SURVEILLANCE VIDEO OBJECT TRACKING WITH DIFFERENTIAL SSIM

Fanglin Wang[a,], Jie Yang[a]*, Xiangjian He[b], Artur Loza[c]*

[a] Institute of Image Processing and Pattern Recognition, Shanghai Jiao Tong University, 200240 China - (hardegg, jieyang)@sjtu.edu.cn
[b] Faculty of Engineering and Information Technology, University of Technology, Sydney – sean@it.uts.edu.au
[c]Department of Electrical and Electronic Engineering, University of Bristol, UK - artur.loza@bristol.ac.uk

KEY WORDS: Tracking, Structural similarity, Gradient ascent

ABSTRACT:

The recently proposed use of the structural similarity measure, in the particle filter-based video tracker has been shown to improve the tracking performance, compared to similar methods using the colour or edge histograms and Bhattacharyya distance. However, the combined use of the structural similarity and a particle filter results in a computationally complex tracker that may not be suitable for some real time applications. In this paper, a novel fast approach to the use of the structural similarity in video tracking is proposed. The tracking algorithm presented in this work determines the state of the target (location, size) based on the gradient ascent procedure applied to the structural similarity surface of the video frame, thus avoiding computationally expensive sampling of the state space. The new method, while being computationally less expensive, performs better, than the standard mean shift and the structural similarity particle filter trackers, as shown in exemplary surveillance video sequences.

1. INTRODUCTION

The development and increased availability of video technology has in recent years inspired a large amount of work on the problem of object tracking in video sequences (A. Hampapur et al, 2005; G. Alper Yilmaz et al, 2006). One of the important tracking applications are the surveillance systems, utilised in a wide range of environments such as: transport systems, public spaces (shopping malls, car parks, etc.), industrial environments, government or military establishments. Employed in such diverse scenarios, a video tracking system faces numerous challenges. In this correspondence we focus on two broad video tracking issues: robust object representation and computational complexity. The objects found in surveillance videos are often being tracked in ’difficult’ environments characterised by the variable visibility (e.g. shadows, occlusions) and the presence of spurious (e.g. similarly-coloured) objects and backgrounds. The ability to track the object in challenging conditions is affected by, among others, the object representation and the image features utilised. Examples of the commonly used region-based object representations are the colour (D. Comaniciu, P. Meer, 2002; K. Nummiaro et al, 2002) and edge histograms (Paul Brasnett et al, 2007) combined with the Bhattacharyya distance. This type of an object representation has been demonstrated to be relatively robust to rotation, shape changes and partial occlusion. On the other hand, it discards completely the spatial information of the image. A. Loza et al(A. Loza et al, 2009) proposed that the Structural SIMilarity Image Quality Index (SSIM) (Zhou Wang et al, 2004) is used to measure the similarity between the target and the candidate regions, based on local luminance, contrast and structure comparison. The tracker, referred to as the SSIM-PF, has proved to be robust to varying light conditions and the presence of the spurious objects/background.

Moreover, due to temporal limitations of the real-time tracking systems, the complexity of the tracking algorithms is of a special importance. In real world applications the speed of the tracking process is affected by, among others, the complexity of the tracking algorithm and the number of objects present in the scene. In this paper we concentrate on a single object tracking and thus the reduction of the computational complexity of the tracking algorithm is our priority. Tracking, i.e. target size and localisation in successive frames, is performed by solving a state-space optimization problem, and both probabilistic and deterministic approaches have been proposed in the past. Some of the most popular probabilistic methods are based on Particle Filter (PF) (K. Nummiaro et al, 2002; M. Isard, A. Blake, 1998) valued for its ability to deal with nonlinear and non-Gaussian estimation problems. PF is a Monte Carlo approach approximating the state space distributions based on their random samples (particles). However, the computational complexity of PF is approximately proportional to the number of particles, and in many cases, the resulting computational load prohibits real-time application. On the other hand, deterministic optimisation methods, although less flexible, are usually less complex than the probabilistic approaches. Among the deterministic techniques, the Mean Shift (MS) algorithm (D. Comaniciu, P. Meer, 2002) is a widely used and a relatively fast adaptive tracking procedure that finds the maximum of the Bhattacharyya coefficient. Other ways of finding the mode of similarity measure in order to localise an object in a video frame, include Differential Earth Movers Distance proposed by Qi Zhao et al(Qi Zhao et al, 2007). Therein, a fast differential formula is proposed to analyse the similarity between the colour distributions of the object template and that of the candidate object.

* Corresponding author. jieyang@sjtu.edu.cn, artur.loza@bristol.ac.uk

In this paper, motivated by the use of the structural information for surveillance video tracking as proposed by A. Loza et al(A. Loza et al, 2009) and the differential treatment of the distance measures (Qi Zhao et al, 2007), a new SSIM-based tracking algorithm is proposed. Unlike A. Loza' method(A. Loza et al, 2009), our method avoids the computationally expensive process of computing the similarity measure at numerous locations in state space, by deriving a gradient ascent method to localize the mode of the SSIM. The algorithm, referred to as the Differential SSIM (DSSIM), while benefiting from the robustness of the original measure to some image distortions, analyses the target–frame similarity based on the gradient of the image by using a differential form of SSIM. This simple form of local optimization of the similarity, while resulting in very good tracking performance in the test video sequences, is shown to be applicable to real time scenarios, due to its reduced computational complexity.

The remainder of this paper is organized as follows. Section 2 presents a description of our approach, including a brief SSIM review, the derivation of the DSSIM, and the corresponding tracking algorithm. The performance and efficiency of the proposed approach is demonstrated in Section 3. Section 4 presents the conclusions and discusses the open issues for future research.

2. DIFFERENTIAL SSIM TRACKER

2.1 Structural similarity measure

A region-based tracking algorithm typically compares the current frame region, I, with the object template, J, by means of a distance or similarity measure. A recently proposed image quality index, SSIM, used in our method, is defined as follows (Zhou Wang et al, 2004)

$$S = \left(\frac{2\mu_I\mu_J + C_1}{\mu_I^2 + \mu_J^2 + C_1}\right)\left(\frac{2\sigma_I\sigma_J + C_2}{\sigma_I^2 + \sigma_J^2 + C_2}\right)\left(\frac{\sigma_{IJ} + C_3}{\sigma_I\sigma_J + C_3}\right), \quad (1)$$

where $C_{1,2,3}$ are small positive constants used for the numerical stability purposes. For $C_3 = C_2/2$, (1) can be simplified to obtain

$$S = \left(\frac{2\mu_I\mu_J + C_1}{\mu_I^2 + \mu_J^2 + C_1}\right)\left(\frac{2\sigma_{IJ} + C_2}{\sigma_I^2 + \sigma_J^2 + C_2}\right), \quad (2)$$

where μ, σ and σ_{IJ} denote the sample mean, standard deviation and covariance, respectively:

$$\begin{aligned} \mu_I &= \frac{1}{N}\sum_{i=1}^{N} I_i, \qquad \sigma_I^2 = \frac{1}{N-1}\sum_{i=1}^{N}(I_i - \mu_I)^2 \\ \sigma_{IJ} &= \frac{1}{N-1}\sum_{i=1}^{N}(I_i - \mu_I)(J_i - \mu_J)\,. \end{aligned} \quad (3)$$

I_i and J_i, denote the pixel intensity of the current and template image regions, respectively. S is symmetric and maps the similarity between two images to the interval (−1, 1] : $S = 1$ iff $I = J$.

This similarity measure has been selected based on its good ability to capture perceptual similarity of images. The SSIM measure simulates the perceptual process of the human visual system by measuring the luminance, contrast and structural similarity of the two images, calculated by the first, second and third term in (1), respectively. Another important feature of the SSIM index is that the three normalised measurements in (1), are sensitive to the relative rather than absolute image distortions (Zhou Wang et al, 2004), thus making this measure suitable to video tracking in varying conditions. The SSIM measure was first successfully applied to particle-filter video object tracking (A. Loza et al, 2009). Therein it was also demonstrated that the structure comparison is more reliable in scenarios when spurious (e.g. camouflaged) objects appear in the scene or when there is not enough discriminative colour information available.

2.2 Differential SSIM Tracking Algorithm

In the PF framework, as proposed by A. Loza et al(A. Loza et al, 2009), the SSIM is computed for each particle. This makes the SSIM-PF method computational expensive for large number of particles. In order to achieve a computationally efficient tracking performance, whilst retaining the benefits of the original measure, a differential SSIM formula is proposed as follows. The object is tracked in the spatial domain of the subsequent video frames by maximising the measure (2), based on its gradient. In order to simplify the subsequent derivation, we choose to analyse the logarithm of (2) by defining a function *ρ(x)*:

Table 1. Pseudocode of the proposed DSSIM tracking algorithm

```
Input:  target location $\mathbf{x}_{k-1}$, previous frame
Output:  target location $\mathbf{x}_k$, current frame
• $k = 0$, Initialise $\mathbf{x}_0$
• FOR $k = 1 : K_{\text{frames}}$
  - Initialise $\mathbf{x}_k^0 = \mathbf{x}_k^1 = \mathbf{x}_{k-1}$
  - WHILE $S(\mathbf{x}_k^1) \geq S(\mathbf{x}_k^0)$
    * Assign $\mathbf{x}_k^0 = \mathbf{x}_k^1$
    * Calculate $\nabla\rho(\mathbf{x}_k^0)$ according to (11)
    * Assign $\mathbf{x}_k^1$ the location of a pixel in $\mathbf{x}_k^0$
      8-connected neighbourhood, along the
      direction of $\nabla\rho(\mathbf{x}_k^0)$
  - END WHILE
  - Assign $\mathbf{x}_k = \mathbf{x}_k^0$
• END FOR
```

$$\rho(\mathbf{x}) = s\log(\mathrm{abs}\,(S(\mathbf{x}))) \quad (4)$$

where *S(x)* denotes the similarity between the object template *J* and a current frame image region *I* centred around the pixel location *x = (x, y)* and *s = sign (S(x)).* The null values of *S* are handled by increasing value of C_2. After a simple expansion of (4) we obtain the expression for the gradient of the function *ρ(x)*

$$\nabla\rho(\mathbf{x}) = s\left(A_1\nabla\mu_I + A_2\nabla\sigma_I^2 + A_3\nabla\sigma_{IJ}\right), \quad (5)$$

where

$$A_1 = \frac{2\mu_J}{2\mu_I\mu_J + C_1} - \frac{2\mu_I}{\mu_I^2 + \mu_J^2 + C_1}, \quad (6)$$

$$A_2 = -\frac{1}{\sigma_I^2 + \sigma_J^2 + C_2}, \quad A_3 = \frac{1}{2\sigma_{IJ} + C_2}\,. \quad (7)$$

The gradients $\nabla_{\mu I}$ and $\nabla^{2}{}_{I}$ can be calculated as follows

Table 2. The performance evaluation measures of the tracking simulations

Seq. name	Image size (pixels)	Template size (pixels)	Speed (fps)			Mean RMSE (pixels)			std RMSE (pixels)		
			MS	SSIM-PF	DSSIM	MS	SSIM-PF	DSSIM	MS	SSIM-PF	DSSIM
cross	720 × 576	54 × 152	27	13	71	37.3	8.3	5.6	62.2	5.1	4.3
man	320 × 240	20 × 46	109	53	315	18.2	8.0	7.0	13.1	6.5	4.9
otcbvs	320 × 240	10 × 21	408	373	1056	29.4	1.9	1.0	15.8	0.8	0.6

$$\nabla\mu_I = \frac{1}{N}\sum_{i=1}^{N}\nabla I_i \,, \tag{8}$$

$$\nabla\sigma_I^2 = \frac{2}{N-1}\sum_{i=1}^{N}(I_i-\mu_I)\nabla I_i \,. \tag{9}$$

A simplified expression for the covariance gradient, ∇_{IJ}, can be obtained, based on the observation that $\sum^{N}_{i=1}(J_i - \mu_J) = 0$:

$$\begin{aligned}\nabla\sigma_{IJ} &= \frac{1}{N-1}\sum_{i=1}^{N}(J_i-\mu_J)(\nabla I_i - \nabla\mu_I)\\ &= \frac{1}{N-1}\sum_{i=1}^{N}(J_i-\mu_J)\nabla I_i\end{aligned} \tag{10}$$

Finally, by defining the gradient of the pixel intensity as $\nabla I_i = (\frac{\partial I_i}{\partial_x}, \frac{\partial I_i}{\partial_y})$, the complete formula for *ρ(x)* is obtained

$$\nabla\rho(\mathbf{x}) = s\sum_{i=1}^{N}\left(\frac{A_1}{N} + \frac{2A_2(I_i-\mu_I)+A_3(J_i-\mu_J)}{N-1}\right)\nabla I_i \,. \tag{11}$$

The proposed algorithm, employing the gradient DSSIM function (11) is summarised in Table 1. In general terms, the estimated target location, x^0_k is moved along the direction the structural similarity gradient by one pixel in each iteration until no further improvement is achieved. The number of SSIM and gradient evaluations depends on the number of iterations needed to find the maximum of the measure *S(x)* and on average does not exceed 5 in our experiments. This makes our approach significantly faster than the original SSIMPF. It should be noted that although the differential framework of the algorithm is based on a reformulation of the scheme proposed by Qi Zhao et al (Qi Zhao et al, 2007), it utilises a distinct similarity measure.

3. EXPERIMENTAL RESULTS

The performance of our method is demonstrated in this section by tracking objects in real-world surveillance video sequences. The DSSIM algorithm has been implemented in C++ programming language and compared with the MS tracker (D. Comaniciu, P. Meer, 2002) and the SSIM-PF method (A. Loza et al, 2009) (see also Section 1 and 2 for a brief description of the methods). The MS tracker, similarly to DSSIM, is a non-parametric technique relying on finding the modes of the underlying target–current frame pdf in the feature space, however, it is based on a different principle and uses the colour histogram. For both MS and DSSIM, the popular scheme for scale adaptation, i.e., varying the object size by 5% and choosing the size giving the highest similarity, has been utilised.

Our simulations consist in tracking a pre-selected object (person) in the following three video sequences. The sequence *cross* (5 sec), taken from the database, contains three people walking rapidly in front of a stationary camera. The tracked region (a person) has similar colour to that of the background and the passers-by. One of the passers-by occludes temporarily the object. The second sequence, *man* (40 sec), is a long recording showing a person walking along a car park. Apart from object's colour similarity to the nearby cars and the shadowed areas, the video contains numerous instabilities, resulting from a shaking camera, fast zoom-ins and zoom-outs, and a wide range of a view angle. The last sequence, *otcbvs*, is a part of a multimedia benchmark dataset collection (J. Davis and V. Sharma). The sequence used in this paper is a colour video of a busy patio, recorded from approximately 3 stories above ground. The small-sized tracked object (see Table 2 for the target sizes) undergoes significant intensity changes as it enters the shadowed areas of the walkway and the entrance of a building.

In order to numerically evaluate the performance of the developed technique, Root Mean Square Error (RMSE) has been used:

$$\mathrm{RMSE}(k) = \left(\frac{1}{M}\sum_{m=1}^{M}(x_k-\hat{x}_{k,m})^2+(y_k-\hat{y}_{k,m})^2\right)^{\frac{1}{2}} \tag{12}$$

where $(\hat{x}_{k,m}, \hat{y}_{k,m})$ and for the upper-left corner coordinates of the tracking box determined by both the object's central position and the scale estimated by the tracker in the frame k. The corresponding ground truth positions of the object, (x_k, y_k), have been generated by manually tracking the object. The mean of RMSE and its standard deviation (std) are presented in Table 2, while the frame-to-frame error plots are shown in Figure 1.

Based on the performance measures in Table 2, it has been concluded that DSSIM outperforms the MS and SSIM-PF, both in terms of the processing speed and the accuracy. It also appears to be more stable than the other two methods (lowest std). Although the error plots in Figure 1 show that in a number of frames the methods perform comparably, it can be seen that in the remaining frames our method achieves the best performance most of the time. Although the difference between the accuracy and the stability of SSIM-PF and DSSIM is not large in some cases, in terms of the computational complexity, our method compares much more favourably with SSIM-PF, as well as with MS. The average tracking speed estimates were computed on PC in the following setup: CPU clock 2.66 GHZ, 1G RAM, MS and DSSIM requiring on average 8 and 5 iterations, respectively, and PF using 100 particles. In terms of the relative computational efficiency, the proposed method has been found to be approximately 3–6 times faster than SSIM-PF and up to 2 times faster than MS (our implementation).

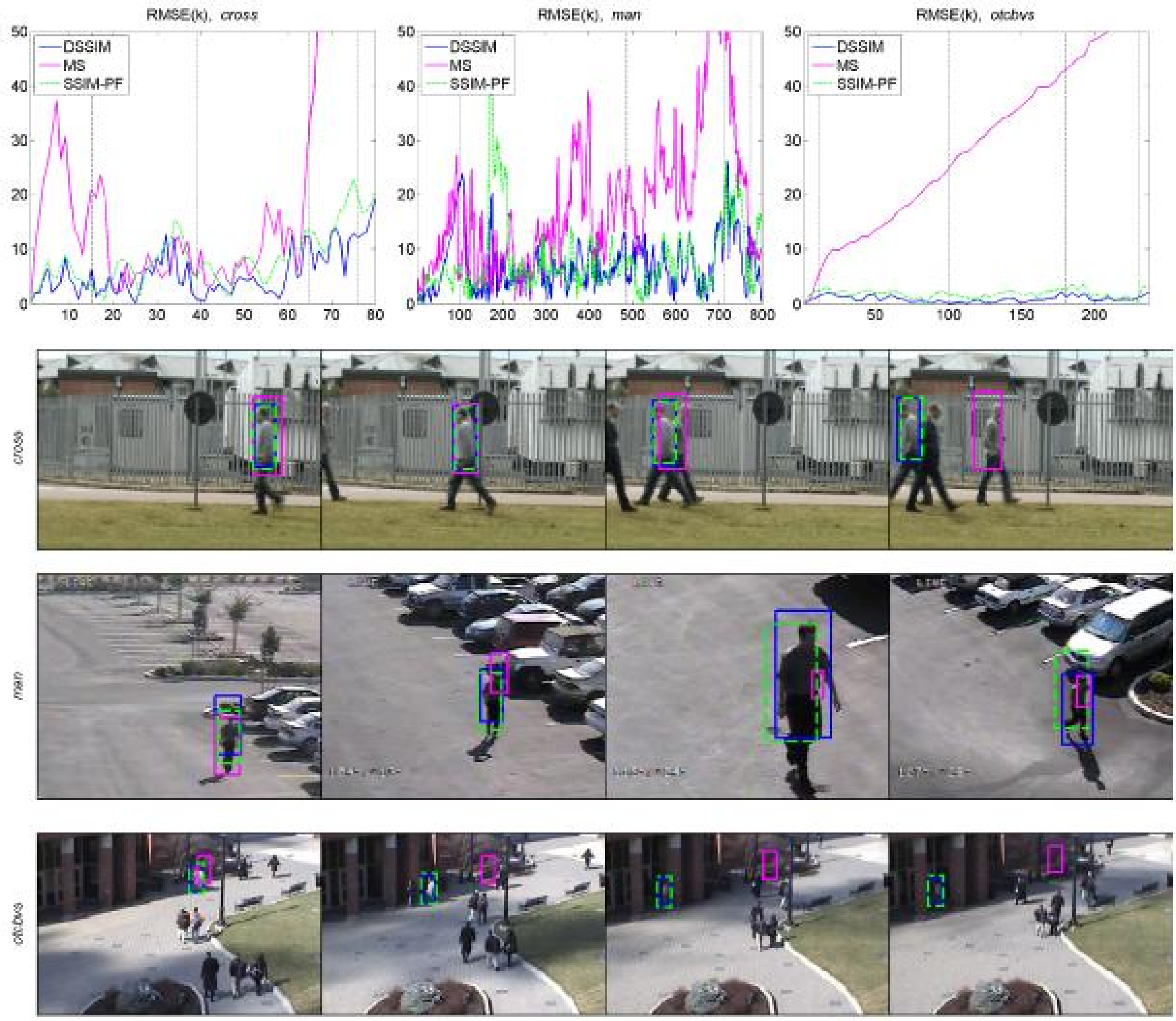

Fig. 1. Top: RMSE plots for the tested trackers. Bottom: Exemplary video frames (marked in the RMSE plots) with tracker output.

The exemplary frames in Figure 1, where the 'difficult' frames have been selected, offer more insight into the performance and robustness of the trackers. In the *cross* sequence, both SSIM-PF and DSSIM are not distracted by the temporary occlusion of the tracked person by other passer-by, whereas the MS tracker locks onto a similar object moving in the opposite direction. In *man* sequence, although all the three trackers manage to follow the target, the proposed method identifies the scale and the position of the object with the best accuracy. Unlike the colour-based MS, both SSIM-based methods track the object throughout the *otcbvs* sequence demonstrating the robustness to the illumination changes.

4. CONCLUSIONS

This work introduces a novel and robust tracking algorithm, DSSIM, in which the distance between the target and candidate is measured by the structural similarity index. The main theoretical contribution in this work is the development of a fast differential algorithm for locally optimal search of the structural similarity surface. The proposed method performs reliably in the exemplary videos under difficult conditions, often occurring in surveillance scenarios: temporal occlusions, nonstationarity of the camera, presence of the spurious objects and illumination changes. The DSSIM has been shown to outperform other deterministic tracking method, mean-shift, and the structural similarity-based PF tracker in terms of the accuracy. Another advantage of the proposed algorithm is its low computation complexity owing to the fast differential algorithm derived in this work, which makes DSSIM applicable to real-time video tracking. Our future investigation will be focused on reliability improvement of the methods, by addressing the local maxima issue of the gradient ascent and by development a template update scheme.

References

A. Hampapur, L. Brown, J. Connell, A. Ekin, N. Haas, M. Lu, H. Merkl, and S. Pankanti, Smart video surveillance: Exploring the concept of multiscale spatiotemporal tracking, *IEEE Signal Processing Mag.*, vol. 22, no. 2, pp. 38–51, Mar. 2005.

G. Alper Yilmaz, Omar Javed, and Mubarak Shah, Object tracking:A survey, *ACM Comput. Surv.*, vol. 38, no. 4, 2006.

D. Comaniciu and P. Meer, Mean shift: A robust approach toward feature space analysis, *IEEE Trans. on Pattern Analysis and Machine Intelligence*, vol. 22, no. 5, pp. 603–619, 2002.

K. Nummiaro, E. B. Koller-Meier, and L. Van Gool, A color-based particle filter, *in Proc. of 1st Intl. Workshop on Generative-Model-Based Vision GMBV'02*, 2002, pp. 53–60.

Paul Brasnett, Lyudmila Mihaylova, David Bull, and Nishan Canagarajah, Sequential Monte Carlo tracking by fusing multiple cues in video sequences, *Image Vision Comput.*, vol. 25, no. 8, pp. 1217–1227, 2007.

A. Loza, L. Mihaylova, D. R. Bull, and C. N. Canagarajah, Structural similarity-based object tracking in multimodality surveillance videos, *Machine Vision and Applications*, vol. 20, no. 2, pp. 71–83, Feb. 2009.

Zhou Wang, A.C. Bovik, H.R. Sheikh, and E.P. Simoncelli, Image quality assessment: from error visibility to structural similarity, *IEEE Trans. on Image Processing*, vol. 13, no. 4, pp. 600–612, Apr. 2004.

M. Isard and A. Blake, Condensation – conditional density propagation for visual tracking, *Intl. J. of Computer Vision*, vol. 28, no. 1, pp. 5–28, 1998.

Qi Zhao, S. Brennan, and Hai Tao, Differential EMD tracking, *in Proc. of IEEE 11th International Conference on Computer Vision*, Oct. 2007, pp. 1–8.

Acknowledgement: This research is partly supported by International research project (No. 2009DFA12870), Ministry of Science and Technology, China.