Weighted Feature Points Extraction based Video Stabilization

K. Madhavi, B. Sreekanth Reddy, Ch.Ganapathy Reddy

Abstract--- ***Camera global motion estimation is critical to the success of video stabilization. An extension of Video stabilization using principal component analysis (PCA) and scale invariant feature transform (SIFT) in particle filter framework is proposed. In the proposed method the feature points are collected from based on Speeded Up Robust Features (SURF). Random Samples Consensus (RANSAC) is used to remove local motion vectors and incorrect correspondences. A particle filter is used to estimate the weight of feature points, solving the issue of Different Depth of Field (DDOF) for feature points weighted least square (WLS) algorithm is applied in the global motion estimation. Finally, a Kalman filter estimates the intentional motion, and the unintentional motion is compensated to obtain stable video sequences. The algorithm has the characteristics of high precision and good robustness.***

Keywords--- particle filter, principal component analysis (PCA), scale invariant feature transform (SIFT), speeded up robust features (SURF).

I. INTRODUCTION

With handheld cameras, camera motion and platform vibrations are difficult to be avoided, which generates unstable video sequences. Therefore, video stabilization becomes an indispensable technique in advanced digital cameras and camcorders. The main contribution of this paper is to develop a novel motion estimation approach based on particle filtering for video stabilization. The key insight of this approach is that, feature points should have different contributions to the estimation results, and good estimation should depend on feature points with similar DOF. In the proposed approach, feature points are weighted, and a WLS algorithm is used to obtain estimation results.

II. BACKGROUND

Video stabilization techniques have been studied for a long time and attracted even great interests in recent years. An automatic image-stabilizing system for camcorders, utilizing only digital signal processing was developed [1]. Then a compact electronic image stabilizer on the basis of scanning area selection of the imager and motion vector detection was realized [2].

This system is suitable for compact video cameras. However, the stabilization rate becomes poor at high frequency. A DIS for video cameras is proposed [3]. It is composed by an edge detection unit, a motion detection unit and a digital zooming unit. The proposed DIS system is designed mainly for hardware minimization in a video camera system. Then extraction and tracking of corner features in order to estimate global motion is done [4]. However, the features are not robust with respect to some image transformations, such as scaling and rotation. To overcome the inefficiency in scaling and rotation, SIFT features [5] and PCA-SIFT [6] are being widely used for global motion estimation. Recently a DIS algorithm based on feature point tracking is presented [7]. The Kanade-Lucas-Tomasi (KLT) tracker to estimate the global motion between two consecutive image frames is used. The motion prediction by the Kalman filter (KF) is incorporated into the KLT tracker to further speed up the tracking process. A novel digital-image stabilization scheme based on independent component analysis (ICA) is proposed [8]. The method utilizes ICA and information obtained from the image sequence to deconvolve the egomotion from the unwanted motion of the sequence. A method to stabilize video for vehicular applications based on feature analysis is proposed [9]. SURF [10] features descriptor is used. For feature matching, KD tree with best-bin-first search significantly reduces the matching time. A damping filer is utilized, predicting the unwanted oscillation.

III. PREVIOUS WORK

Yao Shen and Parthasarathy Guturu had observed the dimensionality of the feature space is first reduced by the principal component analysis (PCA) method using the features obtained from a scale invariant feature transform (SIFT), and hence the resultant features may be termed as the PCA-SIFT features. The trajectory of these features extracted from video frames is used to estimate undesirable motion between frames. A new cost function called SIFT-BMSE (SIFT Block Mean Square Error) is proposed in adaptive particle filter framework to disregard the foreground object pixels and reduce the computational cost. Frame compensation based on these estimates yields stabilized full-frame video sequences. Experimental results show that the algorithm is both accurate and efficient.

The stabilization approaches available in the technique may be classified mainly into two categories: (1) Hardware based methods and (2) Image processing software based methods. Hardware based methods physically avoid camera jerks by adjusting camera motion sensors once unwanted motion is detected. Though this approach performs well in real life applications, the video systems adopting this approach turn out to be very expensive because of the need for sophisticated sensors that measure camera jerks accurately. On the other hand, image processing software based methods, usually known as digital image stabilization methods, are much less expensive. The proposed method uses particle filter (PF) approach for an accurate estimation of undesired motion of the camera.

Manuscript received September, 2013.

K.Madhavi, M.Tech Student, ECE Department, GNITS, Hyderabad, India.

S.Satheesh, Asst. Professor, ECE Department, GNITS, Hyderabad, India.

Ch.Ganapathy Reddy, Professor & HOD, ECE Department, GNITS, Hyderabad, India.

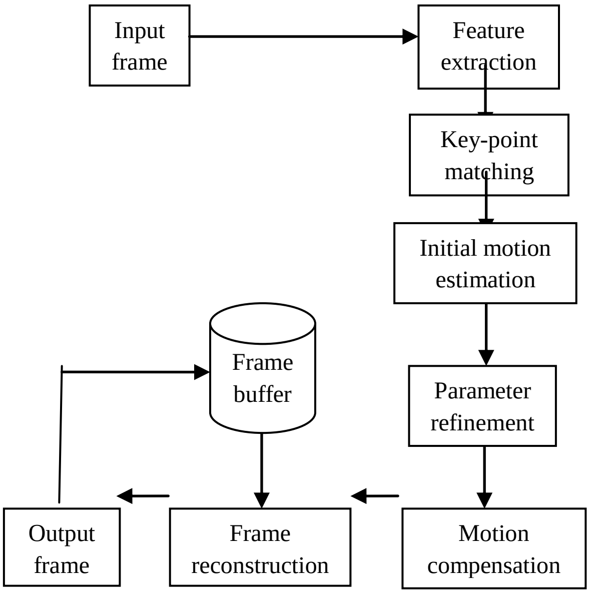


Figure 1: Processing steps of existing technique

Figure1 shows the steps involved in the technique are feature extraction and matching, initial motion estimation using RANSAC, theoretical foundations of particle filter, construction of motion model, SIFT-BMSE cost function, adaptive model noise, particle number and frame reconstruction.

Traditionally, the geometric transformation between two images can be described by a homograph which is a 3D model with eight unknown parameters. However, due to the complexity of homograph, a 2D affine transformation having only four unknown parameters is adopted. Suppose $P_1 = (x, y, 1)^T$ and $P_2 = (x', y', 1)^T$ to be the pixel location of corresponding points in consecutive video frames, the relationship between these two locations can be expressed by following transform.

$$\begin{bmatrix} x' \\ y' \\ 1 \end{bmatrix} = \begin{bmatrix} C_o \cos(\theta_o) & -C_o \sin(\theta_o) & Tx_o \\ C_o \sin(\theta_o) & C_o \cos(\theta_o) & Ty_o \\ 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} x \\ y \\ 1 \end{bmatrix}$$

Here, Tx_o and Ty_o denote translation along x and y axis respectively. C_o and θ_o are scaling and rotation parameters in the image plane.

Although the number of mismatches can be reduced by using PCA-SIFT compared with SIFT, small amount of mismatches still occur, and this may lead to unreliable prediction of motion estimation. Thus, a further check of matching errors is a significant part of the algorithm. RANSAC algorithm is used for both elimination of the outliers in the previous matching and estimation of the four unknown parameters of the above affine transform model. In each iteration of the RANSAC algorithm, minimal sample sets (MSSs) are randomly selected from the input dataset and the affine model parameters are computed using only the elements of this MSS, then RANSAC checks the entire dataset and decides the inliers or outliers depending upon whether an element of input data set fits the model or not. After K iterations, the result that has minimal outliers is used as the initial value of the parameters of affine transform model.

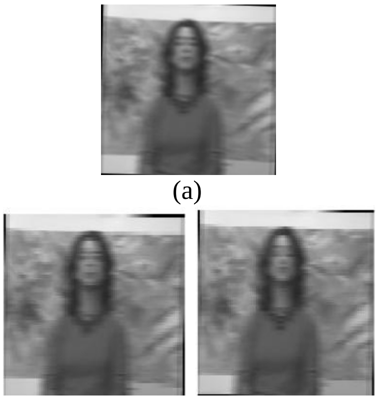
The state vector of S_t can be represented as $S_t = \left[T_x, T_y, \theta, C\right]^T$ where these four elements represent translation along x, y axis, rotation and scaling parameters in affine transform model. The global motion which can be considered as a cumulative motion of previous frame neighbors must be estimated. The state transition equation in our PF model is given by:

$$S_{t+1} = AS_t + U_t \equiv \begin{bmatrix} T_x \\ T_y \\ \theta \\ C \end{bmatrix}_{t+1}$$

$$= \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 \end{bmatrix} \begin{bmatrix} T_x \\ T_y \\ \theta \\ C \end{bmatrix}_t + \begin{bmatrix} N(T_{x_o}, \sigma_x) \\ N(T_{y_o}, \sigma_y) \\ N(\theta_o, \sigma_\theta) \\ N(C_o, \sigma_c) \end{bmatrix} \qquad (1)$$

Here A is the transition matrix and U_t is the process noise following the Gaussian distribution with the variances $\left[\sigma_x, \sigma_y, \sigma_\theta, \sigma_C\right]$ and the means $\left[T_{x_o}, T_{y_o}, \theta_o, C_o\right]$. The components of the mean vector are determined using the RANSAC procedure described. Unlike other parameters, the cumulative result of scaling parameter is a product rather than a summation of the previous values. Hence, in the above transformation, the scaling parameter based on the initial value C_o is computed in the above transformation as computed by RANSAC method without considering the prior state. Another reason for this kind of computation model is that scale changes tend to be small, thus, direct estimation without sacrificing accuracy is possible. The RANSAC approach provides rough estimates of the motion parameters to the PF algorithm, and thereby saves the unnecessary computational cost involved in the generation of useless particle samples.

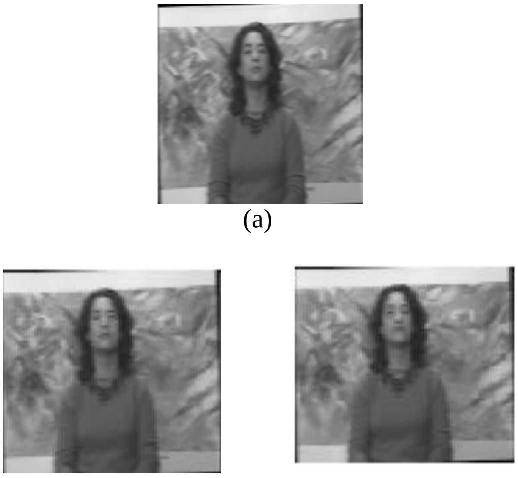
After each frame is compensated by using the above discussed motion estimates, the pixels near the frame boundary may be undefined, and this leads to unacceptable visual effects. Traditionally, many researchers either trim the undefined region or fill in the region with a constant value. This results in information loss and quality degradation especially in the case of frames that undergo large scale translations and rotations because of a jerky camera. To avoid this problem after compensation, information of neighbouring frames can be borrowed to fill the undefined regions though sometimes undefined pixels may still exist. Intuitively, if we increase the number of neighbouring frames, the number of pixels in undefined region will be reduced, but it is computationally intensive. In this technique mosaic method to reconstruct undefined regions using previous stabilized full-frames is applied, since the undefined region of these frames is already determined. However, this may occasionally result in a carry-forward information (information from previous frames) across a considerable number of frames.


(a)

(b) (c)

Figure 2: Input frames extracted (a) Original sample, (b) Frame #1 (c) Frame #2

Figure 2 shows the frames extracted from the input video sequence. Here in this example, 2 frames are extracted from the unstabilized video sequence.



(a)

(b) (c)

Figure 3: Output compensated frames (a) Original sample, (b) Frame #1 (c) Frame #2

Figure 3 shows the frames of output video sequence. As in figure 1, 2 compensated frames are shown.

IV. PROPOSED METHOD

The stabilization method consists of following steps: feature point extraction, RANSAC (Random Samples Consensus) based local motion estimation, particle filtering based global motion estimation, Kalman filtering based intentional motion estimation and image compensation.

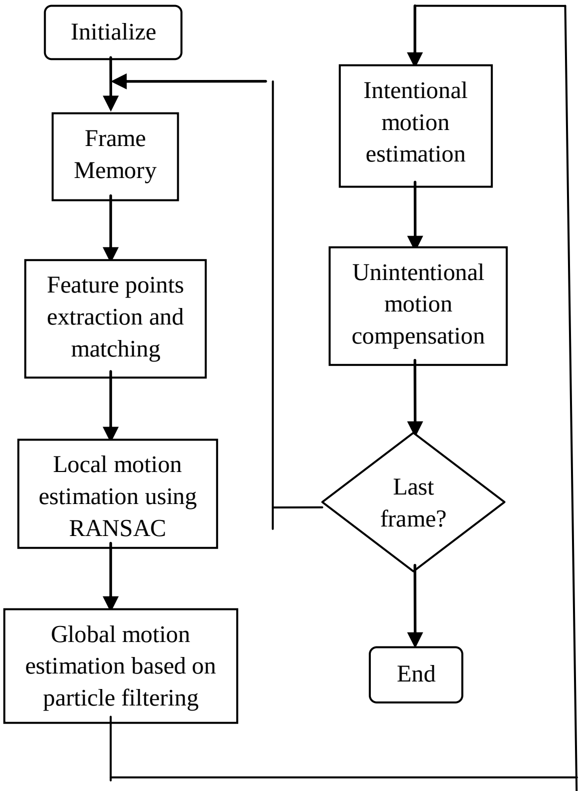

Initialize
Frame Memory
Feature points extraction and matching
Local motion estimation using RANSAC
Global motion estimation based on particle filtering
Intentional motion estimation
Unintentional motion compensation
Last frame?
End

Figure 4: Flow steps of proposed algorithm

The figure 4 shows the steps of the proposed technique the selection of features for motion estimation is very important, since unstable features may produce unreliable estimations with variations in rotation, scaling or illumination. SURF (Speeded Up Robust Features) is a robust image interest point detector SURF descriptor is similar to the gradient information extracted by SIFT (Scale Invariant Feature Transform) and its variants, describing the distribution of the intensity content within the interest point neighbourhood. SURF has similar performance to SIFT, however, faster. The increase in speed is due to the use of integral images, which drastically reduces the number of operations for simple box convolutions that is independent of the chosen scale.

The RANSAC (Random Samples Consensus) algorithm is used to eliminate the outlier feather points. In each iteration of the RANSAC algorithm, minimal sample sets are randomly selected from feature points. Then, RANSAC decides the inliers or outliers depending upon whether an element of input data set fits the model or not. After K iterations, the result that has minimal outliers is used as the initial value of the parameters in the affine transform model.

The global motion contains the intentional and unintentional motion. As we only want to compensate the transformation caused by unintentional camera movements, transformations due to intentional motion should be identified. The KF (Kalman Filter) is applied to estimate the intentional motion of the camera. KF uses measurements that observed over time, containing noise (random variations) and other inaccuracies, and produces values that tend to be closer to the true values of the measurements and their associated calculated values.

Generally the geometric transformation between two images can be described by a 2D or 3D homograph model. Due to similarities in 2D and 3D models, in this paper we adopted a rigid 2D model for convenience. It was expressed as:

$$\begin{bmatrix} x_t \\ y_t \\ 1 \end{bmatrix} = S_t \begin{bmatrix} \cos\theta_t & -sin\theta_t & T_t^x \\ \sin\theta_t & \cos\theta_t & T_t^y \\ 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} x_{t-1} \\ y_{t-1} \\ 1 \end{bmatrix} \qquad (2)$$

or in the form of Y=AX, where (x_t, y_t) is the coordinate at time t, θ_t is the rotation, T_t^x and T_t^y are the translations in the horizontal and vertical direction respectively and S_t is the scaling factor. Nevertheless, more matches can be added under least-square criteria to ensure results more robust:

$$A = [X^T \quad X]^{-1} X^T Y \qquad (3)$$

The framework of particle filtering can be expressed as follows. Supposing a nonlinear discrete dynamic system:

$$S_t = f_t(S_{t-1}) + Q_t \qquad (4)$$

$$Y_t = h_t(S_t) + R_t \qquad (5)$$

Where Y_t, S_t, $Q_n^{'}$ and R_t are the observation, the system state, the process noise, and the measurement noise, respectively f_t and h_t are the system state transition function and the observation function, and t is the index of time step.

Assuming the first order Markov model $p(x_k \mid x_{0:k-1}) = p(x_k \mid x_{k-1})$ for state transition and conditional dependence of Y_t exclusively on S_t, it can be considered as a recursive Bayesian estimation problem to obtain the posterior probability $p(S_t \mid Y_{1:t-1})$ by:

$$p(S_t \mid Y_{1:t}) = \frac{p(Y_t \mid S_t) p(S_t \mid Y_{1:t-1})}{p(Y_t \mid Y_{1:t-1})} \qquad (6)$$

Where $p(Y_t \mid S_t)$ is the likelihood, $p(S_t \mid Y_{1:t-1})$ is the system prior, and $p(Y_t \mid Y_{1:t-1})$ is the evidence. Particle filter estimates $p(S_t \mid Y_{1:t-1})$ by using a set of $p_t = \{S_t^i; w_t^i\}_{i=1.....N}$ weighted particles S_t^i drawn from the proposal distribution $\pi(S_t)$, which is an approximation of $p(S_t \mid Y_{1:t-1})$.

In this proposed method, a particle is composed by a small group of feature points selected randomly. As we use the particle to determine a warped frame using (2), the number of feature points should be no less than 3. Note that a feature point may belong to more than one particle. We generate N particles and the weight of a particle is determined by the point-wise MSE criterion:

$$w_t^i = \frac{\exp\left(-\frac{\left(I_{t-1}-I_t^i\right)^2}{2\sigma_t}\right)}{\sum_{i=1}^{N} exp\left(-\frac{\left(I_{t-1}-I_t^i\right)^2}{2\sigma_t}\right)} \qquad (7)$$

Where w_t^i is the weight of the i^{th} particle at time t, I_{t-1} is the image at time t-1 and σ_t is the variance. I_t^i is the i^{th} warped image of I_t computed based on the i^{th} particle. As warped images have undefined regions, computed only in the center region of the reference frame with size $\frac{w}{2} * \frac{h}{2}$ (w and h are the width and height of the reference frame).

The weight of a feature point is determined by the weight of particles which it belongs to:

$$w_t^j = \frac{1}{N_t^j} \sum_{k=1}^{N_t^j} w_t^k \qquad (8)$$

Where w_t^j is the weight of the j-th feature point at time t, N_t^j is the number of particles the j^{th} feature point belongs to, and w_t^k is the corresponding weight of the particle.

After obtaining the weight of feature points, a WLS algorithm can be used to estimate the global motion of the current frame.

$$A = \lfloor X^T \quad AX \rfloor^{-1} X^T AY \qquad (9)$$

Where, A = *diag* ($W_t^1, W_t^2, ... W_t^N$), N is the total number of feature points in a frame. Increasing the number of feature points in a particle will make the algorithm more robust. However, if particles are composed by the whole dataset of feature points, the weight of feature points would be identical and would degenerate. From experiments in it reveals that the accuracy of the motion estimation will firstly increase and then decrease with the increase of the number of feature points in a particle.

The global motion contains the intentional and unintentional motion. As we only want to compensate the transformation caused by unintentional camera movements, transformations due to intentional motion should be identified. We applied the Kalman filter to estimate the intentional motion of the camera. Kalman filter uses measurements that observed over time, containing noise (random variations) and other inaccuracies, and produces values that tend to be closer to the true values of the measurements and their associated calculated values. Assuming that translation, rotation and scale are independent, and then the four parameters can be modeled separately, which leads to simple state transition and observation models. The state space model is

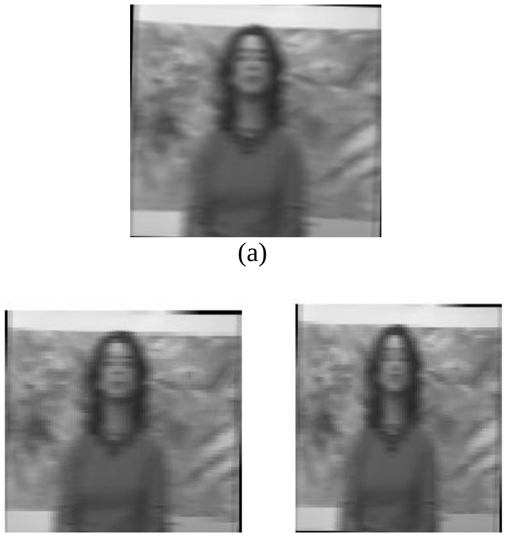
$$\begin{bmatrix} T_x \\ V_x \\ T_y \\ V_y \\ S \\ \theta \end{bmatrix}^t = \begin{bmatrix} 1 & 1 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} T_x \\ V_x \\ T_y \\ V_y \\ S \\ \theta \end{bmatrix}^{t-1} + \begin{bmatrix} 0 \\ N(0,\sigma_{vx}) \\ 0 \\ N(0,\sigma_{vy}) \\ N(0,\sigma_s) \\ N(0,\sigma_\theta) \end{bmatrix}^t \qquad (10)$$

Where T_x and T_y are the translation vectors along x-axis and y-axis, V_x and V_y are the velocity vectors of T_x and T_y, respectively. S and θ are the accumulative scale and rotation factor. $N(0,\sigma_{vx})$, $N(0,\sigma_{vy})$, $N(0,\sigma_s)$ and $N(0,\sigma_\theta)$ are the system noise of V_x, V_y, S and θ respectively and t indicates the time step.

After obtaining the intentional motion estimation, the unintentional motion compensation can be computed as:

$$\begin{bmatrix} x \\ y \\ 1 \end{bmatrix}^t = \hat{S}_t \begin{bmatrix} \cos\hat{\theta} & -\sin\hat{\theta} & -\hat{T}_x \\ \sin\hat{\theta} & \cos\hat{\theta} & -\hat{T}_y \\ 0 & 0 & 1 \end{bmatrix}^t \begin{bmatrix} x \\ y \\ 1 \end{bmatrix}^{t-1} \qquad (11)$$

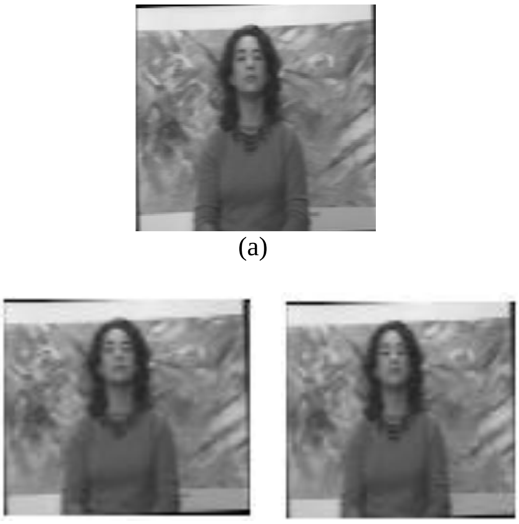
Where $\hat{\theta} = \theta^g - \theta^i, \hat{S} = S^g - S^i, \hat{T}_x = T_x^g - T_x^i$ and $\hat{T}_y = T_y^g - T_y^i$ are the unintentional motion estimation. Superscripts g and i indicate the global motion estimation and intentional motion estimation, respectively.



(a)

(b) (c)

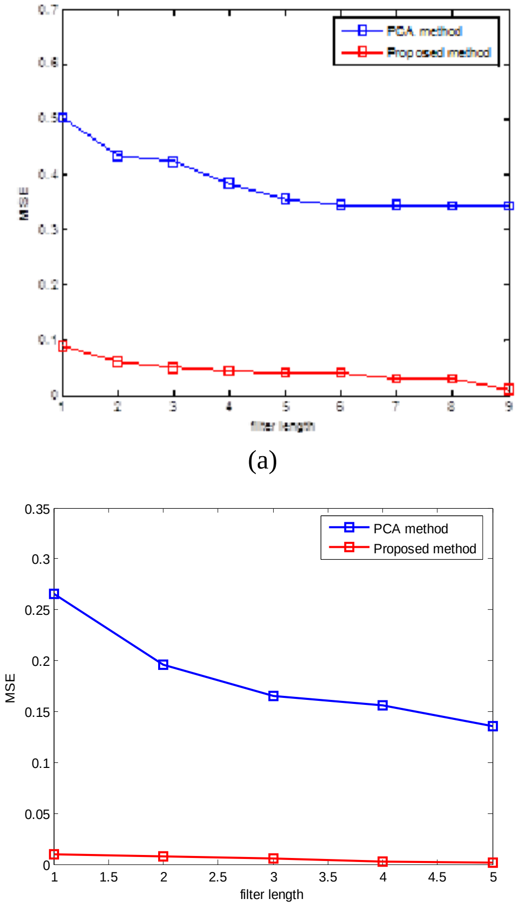
Figure 5: Input frames extracted (a) Original sample, (b) Frame #1 (c) Frame #2



(a)

(b) (c)

Figure 6: Output compensated frames (a) Original sample, (b) Frame #1 (c) Frame #2



(a)

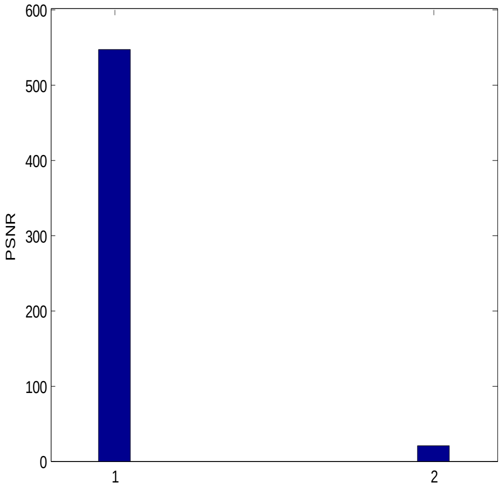
(b)

Figure 7: Graphs for MSE vs filter wavelength for existing and proposed methods of: (a) Frame # 1 and (b) Frame # 2

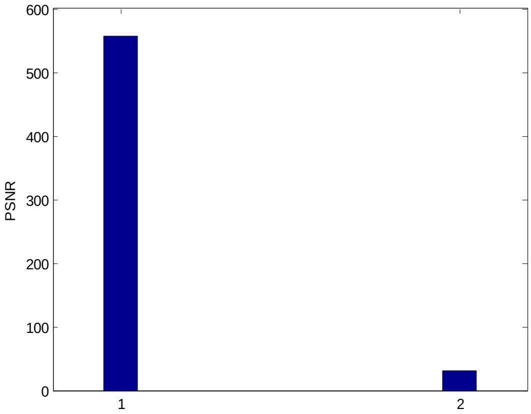
In the first series of experiments, the number of feature points in a particle to see its influence in estimation results. In the second series of experiments, approach was compared with other two video stabilization algorithms based on particle filtering. To evaluate the performance of the approach, ITF (Inter-frame Transformation Fidelity) measure is adopted:

$$ITF = \frac{1}{N_{frame}-1}\sum_{k=1}^{N_{frame}-1} PSNR(K) \qquad (12)$$

where N_{frame} represents the number of video frames.



(a)



(b)

Figure 8: Graphs for PSNR of existing and proposed methods of: (a) Frame # 1 and (b) Frame # 2

$PSNR(K)$ is the Peak Signal-to-Noise Ratio which can be defined as:

$$PSNR(K) = 10\ \log_{10}\frac{I_{max}}{MSE(K)} \qquad (13)$$

Where, I_{max} is the maximum pixel intensity and $MSE(K)$ is the Mean Square Error between consecutive frames. A higher ITF indicates a more accurate estimation.

V. CONCLUSION

From above figures it can be noticed that when compared previous method proposed method is better where the time taken by the video stabilization using PCA and SIFT in particle filter framework algorithm is reduced. The issues leading to the error in the motion estimation are overcome by the proposed robust video stabilization based on particle filtering with weighted feature points technique by taking SURF and applying WLS algorithm. The proposed algorithm has the characteristics of high precision and good robustness.

REFERENCES

[1] K. Uomori, A. Morimura, H. Ishii, T. Sakaguchi, and Y. Kitamura“ Automatic image stabilizing system by full-digital signal processing,” IEEE Trans. on Consumer Electron., vol. 36, no. 3, pp. 510-519, Aug.1990.

[2] T. Kinugasa, N. Yamamoto, H. Komatsu, S. Takase, and T. Imaide, “ Electronic image stabilizer for video camera use,” IEEE Trans. on Consumer Electron., vol. 36, no. 3, pp. 520-525, Aug. 1990.

[3] J. K. Paik, Y. C. Park, and S. W. Park, “An edge detection approach to digital image stabilization based on tri-state adaptive linear neurons,” IEEE Trans. Consumer Electron., vol. 37, no. 3, pp. 521-530, Aug. 1991.

[4] A. Censi, A. Fusiello, and V. Roberto. “Image stabilization by features tracking,” International Conference on Image Analysis and Processing, pp.665- 667, Sep. 1999.

[5] D. Lowe, “Distinctive image features from scale-invariant keypoints,” International Journal of Computer Vision, vol. 60, no.2, pp.91-110, 2004.

[6] Y. Ke, and R. Sukthankar, “PCA-SIFT: A More Distinctive Representation for Local Image Descriptors” International Conference on Computer Vision and Pattern Recognition, vol.2, pp.506-513, Jun. 2004.

[7] C. T. Wang, J. H. Kim, and K. Y. Byun, “Robust digital image stabilization using the Kalman filter,” IEEE Trans. Consumer Electron, vol.55, no.1, pp.6- 14, Feb. 2009.

[8] A. A. Amanatiadis, I. Andreadis, “Digital Image Stabilization by Independent Component Analysis,” IEEE Trans. Instrumentation and Measurement, vol.59, no.7, pp.1755 - 1763, July. 2010,

[9] K. Y. Huang, Y. M. Tsai, C. C. Tsai, L. G. Chen, “Video stabilization for vehicular applications using SURF-like descriptor

and KD-tree," 2010 17th IEEE International Conference on Image Processing (ICIP 2010), pp.3517-3520, Sept. 2010.

[10] H. Bay, T. Tuytelaars and L. V. Gool, "Surf: Speeded UpRobust Features," Computer Vision and Image Understanding (CVIU), vol.110, no.3, pp.346-359, 2008.

K. Madhavi received B.Tech degree in Electronics and Communication Engineering from MLR Institute of Technology in 2011, and presently pursuing masters degree in Digital Electronics and Communication Engineering from GNITS (JNTU, Hyderabad, India). Her research interests are in Image Processing.

B. Sreekanth Reddy received the B.Tech degree in Electronics and Communication Engineering from JBIET, Hyderabad in 2005 and the M.S degree in Electrical Engineering, University of Missouri, Kansas City, USA in 2007. His research interests are in the domain of Low Power VLSI Design.

Ch.Ganapathy Reddy completed B.Tech from RVR and JCOP college of engineering in 1989. Completed ME from OU in Digital systems in 1996. He is the member of IETE, ISTE and IEEE. His research interest are in DSP, Image processing.