore gave promising results and the further extension of the SSIM measure to producing the density function in the mean-shift algorithm will be a subject of the future research.

4.4 Example frames

A closer investigation of the selected output frames illustrates the performance of the different methods. Figures 3–6 show the object tracking boxes constructed from the mean locations and scales estimated during the tests. Additionally, the particles and object location obtained from one of the Monte Carlo trials are shown. Since very similar performance has been obtained for all three *bushes* videos, only the fused sequence, containing complementary information from both input modalities, is shown. The visual difference between contents of the input *bushes* videos (colour information, a person hidden in shaded area) can be seen by comparing the reference frames in Figure 1.

In the sequence *cross*, Figure 3, the 'colour' and 'edges' trackers are attracted by the road sign, which eventually leads to the loss of the object. Then, the first non-occluding passerby causes the 'colour&edges' cue tracker to loose the object (frame 65). The 'structure' tracker is not distracted even by the temporary occlusion (frame 76).

The shaking camera in the sequence *man* (Figure 4, frame 162), has less effect on the 'structure' tracker than on the re-

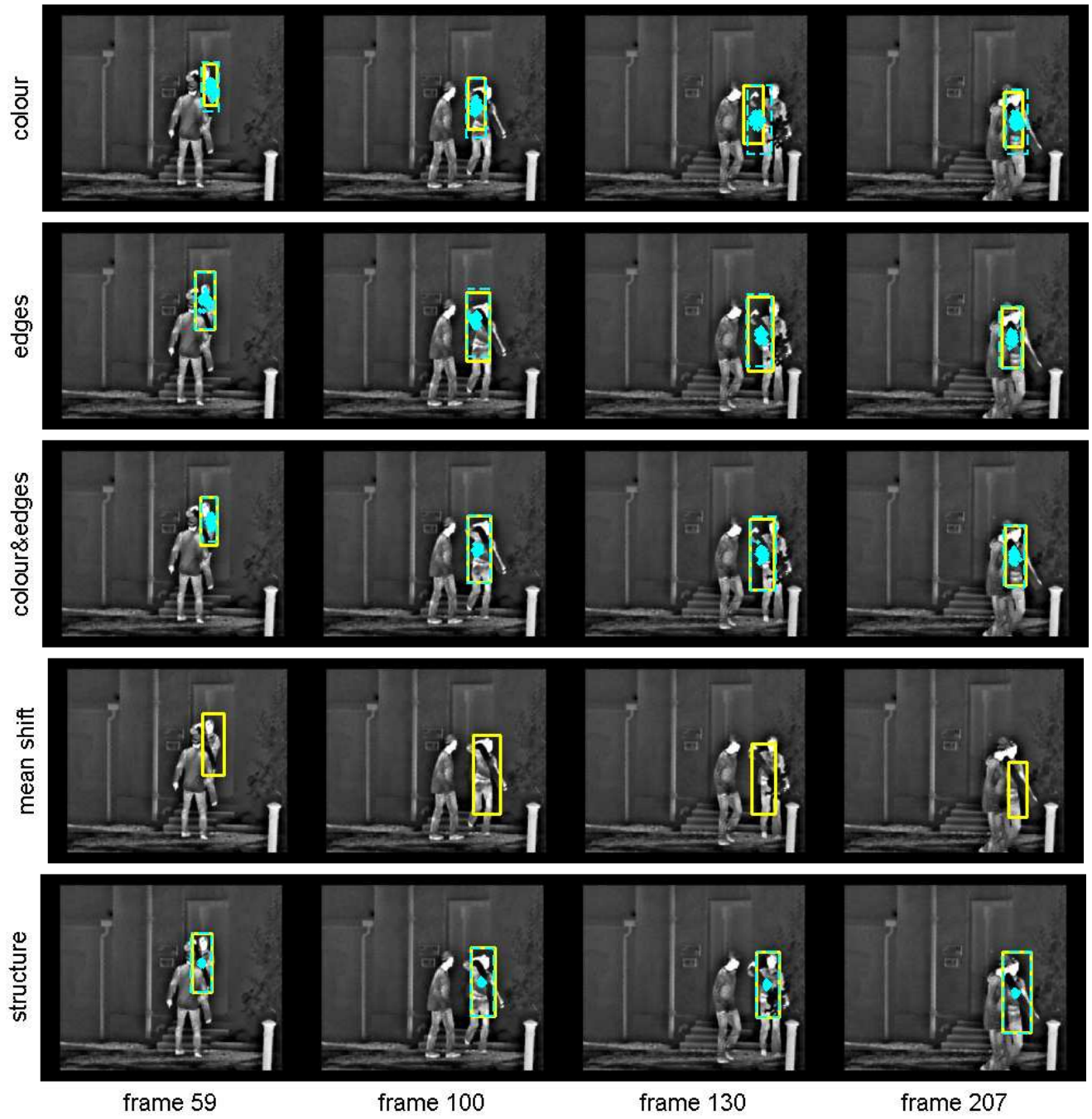

Fig. 5 Example video frames with average output of the tracker (solid line rectangle), a single trial output (dashed line rectangle and particles) superimposed, sequence *doorway_ir*

maining trackers, which appear to choose the wrong scale of the tracking box. Moreover, the remaining trackers do not perform well in case of similar dark objects appearing close-by (shadows, tyres, frame 478, where the 'colour' tracker permanently looses object) and rapid zoom-in (frame 711) and zoom-out of the camera (frame 790). Our method, however, seems to cope well with both situations. It should be noted, however, that 'colour&edges' (and 'edges') trackers show a good ability of recovering from some of the failings.

Although all the methods tested were able to track the person in the sequence *doorway_ir*, Figure 5, the proposed method is the most precise with respect both to position and correct scaling of the tracking box.

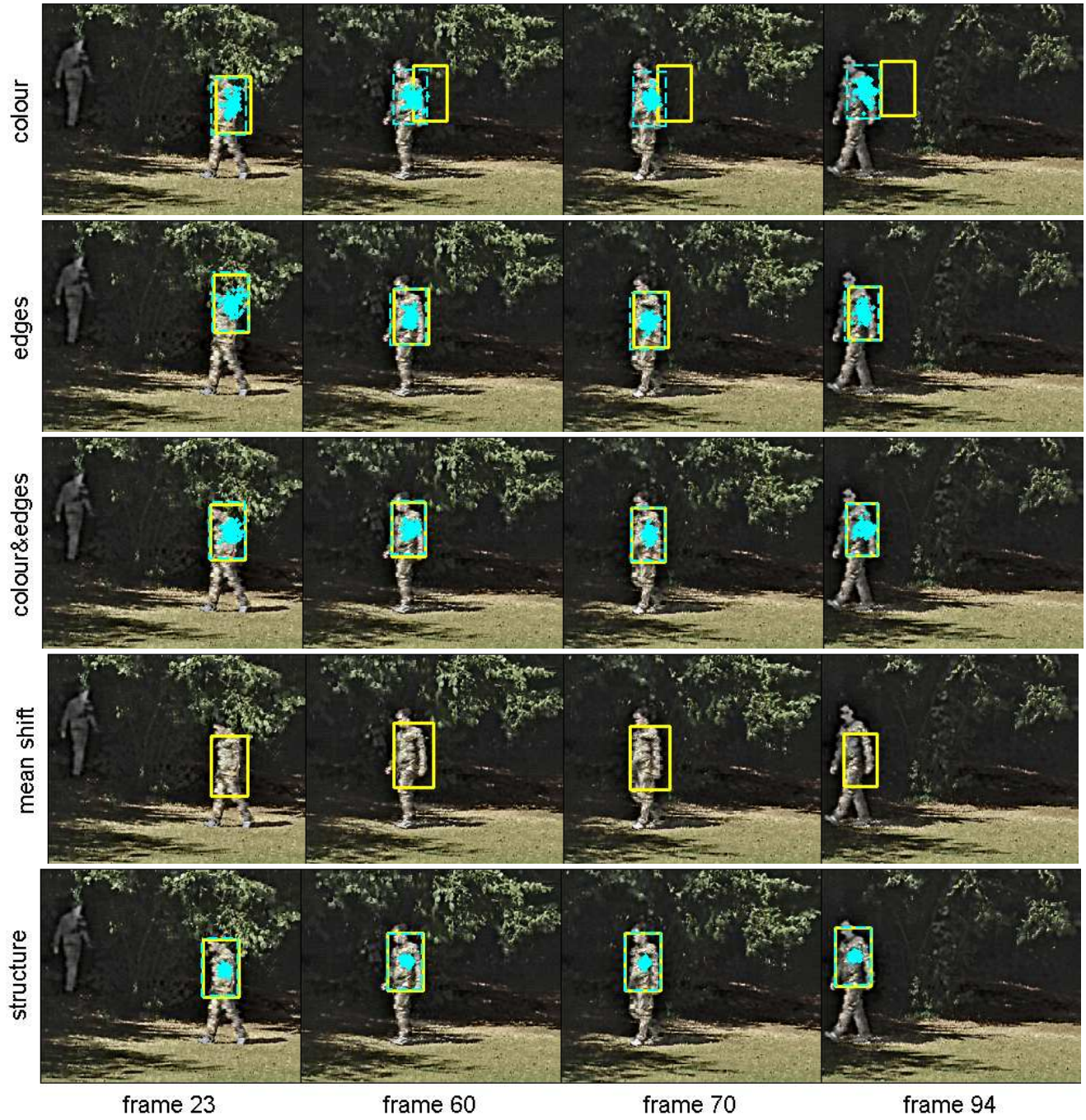

Fig. 6 Example video frames with average output of the tracker (solid line rectangle), a single trial output (dashed line rectangle and particles) superimposed, sequence *bushes_cwt*

Similarly, in the multimodal sequence *bushes*, Figure 6, the proposed 'structure' tracker is the most precise and the 'colour' tracker the least precise (see also Figure 2). The use of the fused video, although resulting in slightly deteriorated performance of the 'edges' tracker, can still be motivated by the fact that it retains complementary information useful both for the tracker and a human operator [21,7]: contextual information from the visible sequence and a hidden object location from the infrared sequence.

A single-trial output shown in Figures 3–6 exemplifies the spread of the spatial distribution of the particles. Typically, in the 'structure' tracker, particles are the most concentrated. Similar features can be observed in the output of the 'colour&edges' tracker. The particle distribution of the remaining PF trackers is much more spread, often at-

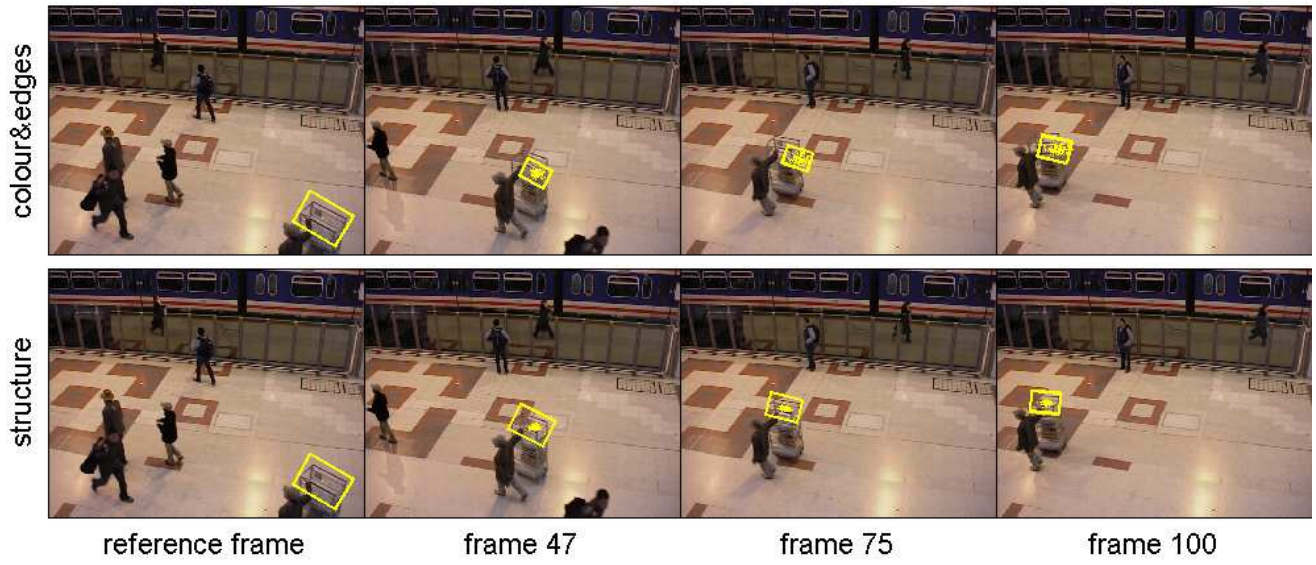

Fig. 7 Example video frames with a tracker output (tracking rectangle and particles) superimposed, sequence *S1-T1-C* containing a rotating object

tracted by spurious objects (see Figures 3 and 4, in particular). It should also be noted that, the actual performance of the tracker varies between realisations, often giving different results compared with the output averaged over all Monte Carlo trials. Also in this respect the proposed method has been observed to be the most consistent, i.e., its results had the lowest variation.

4.5 Possible extensions

In real-life video sequences of long-duration, the properties of the target are likely to change significantly. Variation of the target may be for example due to the change of lighting or viewing angle, motion, or non-rigid transformation of the target.

It can be demonstrated that the proposed structural similarity-based tracking method has low sensitivity to image distortions resulting from lighting changes (see Section 2.2). However, it is generally acknowledged that, as a result of the covariance being computed in the spatial domain, the structural similarity-based tracker is more sensitive to structural distortions as compared to histogram-based trackers. The use of the sliding window applied when estimating the local moments decreases the sensitivity of the SSIM measure to a small displacement of the target (e.g. a person waving hands while walking in *cross* sequence). It is expected that a trade-off between the local and global sensitivity to such distortions can be achieved by varying the window size. This however may not be sufficient for some sequences. In general, in order to be able to track the target over an extended period of time, the tracker should be able to adapt to target alterations, e.g., by updating the target model and/or by being invariant to target's transformations [36]. The adaptive solutions, while beyond the scope of this communication, may include target representation update [24] (this method raises issues as to when and what extent to update). Another approach is to treat model representation as a part of the state [22], updated at each step. Such an extended state space would, however, increase the complexity of the algorithm significantly.

Below, a type of target distortion, for which the state space can be easily extended, is considered: rotation of the target in the plane approximately perpendicular to the camera's line-of-sight. A simple solution to the tracking of the rotating objects is to include an orientation of the target in the state space, by taking $\mathbf{x} = (x_k, y_k, s_k, \alpha_k)^T$, where α_k is the orientation angle. The complexity of the algorithm is increased slightly due to the need to generate the rotated versions of the reference object (which can, possibly, be precomputed). For some video sequences it may also be necessary to increase the number of particles, in order to sufficiently sample the state space. The results of tracking a rotating trolley in a sequence from PETS 2006 Benchmark Data [23], with the use of 150 particles are shown in Figure 7. The figure shows examples of frames from two best-performing trackers, 'colour&edges' and 'structure'. Apart from the rotation scaling of the object, additional difficulty in tracking arose because the object was partially transparent and thus often took on the appearance of the non-stationary background. However, also in this case 'structure' tracker appears to follow the location, scale and rotation of the object more closely than other trackers.

5 Conclusions

A new tracking scheme based on structural similarity has been developed and has been shown to perform reliably under difficult conditions (as often occurs in surveillance vi-

deos), when tested with real-world video sequences. Robust performance has been demonstrated in both low and variable light conditions, and in the presence of spurious or camouflaged objects. In addition, the algorithm copes well with the artefacts that may be introduced by a human operator, such as rapid changes in camera view angle and zoom. This is achieved with relatively low computational complexity, which makes our algorithm potentially applicable to real-time surveillance problems.

Colour cue itself cannot provide stable tracking under changing illumination and when regions with similar colour are present in the scene. Likewise, the combined colour and edge cue, although generally performing well, cannot provide a reliable and precise tracking performance under ambiguous situations, especially with a moving camera (with changes in the pan, tilt and zoom). In contrast, the proposed PF based on the structural similarity measure exhibits the most stable and reliable performance. This is due to the fact that this measure captures the spatial similarity between the regions of interest, independently of the colour. It measures only relative changes in contrast and luminance, and thus is more robust to the changes in the environment. The implemented tracking algorithm uses a changeable size of the tracking window, which makes it suitable for many real-world applications (where the camera–object distance varies significantly).

Future work on the proposed PF tracking approach will be focussed on further improvement of its performance by realising the potential of giving adaptive weights to the three components of the similarity measure in (7). The structural similarity measure-based tracker may not be robust to significant alteration of the tracked object. Thus, the recovery and/or template update techniques will also be investigated in the future to improve reliability of the proposed tracker.

Acknowledgements The authors are grateful to the financial support by the UK MOD Data and Information Fusion Defence Technology Centre, by projects 2.1 'Image and video sensor fusion', 2.2 'Communication optimisation for distributed sensor systems', and the Tracking Cluster project DIF-DTC/CSIPC1/02. The authors would also like to thank Dr Henry Knowles for making his implementation of the Mean-Shift algorithm available to us.

References

1. Aherne, F., Thacker, N., Rockett, P.: The Bhattacharyya metric as an absolute similarity measure for frequentcy coded data. Kybernetica **32**(4), 1–7 (1997)
2. Arulampalam, M., Maskell, S., Gordon, N., Clapp, T.: A tutorial on particle filters for online nonlinear/non-Gaussian Bayesian tracking. IEEE Trans. on Signal Proc. **50**(2), 174–188 (2002)
3. Brasnett, P., Mihaylova, L., Bull, D., Canagarajah, N.: Sequential Monte Carlo tracking by fusing multiple cues in video sequences. Image and Vision Computing **25**(8), 1217–1227 (2007)
4. Brasnett, P., Mihaylova, L., Canagarajah, N., Bull, D.: Particle filtering with multiple cues for object tracking in video sequences. In: Proc. of SPIE's 17th Annual Symposium on Electronic Imaging, Science and Technology, V. 5685, pp. 430–441 (2005)
5. Comaniciu, D., Meer, P.: Mean shift: A robust approach toward feature space analysis. IEEE Transactions on Pattern Analysis and Machine Intelligence **22**(5), 603–619 (2002)
6. Comaniciu, D., Ramesh, V., Meer, P.: Kernel-based object tracking. IEEE Trans. on Pattern Analysis and Machine Intelligence **25**(5), 564–577 (2003)
7. Cvejic, N., Nikolov, S.G., Knowles, H.D., Loza, A., Achim, A., Bull, D.R., Canagarajah, C.N.: The effect of pixel-level fusion on object tracking in multi-sensor surveillance video. In: Computer Vision and Pattern Recognition, 2007. CVPR '07. IEEE Conference on, pp. 1–7 (2007)
8. Doucet, A., Freitas, N., N. Gordon, E.: Sequential Monte Carlo Methods in Practice. New York: Springer-Verlag (2001)
9. Foresti, G., Micheloni, C., Snidaro, L., Remagnino, P., Ellis, T.: Active video-based surveillance system: the low-level image and video processing techniques needed for implementation. IEEE Signal Processing Magazine **22**(2), 25– 37 (2005)
10. Forsyth, D., Arikan, O., Ikemoto, L., Ramanan, D.: Computational Studies of Human Motion: Part 1, Tracking and Motion Synthesis. Foundations and Trends in Computer Graphics and Vision. Hanover, Massachusetts. Now Publishers Inc. (2006)
11. Hampapur, A., Brown, L., Connell, J., Ekin, A., Haas, N., Lu, M., Merkl, H., Pankanti, S.: Smart video surveillance: Exploring the concept of multiscale spatiotemporal tracking. IEEE Signal Processing Magazine **22**(2), 38–51 (2005)
12. Hu, W., Tan, T., Wang, L., Maybank, S.: A survey on visual surveillance of object motion and behaviors. IEEE Trans. on Systems, Man, and Cybernetics – Part C: Applications and Reviews **34**(3), 334–352 (2004)
13. Isard, M., Blake, A.: Contour tracking by stochastic propagation of conditional density. In: European Conf. on Computer Vision, pp. 343–356. Cambridge, UK (1996)
14. Isard, M., Blake, A.: Condensation – conditional density propagation for visual tracking. Intl. J. of Computer Vision **28**(1), 5–28 (1998)
15. Lewis, J.J., Nikolov, S.G., Loza, A., Canga, E.F., Cvejic, N., Li, J., Cardinali, A., Canagarajah, C.N., Bull, D.R., Riley, T., Hickman, D., Smith, M.I.: The Eden project multi-sensor data set. In: Technical Report TR-UoB-WS-Eden-Project-Data-Set. Available at http://www.imagefusion.org (2006)
16. Lewis, J.J., O'Callaghan, R.J., Nikolov, S.G., Bull, D.R., Canagarajah, N.: Pixel- and region-based image fusion with complex wavelets. Information Fusion **8**(2), 119–130 (2007)
17. Liu, J., Chen, R.: Sequential Monte Carlo methods for dynamic systems. Journal of the American Statistical Association **93**(443), 1032–1044 (1998)
18. Loza, A., Dixon, T.D., Canga, E.F., Nikolov, S.G., Bull, D.R., Canagarajah, C.N., Noyes, J.M., Troscianko, T.: Methods of fused image analysis and assessment. In: Proc. of the Advanced Study Institute Conference, Albena, Bulgaria, 16–27 May, pp. 252–259 (2005)
19. Loza, A., Mihaylova, L., Bull, D.R., Canagarajah, C.N.: Structural similarity-based object tracking in video sequences. In: Proc. of the 9th International Conf. on Information Fusion, Florence, Italy, 10-13 July (2006)
20. McKenna, S., Jabri, S., Duric, Z., Rosenfeld, A., Wechsler, H.: Tracking groups of people. Computer Vision and Image Understanding **80**(1), 42–56 (2000)
21. Mihaylova, L., Loza, A., Nikolov, S.G., Lewis, J., Canga, E.F., Li, J., Bull, D.R., Canagarajah, C.N.: The influence of multi-sensor video fusion on object tracking using a particle filter. In: Proc. of the 2nd Workshop on Multiple Sensor Data Fusion, Dresden, Germany,2–6 October, pp. 354–358 (2006)
22. Moreno-Noguer, F., Sanfeliu, A., Samaras, D.: Integration of dependent Bayesian filters for robust tracking. In: Proceedings of the 2006 International Conference on Robotics and Automation, pp. 4081–4067 (2006)
23. PETS 2006 benchmark data. Dataset available on-line at: http://www.pets2006.net (2006)
24. Nummiaro, K., Koller-Meier, E., Gool, L.V.: An adaptive color-based particle filter. Image and Vision Computing **21**(1), 99–110 (2003)

25. Nummiaro, K., Koller-Meier, E.B., Gool, L.V.: A color-based particle filter. In: Proc. of 1st Intl. Workshop on Generative-Model-Based Vision GMBV'02, in conjuction with ECCV'02, pp. 1:53–60 (2002)
26. PerceptiVU, Inc.: Target Tracking Movie Demos. Http://www.perceptivu.com
27. Pérez, P., Vermaak, J., Blake, A.: Data fusion for tracking with particles. Proceedings of the IEEE **92**(3), 495–513 (2004)
28. Piella, G., Heijmans, H.: A new quality metric for image fusion. In: Image Processing, 2003. ICIP 2003. Proceedings. 2003 International Conference on, vol. 3, pp. III–173–6 vol.2 (2003)
29. Ristic, B., Arulampalam, S., Gordon, N.: Beyond the Kalman Filter: Particle Filters for Tracking Applications. Artech House (2004)
30. Shen, C., van den Hengel, A., Dick, A.: Probabilistic multiple cue integration for particle filter based tracking. In: Proc. of the VIIth Digital Image Computing : Techniques and Applications. C. Sun, H. Talbot, S. Ourselin, T. Adriansen, Eds. (2003)
31. Tao, D., Li, X., Wu, X., Maybank, S.: Human carrying status in visual surveillance. In: IEEE Int. Conf. on Computer Vision and Pattern Recognition, vol. 2, pp. 670–677 (2006)
32. Tao, D., Li, X., Wu, X., Maybank, S.J.: General tensor discriminant analysis and Gabor features for gait recognition. Pattern Analysis and Machine Intelligence, IEEE Transactions on **29**(10), 1700–1715 (2007)
33. Toyama, K., Hager, G.: Incremental focus of attention for robust vision-based tracking. Intl. J. of Computer Vision **35**(1), 45–63 (1999)
34. Triesch, J.: Self-organized integration of adaptive visual cues for face tracking. In: In: Sensor Fusion: Architectures, Algorithms, and Applications IV, Belur V. Dasarathy, Editor, Proceedings of SPIE Vol. 4051, pp. 397–406 (2000)
35. Wan, E., van der Merwe, R.: The Unscented Kalman filter. In: S. Haykin (ed.) Kalman Filtering and Neural Networks, chap. 7, pp. 221–280. Wiley Publishing (2001)
36. Wang, X., Xiao, B., Ma, J.F., Bi, X.L.: Scaling and rotation invariant analysis approach to object recognition based on radon and fourier-mellin transforms. Pattern Recognition **40**(12), 3503–3508 (2007)
37. Wang, Z., Bovik, A., Sheikh, H., Simoncelli, E.: Image quality assessment: from error visibility to structural similarity. IEEE Trans. on Image Processing **13**(4), 600–612 (2004)
38. Wang, Z., Bovik, A.C., Simoncelli, E.P.: Structural approaches to image quality assessment. In: A. Bovik (ed.) Handbook of Image and Video Processing, 2nd Edition, chap. 8.3. Academic Press (2005)
39. Webb, A.: Statistical Pattern Recognition. John Wiley & Sons (2003)