Defence Science Journal, Vol. 68, No. 3, May 2018, pp. 307-315, DOI : 10.14429/dsj.68.12209


# Robust Stabilised Visual Tracker for Vehicle Tracking

Kamlesh Verma[#,*], Avnish Kumar[#], and Debashis Ghosh[@]

[#]*DRDO-Instruments Research and Development Establishment, Dehradun - 248 008, India*
[@]*Department of Electronics and Communication Engineering, Indian Institute of Technology Roorkee - 247 667, India*
[*]*E-mail: kamlesh@irde.drdo.in*

**ABSTRACT**

Visual tracking is performed in a stabilised video. If the input video to the tracker algorithm is itself destabilised, incorrect motion vectors will cause a serious drift in tracking. Therefore video stabilisation is must before tracking. A novel algorithm is developed which simultaneously takes care of video stabilisation and target tracking. Target templates in just previous frame are stored in positive and negative repositories followed by Affine mapping. Then optimised affine parameters are used to stabilise the video. Target of interest in the next frame is approximated using linear combinations of previous target templates. Proposed modified L1 minimisation method is used to solve sparse representation of target in the target template subspace. Occlusion problem is minimised using the inherent energy of coefficients. Accurate tracking results have been obtained in destabilised videos.

**Keywords:** Digital video stabilisation; Visual tracking; Occlusion handling; Modified L1 minimisation

## 1. INTRODUCTION

Visual surveillance and tracking information for military operations has always been a prime requirement in all weather, day and night conditions. For doing reconnaissance, three basic electro-optical sensors are needed, i.e. day camera for viewing in day condition, thermal camera for viewing in night condition and eye-safe laser range finder for obtaining the target distance[1]. These three basic sensors are packaged in a mechanical housing called 'gimbal'. This gimbal is line-of-sight stabilised and the whole system is called Stabilised Sighting System (SSS) as shown in Fig. 1. A multifunction display (MFD) is integrated to select the video among day camera, thermal camera and tracked video. When this SSS is integrated on a vehicle, then due to the vehicle platform vibrations, engine vibrations and aerodynamic disturbances, output video is destabilised and shaky which results in reduced tracking accuracy with drift. Hence, there are two major requirements for successful tracking with better accuracy, viz. stabilisation and tracking. This paper proposes an effective and robust solution to meet these two requirements simultaneously in one single algorithm.

Various disturbances may occur due to platform dynamics, engine vibrations, unintentional movements in yaw, roll and pitch directions, aerodynamics etc. Three techniques may be employed to stabilise the video. First technique is optical image stabilisation, wherein a fluid-based optical component is used, which is steered in the opposite direction of disturbance movement. This technique is effective for optical telescope and corrects the low amplitude atmospheric turbulences. Second method is electronic stabilisation in which all electro-optical sensors are mounted in a gimbal, and a rate sensor (gyro) is used to detect the rate of disturbances. To design the compensators for this, frequency response of the gimbal is taken, followed by Bode curve shaping for desired control parameters. The compensator may be a typical PID controller, implemented in the form of an infinite impulse response (IIR) filter on a digital signal processing (DSP) board. A Finite Impulse Response (FIR) filter may also be used, but it will add to the computational cost. Also FIR filter is not required unless a linear phase response is desired. Visual tracking command is given as the input to the block as shown in Fig. 2. However, this scheme demands costly mechanical gimbal and rate sensors. Also, modifications/change in any electro-optical sensor will need new electronics hardware design and hence, is not a flexible method. This scheme results in a coarse stabilisation,

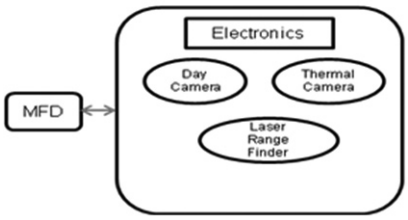


**Figure 1. Gimbal.**

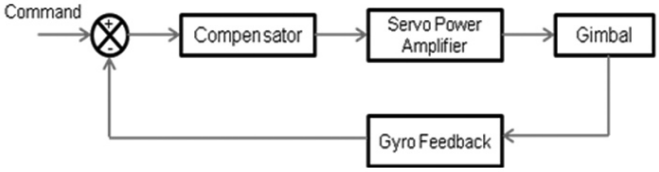


**Figure 2. Electronic video stabilisation.**

Received : 30 October 2017, Revised : 15 March 2018
Accepted : 27 March 2018, Online published : 16 April 2018

which may not be appropriate if accurate tracking is desired. This calls for a second level of stabilisation, termed as digital video stabilisation.

In digital video stabilisation, motion vectors are estimated in each video frame[1,4,5] then each frame is moved in a direction opposite to that of motion and finally all the motion compensated frames are combined to get stabilised video as shown in Fig. 3 and Fig. 4.

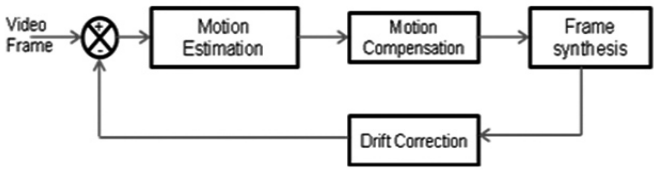


**Figure 3. Digital video stabilisation.**

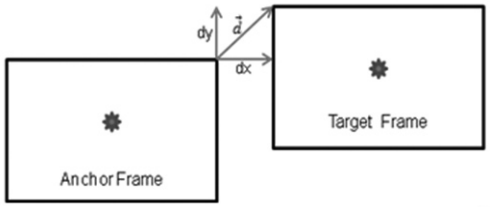


**Figure 4. Destabilised frame $\vec{d}$.**

This technique is very flexible, as it requires only digital video and all image processing is carried out using software. Once the input video is stabilised, tracking operation can be performed reliably. Due to the challenges involved in target detection and tracking[8-10], visual tracking has drawn the attention of many researchers in recent times. Some of the major challenges are scene convergence from 3D to 2D, complex motion of targets, non-rigid structure of objects, illumination variation, full or partial occlusion, real-time processing and noise. Further, in the defence scenario, a generic algorithm may not suit equally for land/air/naval systems as they operate in totally different environments. Also, in most of the existing techniques for video stabilisation, global motions are utilised for video stabilisation while local motions are used for target tracking. In this paper a novel method is proposed wherein local motion vectors are preserved during stabilisation, resulting in the rejection of aerodynamic disturbances even when tracking the target of interest.[13,28-31]

## 2. RELATED WORK

### 2.1 Digital Video Stabilisation

Various research work in the field of video stabilisation have been reported in the literature. An integral curve warping projection technique is used[1] to stabilise the video using a novel mechanism to correct very fine drift even upto 1/15th pixel. Global motion energy computation based technique is presented[2] in which maximum a posteriori estimation of global motion vectors is carried out. The motion path so obtained, is smoothened as a continuous curve over successive frames which finally provides the stabilised video. But this does not correct the 3D rotation. This problem was solved[3] where the authors have used a 3D rotation modelling using Riemannian geometry to smooth the global motion as a geodesic-convex constrained regression and calculated gradient and Hessian of the objective function. A block matching based algorithm is proposed[4] where those motion vectors which did not reliably contribute to the smoothness of global motion vector are discarded by ad-hoc rules based on local similarity. A voting based image alignment[5] performs the video stabilisation using a local motion estimator. A fast feature based algorithm detects potential features from each frame using features from accelerated segment test[6]. These features are then matched with consecutive frames using oriented binary robust independent elementary features. A[7] classified background feature point method is proposed to stabilise the video jointly with global motion estimation. This method is well suited when the moving foreground objects cover a very large area of the scene. However L1 tracker based video stabilisation method is used in the proposed algorithm which compensates the global motion to stabilise the video and preserves the local motion vector for tracking.

### 2.2 Visual Tracking

Once the video is stabilised, it is ready for visual tracking. Few extensive surveys on target tracking are available[8-10] that give a deep insight into target detection, representation, tracking methodologies with their advantages and disadvantages. Colour histogram and local steering kernel descriptors are used[11] to adaptively model the target for tracking, thereby providing robustness against scale and rotation changes, as well as for partial occlusion. Authors[12] proposed a multi-view support vector machines for robust representation of target appearance. The support vector is trained by three different features, viz. gray scale value, histogram of oriented gradients (HoG) and local binary pattern. Sparsity-induced subspace learning method is discussed[13] which exploits local mutual relations between future observations while learning. Combining different trackers addressing different tracking challenges is proposed[14].

Subsequently, two different Bayesian trackers are fused to get the robust tracker. Observation dependent hidden Markov model framework is used[15] where observation dependency was calculated by structure complexity between consecutive frames with updated object model. Visual tracking is achieved[16] by using locally structured Gaussian progress regression where spatially local structure is used for temporal targets and a joint representation with time series function is carried out for tracking. A relative discriminative HoG under the particle frame network is proposed[17] for surveillance, prediction and tracking of vehicle in traffic environment. This algorithm updates the object model in extreme brightness conditions thereby reducing the effect of tracking drift. Supervised[18] learning of discriminative classifiers is extended to online multiple instance learning (MIL) method where classifier separates foreground and background using features based on bag likelihood function. But this may yield poor result in case of complex background. The problem is solved[19], where more informative active features are selected by using Fisher criteria that measures the uncertainty in classification of object model. Another problem with[18] is that tracking may be slow

and less efficient if less important samples are present in the learning procedure. This problem is solved[20] by assigning sample importance (weight) while training classifier, called as weighted multiple instance learning (WMIL).

Online discriminative features (ODFS) is used[21] for real time target tracking where prior information of instance label for most correct positive instance is used, thereby minimising the drift in the MIL-based tracker. Here objective function is optimised in the steepest ascent direction, for positive samples while in the descent direction for negative samples. Real time compressive tracking is used[22] where important features are extracted from multi-scale image feature space with data independent basis using sparse measurement matrix for the appearance model. Online updation is carried out in compressed domain while visual tracking is performed by naive Bayes classifier. For multiple target tracking, a multi-path search is adopted[23]. Here track interactions were modeled like object spatial layout consistency or mutual occlusion resulting in optimised best view video synthesis output. A combination of local steering kernel descriptors and colour histogram is used[24] for target tracking. Colour histogram information separates the background from the foreground. This technique enables visual tracking under scale and rotation variations while solving partial occlusion problem also. If the target of interest lies in a cluttered background, then visual tracking method demands a hybrid strategy.

This is explained[25] where low amplitude descriptors are exploited to track the surrounding of object thereby helping in management of track trajectory. This technique is beneficial where multiple objects vary in their shapes. Multi-object detection and tracking in aerial videos is presented[26] where candidate targets are detected and extracted from tracking trajectory, followed by mean shift tracking fused with detection results to update the tracking models. A motion model based on temporal information of features is exploited[27] to track vehicles. Hidden Markov model is used to separate target vehicles from the background and tracking is carried out using probabilistic method. Mei[28], *et al.* have used sparse approximation of target candidate by solving L1 regularised least square problem under particle filter framework. To solve the minimisation problem, author used preconditioned conjugate gradients (PCG) method which is computationally heavy. A non-negativity constraint with dynamic template scheme was used by authors[29] to reduce the tracking drift under the same L1 minimisation problem. A bounded particle resampling L1 tracker was used[30] for further reduction of tracker drift problem. Considering above tracking methods, a novel method is presented here which solves the L1-regularisation problem with minimisation of energy of coefficients resulting in minimised tracking drift in stabilised frame. Zeng[31], *et al.* analyse the target of interest in two steps. In first step, discriminative model is used to separate the target from the cluttered background in a global approach method. Then using sparse representation of positive and negative target dictionaries, target is identified. In next step, authors use a set of filters and convolution is carried out between target template and each region of target providing the local features. Then final tracker is invoked using the integration of outputs of both steps.

### 2.3 Outcome of Related Work

Based upon the related work carried out in proposed fields, it can be inferred that visual tracker needs the local motion of target[8-31]. If the frame itself has a global motion due to destabilisation[1-7], then combined effect after vector addition of local and global motion vectors will cause a serious drift in the tracker and finally after few frames tracker will fail. This has motivated to develop an algorithm where robust Affine parameters drive the stabilisation and Sparse solution of L-1 tracker[28-31] needs modification to handle destabilisation.

## 3. PROPOSED VIDEO STABILISATION AND TRACKING METHOD

The problem at hand is to track a target of interest in a video captured by a sighting system camera mounted in a gimbal fixed on a vehicle platform; the video being jerky and destabilised due to several disturbances, as mentioned above. As a solution to the problem, this destabilised input video is to be stabilised followed by target tracking with good accuracy. To prove the algorithm, only gray level video is processed. If the processing needs to be carried out for colour video, then for each frame the proposed algorithm needs to be run for three times, i.e. red, green and blue colour matrices. Performance of algorithm will not be affected, but it will be slowed down if dedicated hardware is not used. FPGA is a good solution for colour image processing due to parallelism, if speed of processing is needed to be computed in real time. A target is identified in the first frame and a window of suitable area is placed around the target. This target identification may be done manually or automatically. In the case of automatic target selection, manual target selection in the very first frame need not to be considered. In the proposed work, the target is manually selected. Next, $n$ number of image templates are taken around this window. Three coordinates of window are defined, say $P_1$(top left), $P_2$ (bottom left) and $P_3$ (top right). These coordinates are randomly shifted by one pixel in all four directions, i.e. left, right, up and down, to generate $n$ number of templates. Video stabilisation and target tracking are performed simultaneously from the second frame onwards. The target in the next frame is called 'target candidate' which is represented as a linear combination of the above $n$ templates (both positive and negative templates are considered to take care of occlusion and to increase robustness of the algorithm). Now, target tracking boils down to finding the sparse approximation of target candidate in the template subspace. L1 minimisation tool is used to find the sparse solution while employing affine image warping motion model between two consecutive frames. To verify whether the input image sequence is stabilised or destabilised, motion vector calculated through affine solution is checked. A variation in the motion vector indicates that the target frame is destabilised by an amount $\vec{d}$ pixels, as shown in Fig. 4. Using the same computation, target frame is subsequently stabilised before tracking. It is to be noted here that local motion of the target is preserved during tracking. The L1 norm minimisation criterion is modified to achieve robustness. After fixing the target under the window with stabilisation, the image templates are updated for tracking. In our simulation experiments, the proposed algorithm was

compared with four other classic tracking algorithms and it was found that proposed algorithm outperforms others whenever there is video destabilisation for tracking.

### 3.1 Template Generation

The posterior distribution of random variables related to Markov chain may be estimated by particle filter method. For estimating the object in the next image sequence without any information on the observation probability of object, particle filter is an excellent tool. The target object is first predicted in nearby regions of the previous target location and then, after localising the target in the target frame, the coordinates of the object are updated. In our approach, the target of interest is initially located manually in the first frame of the input video. A rectangular window is placed around that target keeping in view that at least 90 per cent of the target is covered by that window. The three co-ordinates $P_1$, $P_2$ and $P_3$ of the window are then initialised in the first frame. Next, to extract the templates, a random value generator is used to shift the coordinates of $P_1$, $P_2$ and $P_3$ one at a time. The coordinate values are moved by one pixel in four possible directions, but one at a time, to generate one template. This provides with $n$ number of templates. If number of templates are increased, then computational cost will increase but on the other hand accuracy of target detection will increased. All the templates are put in a 1D array **T**. The candidate target in the next frame is obtained as the linear combination of these templates. It is to be noted that this algorithm will also work for infra red (IR) videos. Templates can be generated as described earlier. Further since IR videos are always in gray scale, so only one dimensional processing will be required.

### 3.2 Normalisation of Template

In order to take care of varying illumination during image acquisition, the templates are normalised by subtracting the mean pixel value from each pixel in a template and then dividing by the pixel variance. Let the template image is denoted by *TI*, mean value of *TI* as μ and variance $\sigma^2$. Then the normalised template image is given as

$$NTI = \frac{(TI - \mu)}{\sigma^2} \tag{1}$$

### 3.3 Affine Motion Model

The affine transformation parameters of each template are calculated. The state variable is defined as $x_t$, which is modeled by six affine motion parameters $(\beta_1, \beta_2, \beta_3, \beta_4, t_1, t_2)$ ,where $(\beta_1, \beta_2, \beta_3, \beta_4)$ are the deformation parameters and $(t_1, t_2)$ are the 2D position parameters. Three coordinates of each template are taken as input points $P_1(x_1, y_1)$, $P_2(x_2, y_2)$ and $P_3(x_3, y_3)$. These points will be transformed into three outpoint coordinates i.e. inpoints= transformation matrix x outpoints. If template has number of rows *r* and number of columns *c* then

$$\text{inpoints}=\begin{pmatrix} y_1 & y_2 & y_3 \\ x_1 & x_2 & x_3 \\ 1 & 1 & 1 \end{pmatrix}; \quad \text{outpoints}=\begin{pmatrix} 1 & r & 1 \\ 1 & 1 & c \\ 1 & 1 & 1 \end{pmatrix};$$

$$\text{outpoints}=[R]^{-1}.\text{inpoints} \tag{2}$$

The corresponding affine objects will be [*R*(1,1), *R*(1,2), *R*(2,1), *R*(2,2), *R*(1,3), *R*(2,3)]. Thus, $x_t$ is obtained. All templates are mapped into next frame templates through affine modelling.

### 3.4 Particle Filter Framework

The original affine parameters of manually selected target in first frame are called as first particle. The region of interest is cropped in the next frame. Then using the method outlined in Section 3.3, authors calculated affine parameters of templates of next frame and find the best-matched target coordinates, i.e. the second particle. This way the coordinates of all the particles (frames) are found. Say, that the observation of all previous frames, i.e. from first frame to frame $(t-1)$ is denoted as $\theta_{1:t-1} = \{\theta_1, \theta_2, ..., \theta_{t-1}\}$, where $\theta_i$ gives affine parameters of the target in the $i^{\text{th}}$ frame (particle). The predicting distribution of $x_t$ provides all observations upto time (*t*-1) denoted by $p(x_t \mid \theta_{1:t-1})$ which is recursively computed as

$$p(x_t \mid \theta_{1:t-1}) = \int p(x_t \mid x_{t-1})\, p(x_{t-1} \mid \theta_{1:t-1})\, dx_{t-1} \tag{3}$$

where $p(x_t \mid x_{t-1})$ is the state transition probability. The posterior probability is defined as

$$p(x_t \mid \theta_{1:t}) = \frac{p(\theta_t \mid x_t)\ p(x_t \mid \theta_{1:t-1})}{p(\theta_t \mid \theta_{1:t-1})} \tag{4}$$

Direct calculation of the above two probabilities is practically not possible. Therefore, the posterior probability $p(x_t \mid \theta_{1:t})$ is approximated from a set of *N* particles (i.e. *N* frames), $\Phi_t = \{x_t^i\}_{i=1}^N$ with importance weights given as $W = \{w_t^1, w_t^2, ..., w_t^N\}$. The weights of the samples are updated as:

$$w_t^i = w_{t-1}^i \frac{p(\theta_t \mid x_t^i)\ p(x_t^i \mid x_{t-1}^i)}{\Omega(x_t \mid \theta_{1:t}, x_{1:t-1})} \tag{5}$$

here, $\Omega(x_t \mid \theta_{1:t}, x_{1:t-1})$ is the sequential importance distribution. If in a particular case $\Omega(x_t \mid \theta_{1:t}, x_{1:t-1}) = p(x_t \mid x_{t-1})$ then equation (4) will be simply $w_t^i = w_{t-1}^i\, p(\theta_t \mid x_t)$. This implies that weight of some particles may keep increasing. This will lead to degenerative condition. To avoid this situation, all samples are re-sampled having same weight as per their weight distribution. For this, transformation samples are drawn from Gaussian distribution.

Defining $\delta = \sqrt{\frac{(\beta_1^2 + \beta_2^2 + \beta_3^2 + \beta_4^2)}{2}}$ the new resampled affine parameters are redefined as $[\beta_1 + \delta.\beta_2 + \delta.\beta_3 + \beta_4 + \delta + \delta]$.

### 3.5 Sparse Representation

The global appearance of candidate target under varying illumination conditions and pose changes will approximately lie in the low dimensional subspace spanned by the previously captured templates. Before video stabilisation and tracking, a candidate target image patch is cut and normalised so that the size of the candidate target matches with previous templates as discussed in Section 3.2. All the templates are stored into a 1D array and named as $T = [t_1, t_2, ..., t_n]$, where *n* is the number of templates. A target coefficient vector is defined as

$\alpha = (\alpha_1, \alpha_2, ..., \alpha_n)$. Now as discussed before, the candidate target is represented as an approximate linear combination of templates:

$$y \triangleq T.\alpha = \alpha_1 t_1 + \alpha_2 t_2 + ... + \alpha_n t_n \quad (6)$$

It is observed that candidate target may be affected by occlusion. These occlusions may be scattered or connected regions of pixels. The second case is taken and solved which is more likely to appear, using a set of unit vectors. These unit vectors are called trivial templates. Here a trivial template is a vector with only one nonzero entry at the $i_{th}$ location. Then these trivial templates are arranged in a fashion so that resultant is an identity matrix. These trivial templates are stored in $I = [i_1, i_2, ..., i_p]$ and trivial coefficients are stored in $c = (c_1, c_2, ..., c_p)^T$ where 'p' is the number of pixels in the template. The target of interest can be well represented using target templates and trivial templates in a linear combination. It is noted that the number of positive trivials will be equal to the multiplication of number of rows and number of columns of template. To counter occlusion ***y*** is redefined as

$$y = \begin{bmatrix} T & I \end{bmatrix} \begin{bmatrix} \alpha \\ c \end{bmatrix} \quad (7)$$

Theoretically, trivial coefficient vector will be any real number, but during tracking of the candidate target, which is approximated by the immediate previous target template, it will correspond to the non-negative trivial coefficient. Since the present target template is updated, the next candidate target will be best represented by using non-negative trivial coefficients. However, in situations like shadow, intensity pattern will be reversed and non-negative constraint may result in tracking drift. To avoid such conditions, negative trivial templates are also considered.

$$y = \begin{bmatrix} T & I & -I \end{bmatrix} \begin{bmatrix} \alpha \\ c^+ \\ c^- \end{bmatrix} \triangleq M\, b \qquad \text{s.t. } b \geq 0 \quad (8)$$

where $\geq$ denotes component-wise inequality**.** Here $c^+$ and $c^-$ denote the positive and negative trivial coefficient vectors. Here $M = \begin{bmatrix} T & I & -I \end{bmatrix}$ and $b = \begin{bmatrix} \alpha & c^+ & c^- \end{bmatrix}^T$ are non-negative coefficient vectors. It is noted that the solution of equation (8) is underdetermined and one cannot get a unique solution for ***b***. If there is some error due to noise and occlusion, then those corrupted pixels will correspond to zero coefficients or it may be said that for a best-matched target candidate, very few coefficients of $c^+$ and $c^-$ will be nonzero. This leads to a sparse solution for equation (8). The L1 regularised least-square solution is used as below[28-30].

$$\min \| M\, b - y \|_2^2 + \lambda \| b \|_1 \quad , c \geq 0 \quad (9)$$

The largest coefficient value will correspond to the best target match for video tracking. It is also emphasised to save the target coordinate for video stabilisation as discussed below.

### 3.6 Digital Video Stabilisation

It is to be noted that in the literature all minimisation problems were solved assuming that there was no unintentional global motion. However, in the presence of unwanted global motion, that is when the frame is destabilised during the movement of vehicle, the classical electronics stabilisation method is not always accurate. Therefore, a second level software video stabilisation is proposed while tracking. The appearance model of templates has been defined already. There are two types of templates defined, viz. target templates and trivial templates. Target templates represent target for visual tracking and are updated dynamically, while trivial templates represent the occlusion, noise and background. There may be many instances when these trivial templates do not fit the target effectively if the frames are destabilised and causes drift. Experiments have shown that this local drift in covering target region can be effectively represented by affine parameters. The affine parameters of target candidate are already matched with previous templates and the coordinate obtained, as in Section 3.5. Since this algorithm corrects only horizontal and vertical global motion, therefore only first four Affine transformation parameters are chosen. Suppose target model has translation affine parameters as $(h^1_{t,x}, h^4_{t,y})$ in frame *t* and $(h^1_{t+1,x}, h^4_{t+1,y})$ in the next consecutive frame. Here superscripts '1' and '4' correspond to the array positions in affine parameter model. Then before invoking tracking function, the frame can be stabilised in opposite directions by calculating the shift in horizontal (*dx*) and vertical (*dy*) directions as

$$dx = |(h^1_{t+1,x} - h^1_{t,x})| \quad and \quad dy = |(h^4_{t+1,y} - h^4_{t,y})| \quad (10)$$

Before taking the mod, sign of *dx* or *dy* is preserved. If it is positive, then frame is shifted left(down) and if it is negative, then frame is shifted right(up). A modified version of Eqn. (9) is presented by analysing that when there is no destabilisation phenomenon, the energy of the coefficients in **b** associated with target template will be small. If there is destabilisation, then **b** will have some more energy due to the displacement vector. This destabilised factor is associated in Eqn. (9). Secondly, when there is no occlusion, then the target in consecutive frames may be well approximated by a linear combination of sparse representation and hence, the energy of trivial coefficients associated with **b** will be minimum. Thirdly, if there exists occlusion, then the trivial coefficients will have more energy. The minimisation of Eqn. (9) is not appropriate as it does not consider these three cases under destabilisation case.

Motivated by the above discussion, an adaptive control on the energy of target template as well as on trivial coefficients is presented. When the destabilisation and occlusion are negligible, the energy associated with target template and trivial template should be small, while if there is sufficient amount of destabilisation and occlusion, the energy is allowed to be large. Now, a robust minimisation model for L1 tracker is defined as follows:

$$\min \frac{1}{2} \| M\, b - y \|_2^2 + \lambda \| b \|_1 + \frac{\pi_t}{2} \| b_I \|_2^2 + \frac{\rho_t}{2} \| b_T \|_2^2$$
$$s.t. \;\; b_T \geq 0 \quad (11)$$

where, $b = \begin{bmatrix} b_T & b_I \end{bmatrix}$ are coefficients associated with target templates and trivial templates respectively. $\pi_t$ and $\rho_t$ control the energy of coefficients in trivial and target templates. While implementing the above, care is taken that if there is no

occlusion, then $\pi_t$ =0; if there is no destabilisation detected, $\rho_t$ =0; otherwise $\pi_t$ and $\rho_t$ take some predefined constants. It is noted that for a stabilised video input, the minimisation Eqn. (9) is a special case of Eqn. (11). The last two terms in Eqn. (11) is the average of consecutive displacements with occlusion.

### 3.7 Updation of Templates

The templates have to be updated to cope up with recent changes in video like pose or illumination level. If template is updated too frequently, then each time there will also present some small error which will cause tracking and stabilisation drift due to accumulation of errors. If template is updated less often, then the target may be lost after some time due to appearance variation. Therefore a midway level is taken and dynamic template updation scheme is applied. An importance weight $w_i$ is assigned with the template $t_i$ for stability. Larger weight will correspond to more important template, while smaller weight will correspond to rarely used template. Therefore these rarely used templates can be discarded and new template is taken and template vector is updated dynamically. While solving Eqn. (9), it is observed that larger the norm $\| b \|_1$, smaller the coefficient for approximation, favoring the template. Therefore this feature may be used for assigning the weight. For this, the second norm of template i.e. $\| t_i \|_2$ is taken and assigned to $w_i$ of template $t_i$. If candidate target does not match the target template, the weight of that respective template can be re-assigned as the mean of all the weights. This will prevent the newly added templates to dominate old templates. In this way the weights of templates resembling the target closely will increase and decrease otherwise.

### 3.8 Computational Complexity

When original L1 tracker is solved using preconditioned conjugate gradients (PCG) method to solve L1 minimisation function, a matrix vector product is solved which is the most costly step having computational complexity of $O(d^2 + dn)$, where hundreds of such steps are used. In the present case to solve the Eqn. (11) using Cholesky factorisation method, computational complexity is given by $dn^2 + (1/3)n^3$. For example in the car movie, template size d=15x12=180, number of templates, n=10. The cost of Cholesky function is $180 \times 100 + (1/3) \times 1000 \approx 18000$. Similarly the cost for PCG step is $180 \times 180 + 180 \times 10 = O(32400)$, and there will be few hundred such steps. With this it is concluded that the proposed method is at least two order faster than the original one. It will provide 15 to 16 frames per second rate.

## 4. EXPERIMENTS AND RESULTS

HP Z840 workstation machine having 64 bit operating system and 64 GB RAM installed with MATLAB R2014a version was used to implement and performance analysis was carried out in the proposed algorithm. A number of destabilised visual tracking videos were recorded and used as input in proposed algorithm. Fast and robust stabilised visual tracking result could be obtained with promising results.

### 4.1 Ground Truth Generation

To quantify the amount of destabilisation and to measure the tracking accuracy, a sample input video was taken. This video was destabilised with known parameters using two random number generators (h,v). Each frame is shifted by this amount in horizontal (h pixels) direction and in vertical (v pixels) direction. To give a smooth motion, this randomly destabilised video needs to pass in a low pass filter. A three tap moving average filter was effectively used to smooth the video. More number of taps in low pass filter will increase computational cost. This input video with known shifts in both directions is the ground truth for comparison with different visual trackers. To retain the same environment, a sample 'car. avi' destabilised video was taken. All video frames are in gray scale.

### 4.2 Performance Analysis

A novel approach has been presented here combining digital video stabilisation followed by visual tracking. It is brought out that such type of algorithm could not be obtained to compare with the proposed one. Therefore the performance analysis of the proposed algorithm is carried out in two parts.

The new algorithm is compared with four benchmark tracking algorithms; these are online multiple instance learning[18], weighted multiple instance learning[20], online discriminative feature selection[21] and compressive tracking[22].

#### *4.2.1 Qualitative Performance*

A comparison of the proposed RSVT algorithm is as shown in Fig. 5. It is seen that for destabilised videos, RSV is much more robust than other four algorithms. Tracking drift can be seen in all other algorithms. WMIL performs worst when compared with RSVT.

#### *4.2.2 Horizontal and Vertical Error Measurement*

Horizontal and vertical error measurement of tracking coordinates for first 50 frames were taken using CT, MIL, ODFS and WMIL w.r.t. RSVT is as shown in Fig. 6. where x-axis is number of frames and y-axis is the magnitude in pixel.

#### *4.2.3 Statistical Error Measurement*

A statistical measurement of standard deviation for x and y errors, comparative to RSVT algorithm is placed in Table 1.

**Table 1. Standard deviation error**

| Standard deviation error value | 'x' error | 'y' error |
|---|---|---|
| MIL algorithm | 26 | 9 |
| WMIL algorithm | 53 | 17 |
| ODFS algorithm | 21 | 12 |
| CT algorithm | 18 | 15 |

#### *4.2.4 Root Mean Square Error Measurement*

Root Mean Square Error (RMSE) is given by the formula as follows:

$$RMSE = \sqrt{\frac{1}{N}\sum_{n=1}^{N}[(x_n - x_d)^2 + (y_n - y_d)^2]} \qquad (12)$$

where $N$ is the number of frames, $(x_n, y_n)$ is the tracking coordinate of MIL, WMIL, ODFS and CT algorithms,

**Figure 5. Comparison with bench mark visual trackers : (a) RSVT, (b) MIL, (c) WMIL, (d) ODFS, (d) CT.**

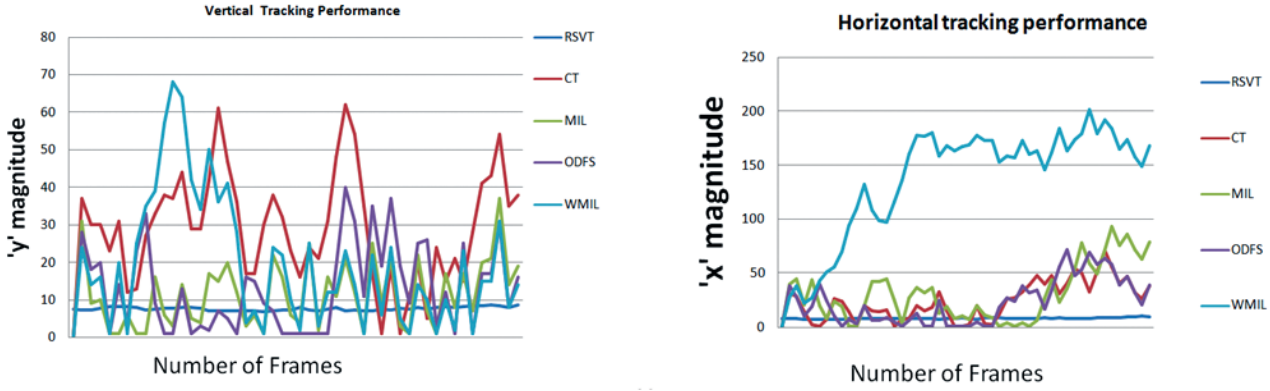


**Figure 6. Horizontal and vertical error graph.**

respectively, and $(x_d, y_d)$ is the tracking coordinate of RSVT algorithm. A comparison for first fifty frames is as given in Table 2.

**Table 2. RMSE comparison**

| Algorithm | RMSE |
|---|---|
| RSVT | 11 |
| MIL | 43 |
| WMIL | 148 |
| ODFS | 36 |
| CT | 44 |

It is noted that the proposed RSVT algorithm is more robust than other bench mark algorithms for a given destabilised visual tracking problem.

### *4.2.5 Video Stabilisation PSNR Performance Analysis*

A destabilise video will have poor quality. The quality of stabilised video is measured in terms of Peak Signal to Noise Ratio (PSNR) between two successive frames. PSNR value will be always higher for stabilised frames than destabilised frames. PSNR is expressed in terms of power. Higher the PSNR value, higher will be the fidelity of video. PSNR between two consecutive frames is expressed as follows:

$$PSNR = 10 \log_{10} \frac{I_{\max}^2}{MSE(k)} \tag{13}$$

where $I_{\max}$ is the value of maximum intensity level, which is 255 for gray level case. MSE is the Mean Square Error between two successive frames which tells the average departure per pixel from the desired stabilised result and $k = 1 \ldots N_{frames}$. A comparative graph between RSVT and other four bench mark algorithm as discussed in section 4.2 is as shown in Fig. 7 which clearly shows the superiority of RSVT algorithm among other bench mark algorithms due to stabilisation.

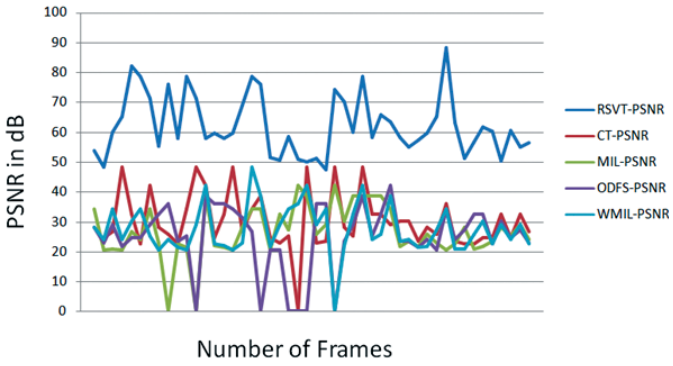


**Figure 7. PSNR graph.**

### *4.2.6 Video Stabilisation ITF Performance Analysis*

Inter Transform Fidelity (ITF) is another criteria to quantify the stabilisation of a video. Higher the ITF value, higher is the quality of stabilisation. ITF is give as

$$ITF = \frac{1}{(N_{frames} - 1)} \sum_{k=1}^{(N_{frames}-1)} PSNR(k) \tag{14}$$

ITF is calculated for the videos and compared with other four bench mark algorithms. As expected RSVT has much higher value of ITF due to stabilisation and tracking features.

**Table 3. ITF comparison**

| Algorithm | ITF (dB) |
|---|---|
| RSVT | 60.90587 |
| CT | 29.00457 |
| MIL | 26.18068 |
| ODFS | 24.40022 |
| WMIL | 26.95513 |

## 5. CONCLUSIONS

In this paper, a novel approach is presented to solve the two problems of digital video stabilisation and visual tracking simultaneously. A robust algorithm RSVT was developed with new L1 minimisation model which is also capable of taking destabilised video and perform visual tracking after stabilising it using affine transformation parameters. This technique is very useful in defence applications for electro-optical tracking equipment used for military surveillance. A second level finer video stabilisation is achieved followed by robust visual tracking. A performance analysis of the proposed RSVT algorithm was carried out with well known bench marked visual trackers and it was found that in the presence of destabilised video input, proposed RSVT outperforms and is more robust. The global as well as local motion could be retained while simultaneously stabilising and tracking the video. It is observed that while stabilising the video, some information is lost and this resulted in varying width of black strips in horizontal as well as in vertical directions depending upon the destabilisation magnitude. The lost information can be recalled from previous frame and new frame can be displayed by mosaicing the missing information. Image stitching is proposed to retain lost information in the final stabilised video as a future work.

## REFERENCES


1. Verma, K.; Kumar, A.; Gupta, S. & Venkatesh, K.S. Fast and robust real time digital video stabilisation with smear removal using integral projection curve warping technique. *In* IEEE Twentieth National Conference on Communications, 2014. doi: 10.1109/NCC.2014.6811276
2. Tsai, T.; Lun, C. & Chuang, F.H. Design and implementation of efficient video stabilisation engine using maximum a posteriori estimation and motion energy smoothing approach. *IEEE Trans. Circuits Syst. Video Technol.*, 2012, **22**(6), 817-830. doi: 10.1109/TCSVT.2011.2177179
3. Jia, C. & Evans, B.L. Constrained 3D rotation smoothing via global manifold regression for video stabilisation. *IEEE Trans. Signal Process.*, 2014, **62**(13), 3293-3304. doi: 10.1109/TSP.2014.2325795
4. Battiato, S.; Bruna, A. & Puglisi, G. A robust block-based image/video registration approach for mobile imaging devices. *IEEE Trans. Multimedia*, 2010, **12**(7), 622-635. doi: 10.1109/TMM.2010.2060474
5. Puglisi, G. & Battiato, S. A robust image alignment algorithm for video stabilisation purposes. *IEEE Trans. Circuits Syst. Video Technol.*, 2011, **21**(10), 1390-1400. doi: 10.1109/TCSVT.2011.2162689
6. Xu, J.; Chang, H.; Yang, S. & Wang, M. Fast feature-based video stabilisation without accumulative global motion estimation. *IEEE Trans. Consum. Electron.*, 2012, **58**(3), 993-999. doi: 10.1109/TCE.2012.6311347
7. Kim, S.; Kang, S.; Wang, T. & JeaKo, S. Feature point classification based global motion estimation for video stabilisation. *IEEE Trans. on Consum. Electron.*, 2013, **59**(1), 267-272. doi: 10.1109/TCE.2013.6490269
8. Yilmaz, A.; Javed, O. & Shah, M. Object tracking: A survey. *ACM Comput. Surveys*, 2006, **38**(4), Article 13.
9. Jalal, A.S. & Singh, V. The state-of-the-art in visual object tracking. *Informatica (Slovenia)*, 2012, **36**(3), 227-248.
10. Hu, W.; Tan, T.; Wang, L. & Maybank, S. A survey on visual surveillance of object motion and behaviors. *IEEE Trans. Syst., Man, Cybern.- Pt C: Appl. and Rev*., 2004, **34**(3), 334-352. doi: 10.1109/TSMCC.2004.829274
11. Zoidi, O.; Tefas, A. & Pitas, I. Visual object tracking based on local steering kernels and color histograms. *IEEE Trans. Circuits Syst. Video Technol.*, 2013, **23**(5). doi: 10.1109/TCSVT.2012.2226527
12. Zhang, S.; Yu, X.; Sui, Y.; Zhao, S. & Zhang, Li. Object tracking with multi-view support vector machines. *IEEE Trans. Multimedia*, 2015, **17**(3), 265-278. doi: 10.1109/TMM.2015.2390044
13. Sui, Y.; Zhang, S. & Zhang, Li. Robust visual tracking via sparsity-induced subspace learning. *IEEE Trans. Image Process.*, 2015, **24**(12), 4686-4700. doi: 10.1109/TIP.2015.2462076
14. Biresaw, T.A.; Cavallaro, A. & Regazzoni, C.S. Tracker-level fusion for robust bayesian visual tracking. *IEEE Trans. Circuits Syst. Video Technol.*, 2015, **25**(5), 776-789. doi: 10.1109/TCSVT.2014.2360027
15. Yuan, Y.; Yang, H.; Fang, Y. & Lin, W. Visual object tracking by structure complexity coefficients. *IEEE Trans. Multimedia*, 2015, **17**(8), 1125-1136. doi: 10.1109/TMM.2015.2440996
16. Sui, Y. & Zhang, Li. Visual tracking via locally structured gaussian process regression. *IEEE Signal Process. Lett.*, 2015, **22**(9), 1331-1335. doi: 10.1109/LSP.2015.2402313
17. Wu, B.F.; Kao, C.; Jen, C.L.; Li, Y.F.; Chen, Y.H. & Juang, J.H. A Relative-discriminative-histogram-of-oriented-gradients-based particle filter approach to vehicle occlusion handling and tracking. *IEEE Trans. Ind. Electron.*, 2014, **61**(8), 4228 – 4237. doi: 10.1109/TIE.2013.2284131
18. Babenko, B.; Yang, M.H. & Belongie, S. Robust object tracking with online multiple instance learning. *IEEE Trans. Pattern Anal. Mach., Intell.*, 2011, **33**(8), 1619-1632. doi: 10.1109/TPAMI.2010.226
19. Zhang, K.; Zhang, L.; Yang, M.H. & Hu, Q. Robust object tracking via active feature selection. *IEEE Trans. Circuits Syst. Video Technol.,* 2013, **23**(11), 1957-1967. doi: 10.1109/TCSVT.2013.2269772
20. Zhang, K. & Song, H. Real-Time Visual tracking via online weighted multiple instance learning. *Pattern Recog.,* 2013, **46**(1), 397–411.
21. K. Zhang, L. Zhang, & M. H. Yang. Real-time object tracking via online discriminative feature selection. *IEEE Trans. Image Process.,* 2013, **22**(12), 4664-4677. doi: 10.1109/TIP.2013.2277800

22. Zhang, K.; Zhang, L. & Yang, M.H. Real-time compressive tracking. *In* ECCV 2012, Proceeding of the 12th European Conference on Computer Vision, Part III, 2012, LNCS 7574, 866–879. doi: 10.1007/978-3-642-33712-3_62
23. Jiang, H.; F.S. & Little, J.J. Optimizing multiple object tracking and best view video synthesis. *IEEE Trans. Multimedia*, 2008, **10**(6), 997-1012. doi: 10.1109/TMM.2008.2001379
24. Zoidi, O. & Pitas, A.I. Visual object tracking based on local steering kernels and color histograms. *IEEE Trans. Circuits Syst. Video Technol.*, 2013, **23**(5), 870-882. doi: 10.1109/TCSVT.2012.2226527
25. Cavallaro, A.; Steiger, O. & Ebrahimi, T. Tracking video objects in cluttered background. *IEEE Trans. Circuits Syst. Video Technol.,* 2005, **15**(4), 575-584. doi: 10.1109/TCSVT.2005.844447
26. Shen, H.; Li, S.; Zhang, J. & Chang, H. Tracking-based moving object detection. *In* 20th IEEE International Conference on Image Processing*,* 2013, pp. 15-18. doi: 10.1109/ICIP.2013.6738637.
27. Jazayeri, A.; Cai, H.; Yu Zheng, J. & Tuceryan, M. Vehicle detection and tracking in car video based on motion model. *IEEE Trans. Intell. Transp. Syst.*, 2011, **12**(2), 583-595. doi: 10.1109/TITS.2011.2113340.
28. Mei, X. & Ling, H. Robust visual tracking and vehicle classification via sparse representation. *IEEE Trans. Pattern Anal. Mach. Intell.,* 2011, **33**(11), 2259-2272. doi: 10.1109/TPAMI.2011.66
29. Mei, X. & Ling, H. Robust visual tracking using L1 minimization. 1436-1443, ICCV, 2009.
30. Mei, X.; Ling, H.; Wu, Y.; Blasch, E. & Bai, L. Minimum error bounded efficient L1Tracker with occlusion detection. *In* IEEE Conference on Computer Vision and Pattern Recognition, 2011*,* doi: 10.1109/CVPR.2011.5995421
31. Zeng, X.; Xu, L.; Ma, L.; Zhao, R. & Cen, Y. Visual tracking using global sparse coding and local convolutional features. *J. Dig. Sig. Process.*, 2017, **72**, 115-125. doi: 10.1016/j.dsp.2017.10.007

## ACKNOWLEDGEMENTS

First author extends his deep sense of gratitude to Mr Rajeev Marathe, Scientist 'G', for his support and encouragement for this work. Authors would like to sincerely thank Mr Benjamin Lionel, Director IRDE for encouragement and permission to publish the work.

## CONTRIBUTORS

**Mr Kamlesh Verma** received his MTech from IIT Kanpur, in 2010 and currently pursuing his PhD from IIT Roorkee. He is serving as a senior Scientist in Instruments Research and Development Establishment, DRDO, Dehradun. His research area of interest is design and development of embedded hardware and software for line-of-sight stabilisation technology, computer vision, image and video processing, target detection and tracking.
Contribution in the current work: He has carried out exhaustive literature survey on video stabilisation and target tracking and developed the RSVT algorithm. He has prepared the manuscript of this paper.

**Mr Avnish Kumar** received his BTech and MTech (Electrical engineering) from IIT Roorkee. Currently, he is working at IRDE, Dehradun, and is guiding a team of scientists for the design and development of LOS stabilisation, electro-optical surveillance, and fire-control systems for various platforms.
Contribution in the current, he has given his valuable inputs for the development of algorithm based on his rich practical experience for visual tracking and stabilisation.

**Dr Debashis Ghosh** received his MS and PhD in Communication Engg. from IISc Bangalore, in 1996 and 2000, respectively. He is currently Professor & Head in the Dept. of Electronics & Communication Engg., IIT Roorkee. His teaching and research interests include communication systems, signal processing, image/video processing, computer vision, and pattern recognition.
Contribution in the current study, he has guided and supervised this work.