Automatic Feature-Based Stabilization of Video with Intentional Motion through a Particle Filter

Carlos R. del-Blanco*, Fernando Jaureguizar, Luis Salgado, and Narciso García

Grupo de Tratamiento de Imágenes, Universidad Politécnica de Madrid,
28040, Madrid, Spain
{cda,fjn,lsa,narciso}@gti.ssr.upm.es
http://www.gti.ssr.upm.es

Abstract. Video sequences acquired by a camera mounted on a hand held device or a mobile platform are affected by unwanted shakes and jitters. In this situation, the performance of video applications, such us motion segmentation and tracking, might dramatically be decreased. Several digital video stabilization approaches have been proposed to overcome this problem. However, they are mainly based on motion estimation techniques that are prone to errors, and thus affecting the stabilization performance. On the other hand, these techniques can only obtain a successfully stabilization if the intentional camera motion is smooth, since they incorrectly filter abrupt changes in the intentional motion. In this paper a novel video stabilization technique that overcomes the aforementioned problems is presented. The motion is estimated by means of a sophisticated feature-based technique that is robust to errors, which could bias the estimation. The unwanted camera motion is filtered, while the intentional motion is successfully preserved thanks to a Particle Filter framework that is able to deal with abrupt changes in the intentional motion. The obtained results confirm the effectiveness of the proposed algorithm.

1 Introduction

Recently, the number of industrial and military applications based on video cameras have dramatically increased due mainly to two factors: the decrease in cost of both video cameras and processing hardware, and their higher processing power that has allowed using complex and efficient algorithms, previously restricted to simulation environments. Many of these applications mount the camera on a hand held device or a mobile platform (car, airplane, etc.), that causes that the acquired video sequences are affected by unwanted shakes and jitters. In this situation the performance of the applications may decrease significantly. To overcome this problem both hardware and digital processing approaches to stabilize the video sequence have been developed. The hardware based approaches use sophisticated motion sensors and an active optical system to compensate the unwanted camera motion. Despite they are the most powerful, their high cost prevent their incorporation in a broad range of applications. The second approach, that is the focus of the work in this paper, is based only on digital analysis of the acquired video sequence, thus reducing significantly the cost. Different techniques have been proposed for digital video stabilization, which differ in the method they use to compute the image motion. Block-matching [1] [2] methods divide a frame into blocks, and compute a motion vector for each one through the searching of the more similar block in the next frame. However, the motion estimation could be biased in low-textured image regions due to the aperture problem [3]. Feature-based methods overcome this problem by computing the motion only in regions that stand out according to a specific image feature. In this context, SIFT features [4] [5] [6] have recently been very popular because of their high efficiency in registration applications. Nevertheless, similar objects in the scene with different rotations or scales could generate erroneous motion estimations, since the SIFT features are invariants to these dimensions.

Once the inter-frame motion has been estimated, it is compensated to stabilize the video sequence. However, the camera motion in a video sequence is a combination of the displacement of the camera, i.e. the intentional motion, and the undesired shaking or jitter, which are the only ones to be filtered to achieve a successfully video stabilization. Several techniques have been proposed to filter the shaking from the intentional motion such as Kalman filter [5] [7] and Motion Vector Integration [6]. However, they do not work properly when the intentional camera motion is fast and abrupt or when the magnitude of the camera shaking is variable along the time. In addition, these techniques typically depend on several user parameters that need a particular setting for each sequence, that severely restricts their applicability.

In this paper a novel digital video stabilization algorithm is proposed, which overcomes the previous problems by computing a robust motion estimation through a variation of the SIFT algorithm adapted to video sequences to be discriminative to scale and orientation, and by performing an automatic camera motion filtering that is able to preserve the abrupt and variable intentional motion. The steps involved in the video stabilization algorithm are: local inter-frame motion is computed by means of a robust feature-based technique. Then, global inter-frame motion is accurately inferred from the estimated local motion through a RANSAC framework robust to erroneous motion estimations. Global motion estimation between the last stabilized frame and the current one is computed from previous global inter-frame motion estimations. This estimation is refined by a global motion minimization technique, that corrects the uncertainties related to the accumulation of the global inter-frame estimations. Finally, global motion is analyzed by means of a Particle Filter framework that automatically infers the intentional motion. As a result, the video sequence is stabilized from camera shakings and jitters, leaving the intentional motion unfiltered.

2 Local Inter-frame Motion

Image motion is computed through a fast feature-based motion estimation technique (FFME) [8], which is robust to noise, aperture problem, illumination changes and small variations of 3D viewpoint. But, unlike the SIFT algorithm, it is not invariant to abrupt scale and rotation changes, but this is an advantage in video motion estimation, since, taking into account that the variations in a video sequence are enough smooth due to the high temporal correlation, the scale and orientation are used as discriminative features, and thus improving the overall motion estimation.

The main steps of FFME [8] are briefly exposed in the following lines. The selection of features or singular points is accomplished by means of three different restrictions. The first one selects image points with a significant gradient magnitude value. Among those selected, the second restriction rejects the image points with low cornerness, that are those located in straight edges, and therefore still affected by the aperture problem. As a result those points that globally stand out by their gradient magnitude and cornerness are retained. The final selection is obtained after applying the third restriction, a non-maximal suppression in the cornerness space that removes those points that are not very significant according to their neighborhood. This final set of singular points is considered the most reliable to estimate the image motion, since the image points that were specially sensitive to the noise and to the aperture problem have been discarded. Each singular point is characterized by a sophisticated descriptor robust to illumination changes and small variations in the 3D viewpoint. To compute the descriptor, an array of gradient phase histograms in the neighborhood of each singular point is calculated. All the histograms are concatenated in a vector to form the descriptor, which is normalized to make it invariant to brightness and contrast changes. Singular points of consecutive images are matched using as the similarity function the Euclidean distance between the corresponding descriptors. Erroneous correspondences are discarded if its similarity function value is too close to the one belonging to the second best correspondence. Finally, motion vectors $\mathbf{mv}^j$ are obtained from the correspondences that fulfill the previous condition. This set of motion vectors forms an accurate and sparse motion vector field, MVF, representing the motion in the image.

3 Global Inter-frame Motion

The global motion is modeled by an affine transformation that relates the pixel coordinates between consecutive images. This geometric model is a suitable approximation for the projective camera model provided that the depth relief of the objects in the scene is small enough compared to the average depth, and the field of view is also small [9].

The backward affine transformation $\mathbf{T}_k$ at time step k is robustly estimated from the set of inlier motion vectors MVF_{In}, which are motion vectors close to the true motion but affected by slight uncertainties. MVF_{In} is computed by means of the combination of RANSAC (Random Sample Consensus) and LMedS (Least Median of Squares) [10] [11], which are robust estimation techniques that successfully discard the outlier motion vectors of MVF (outliers that arise as a consequence of independent moving objects in the scene). The estimation process starts creating N_S subsets of MVF, each one composed by 3 motion vectors randomly selected from MVF. N_S is computed by Equ. (1), which assures, with probability p_s, that at least one of the subsets is free of outliers:

$$N_S = \frac{\log\left(1 - p_s\right)}{\log\left[1 - (1 - \varepsilon)^{N_{MV}}\right]} \tag{1}$$

where ε is the expected maximum fraction of outliers, and N_{MV} is the total number of motion vectors in MVF. For each subset of 3-motion vectors a candidate affine transformation $\widehat{\mathbf{T}}_m; m = 1, ..., N_S$ is obtained by solving the linear equation system:

$$\begin{bmatrix} x_{k-1}^j \\ y_{k-1}^j \\ 1 \end{bmatrix} = \begin{bmatrix} a\ b\ t_x \\ c\ d\ t_y \\ 0\ 0\ 1 \end{bmatrix} \cdot \begin{bmatrix} x_k^j \\ y_k^j \\ 1 \end{bmatrix} \qquad j = 1, 2, 3 \tag{2}$$

where a, b, c, d, t_x and t_y are the parameters of $\widehat{\mathbf{T}}_m$; and x_{k-1}^j, y_{k-1}^j, x_k^j, y_k^j are the coordinates of the points that form j^{th} motion vector $\mathbf{mv}^j = (x_{k-1}^j - x_k^j, y_{k-1}^j - y_k^j)$. Then, the fitting error e_m associated to each $\widehat{\mathbf{T}}_m$ is computed as:

$$\begin{aligned} e_m &= \text{median}(r_l^2) \qquad l = 1, ..., N_{MV} \\ r^l &= d(p_{k-1}^l, \widehat{\mathbf{T}}_m \cdot p_k^l) \end{aligned} \tag{3}$$

where $p_k^l = (x_k^l, y_k^l)$ and $p_{k-1}^l = (x_{k-1}^l, y_{k-1}^l)$ are the points related to the motion vector $\mathbf{mv}^l = p_k^l - p_{k-1}^l$, and r^l is the residual distance computed through the Euclidean distance $d()$. The set of inlier motion vectors related to $\widehat{\mathbf{T}}_m$ is finally calculated by:

$$MVF_m^{In} = \{\mathbf{mv}^l \in MVF | (r^l)^2 \leq th_{In}\} \qquad l = 1, ..., N_{MV} \tag{4}$$

where th_{In} is a threshold that has been computed analyzing the mean square residual distance $\sum_l (r^l)^2$ in sequences without motion, in which, theoretically, the same feature points should have been detected in all frames.

Finally, $\mathbf{T}_k$ is obtained by solving Equ. (2) through the Least Mean Squares algorithm (LMS), as the number of elements of MVF_m^{In} is generally greater than 3. This improves the accuracy of the affine transformation estimation since the uncertainty associated to each inlier motion vector is averaged.

The set of MVF_m^{In} with the least e_m is chosen to be MVF^{In} , i.e. the best set of inlier motion vectors to accurately compute the affine transformation $\mathbf{T}_k$ between the time steps $k - 1$ y k.

4 Global Motion between Distant Frames

The current frame I_k is aligned with respect to a reference frame I_r by multiplying in cascade all the affine transformations corresponding to each intermediate time steps:

$$I_r(\mathbf{c}) = I_k(\mathbf{T}_{r+1} \cdot ... \cdot \mathbf{T}_{k-1} \cdot \mathbf{T}_k \cdot \mathbf{c}) = I_k(T_{(r+1:k)}\mathbf{c}) \tag{5}$$

where $\mathbf{c}$ is a vector representing image pixel coordinates.

However, because of the accumulation of small inaccuracies associated to each affine transformation, the quality of the estimated global motion estimation between distance frames decreases. This is solved by computing an additional affine transformation $\mathbf{T}^c$ that corrects the global motion estimation, as is shown in:

$$I_r(\mathbf{c}) = I_k(\mathbf{T}^c \cdot \mathbf{T}_{(r+1:k)} \cdot \mathbf{c}) = I_k(\mathbf{T}^c_{(r+1:k)} \cdot \mathbf{c}) \tag{6}$$

$\mathbf{T}^c$ is computed through a gradient descent iterative technique based on Gauss-Newton [12] that uses MAD_M as cost function a modified version of the Mean Absolute Deviation (MAD). MAD_M has been designed to be robust to the motion of independent objects that could bias the global motion parameter estimation. In order to compute MAD_M, the image is divided into 8×8 pixel blocks, and the MAD is computed for each block MAD_b. Then, MAD_M is calculated as:

$$MAD_M = \sum_{b=1}^{N_b} MAD_b \cdot w_b \tag{7}$$

where N_b is the total number of blocks in the image, and w_b is a binary weighting factor defined as:

$$w_b = \begin{cases} 1 & \text{if } MAD_b < 2.5 \cdot Th_{med} \\ 0 & \text{otherwise} \end{cases} \tag{8}$$

where $Th_{med} = \text{median}\{MAD_b\}, b = 1, ..., N_b$.

The Gauss-Newton based minimization of MAD_M is restricted to a reduced subspace of affine parameters in the environment of $\mathbf{T}_{(r+1:k)}$, since $\mathbf{T}^c$ is expected to be quite close to it. This allows dramatically reducing the computational cost, and only a few number of iterations is necessary to reach a proper solution.

5 Intentional Motion

The global motion is a combination of intentional motion, that arises from the movement of the user or camera platform, and undesired shaking or jitter. To achieve a successfully video stabilization, a Particle Filter framework [13] is used to filter the undesired camera motion, and thus obtaining only the intentional motion. The proposed approach allows automatically addressing smooth and abrupt intentional motions, unlike Kalman-based approaches. Each affine parameter related to the intentional motion is addressed independently by a particle filter. Following the probabilistic state-space formulation, the state vector $\mathbf{x}_k = [a, da]$ contains respectively an affine parameter (the parameter 'a' of the affine matrix has been chosen as an example) and its corresponding temporal derivative at each time step k. The sequence of state vectors $\mathbf{x}_k, k \in \mathbb{N}$ represents the evolution of an affine parameter related to the intentional camera motion along the video sequence. The affine parameter at time step k is estimated by the MAP (Maximum A Posteriori):

$$\mathbf{x}_k^{MAP} = \arg\max_{x_k} p(\mathbf{x}_k|\mathbf{z}_{1:k}) \tag{9}$$

where $\mathbf{z}_{1:k}$ are the sequence of measures from the time step 1 until k, i.e. the global motion estimations and their corresponding temporal derivatives; and $p(\mathbf{x}_k|\mathbf{z}_{1:k})$ is the posterior pdf (probability density function), which according to the Particle Filter Framework is approximated by a set of particles $\{\mathbf{x}_k^i, i = 0, 1, ..., N_p\}$ with associated weights $\{w_k^i, i = 0, 1, ..., N_p\}$ as shown by:

$$p(\mathbf{x}_k|\mathbf{z}_{1:k}) \approx \sum_{i=1}^{N_p} \omega_k^i \cdot \delta(\mathbf{x}_k - \mathbf{x}_k^i) \tag{10}$$

where $\delta(\mathbf{x})$ is the Kronecker delta function. The weights are recursively computed as:

$$\omega_k^i = \omega_{k-1}^i \cdot \frac{p(\mathbf{z}_k|\mathbf{x}_k^i)p(\mathbf{x}_k^i|\mathbf{x}_{k-1}^i)}{q(\mathbf{x}_k^i|\mathbf{x}_{k-1}^i, \mathbf{z}_k)} \tag{11}$$

where $p(\mathbf{x}_k^i|\mathbf{x}_{k-1}^i)$ is the prior pdf that models the system dynamics, i.e. predicts the expected evolution of the affine parameter for the next time step; $p(\mathbf{z}_k|\mathbf{x}_k^i)$ is the likelihood function that describes the system measure model, which is used to correct the affine parameter prediction; and $q(\mathbf{x}_k^i|\mathbf{x}_{k-1}^i, \mathbf{z}_k)$ is the importance density function that, taking into account the last measure (the corresponding affine parameter related to the last global motion estimation), performs a particle resampling to assure that the posterior pdf will be correctly approximated.

Since $p(\mathbf{z}_k|\mathbf{x}_k^i)$ is not normalized for each particle, the computed weights must be normalized at the end of each iteration such that $\sum_{i=1}^{N_p} \omega_k^i = 1$. In addition, at the beginning of each iteration the particles are resampled [13] to avoid the degeneracy problem, in which all but one particle have negligible weight after a few iterations.

The prior pdf $p(\mathbf{x}_k^i|\mathbf{x}_{k-1}^i)$ models the three different situations that can occur in real video stabilization applications: no intentional motion, smooth intentional motion and abrupt intentional motion. Equ. (12) computes the particle predictions $\widehat{\mathbf{x}}_k^i$, which represent the discrete approximation of $p(\mathbf{x}_k^i|\mathbf{x}_{k-1}^i)$.

$$\begin{aligned}
\widehat{\mathbf{x}}_k^i &= \begin{cases} \mathbf{A}_k^{nim} \cdot \mathbf{x}_{k-1}^i + \frac{d_e}{2} \cdot p_{aim} \text{ if } i \leq N_p \\ \mathbf{A}_k^{sim} \cdot \mathbf{x}_{k-1}^i + \frac{d_e}{2} \cdot p_{aim} \text{ otherwise} \end{cases} \\
\mathbf{A}_k^{nim} &= \begin{bmatrix} 1 & 0 \\ 0 & 0 \end{bmatrix} \\
\mathbf{A}_k^{sim} &= \begin{bmatrix} 1 & 1 \\ 0 & 1 \end{bmatrix}
\end{aligned} \tag{12}$$

where A_k^{nim} is the system dynamics in the case of no intentional motion; A_k^{sim} is the system dynamics in the case of smooth intentional motion; $i \leq N_p$ gives out the particles between the situations of no intentional camera motion and smooth intentional camera motion; d_e is the Euclidean distance between the affine parameters related to the estimated intentional motion and the global motion in the previous time step; and p_{aim} is the probability that an abrupt intentional motion occurs, computed as:

$$p_{aim} = \begin{cases} 1 \text{ if } d_e > 2.5\sqrt{\mathbf{R}_k(0,0)} \\ 0 \text{ otherwise} \end{cases} \tag{13}$$

where $\mathbf{R}_k$ is the noise covariance matrix related to the measure (global motion) at the time step k, which is calculated as:

$$\mathbf{R}_k = \text{diag}\left(\frac{1}{N_w}\sum_{h=0}^{N_w-1}(\mathbf{z}_{k-h} - \mathbf{z}_{k-h-1})^2\right) \tag{14}$$

where diag() creates a diagonal matrix, and N_w is the number of previous consecutive measures used to computed $\mathbf{R}$. N_w is the only user-defined parameter of the Particle Filtering framework, that determines the maximum length of the shaking that is able to successfully filter. An example is presented in Section 6 to show the relation between N_w and the corresponding motion filtering.

The importance density function $q(\mathbf{x}_k^i|\mathbf{x}_{k-1}^i, \mathbf{z}_k)$ is computed as a filtered version of $p(\mathbf{x}_k^i|\mathbf{x}_{k-1}^i)$ by means of a Kalman filter. According to this, the resampling $\mathbf{x}_k^i \sim q(\mathbf{x}_k^i|\mathbf{x}_{k-1}^i, \mathbf{z}_k)$ is carried out by:

$$\begin{aligned}
&\widehat{\mathbf{x}}_k^i \sim p(\mathbf{x}_k^i|\mathbf{x}_{k-1}^i) \\
&\mathbf{Q}_k = \mathbf{R}_k + \mathbf{U}[0, 10^{-6}] \\
&\widehat{\mathbf{P}}_k = \mathbf{A}_k\mathbf{P}_{k-1}\mathbf{A}'_k + \mathbf{Q}_k \\
&\mathbf{K}_k = \hat{\mathbf{P}}_k\mathbf{H}'(\mathbf{H}\widehat{\mathbf{P}}_k\mathbf{H}' + \mathbf{R}_k)^{-1} \\
&\mathbf{x}_k^i = \widehat{\mathbf{x}}_k^i + \mathbf{K}_k(\mathbf{z}_k - \mathbf{H}\widehat{\mathbf{x}}_k^i); \\
&\widehat{\mathbf{P}}_{k+1} = (\mathbf{P}_k - \mathbf{K}_k\mathbf{H})\widehat{\mathbf{P}}_k;
\end{aligned} \tag{15}$$

where $\mathbf{Q}$ is the noise covariance matrix related the system dynamics; $\mathbf{U}[0, 10^{-6}]$ is a uniform random variable that allows a proper camera motion filtering independently of the level of $\mathbf{R}_k$ and its variations along the time; $\mathbf{P}_k$ is the error covariance matrix; $\widehat{\mathbf{P}}_k$ is the prediction of the error covariance matrix; $\mathbf{K}_k$ is the Kalman gain; $\mathbf{A}_k$ can be $\mathbf{A}_k^{sim}$ or $\mathbf{A}_k^{nim}$; and $\mathbf{H}$ is the measure model given by the identity matrix.

The likelihood function $p(\mathbf{z}_k|\mathbf{x}_k^i) = p_{ma}^i p_{med}^i$ is designed to give more relevance to the particles that are closed to the moving average of the last N_t time steps, that is represented by p_{ma}. While this provides satisfactory results for moderate variations of the global motion, its behavior is not suitable for high variations which can bias the estimation. For this reason, it is combined with a moving median p_{med}, which is able to filter the global motion peaks. The expressions for p_{ma}^i and p_{med}^i are given by the Gaussian functions:

$$\begin{aligned}
&p_{ma}^i = \frac{1}{\sqrt{2\pi\mathbf{R}_k}} \cdot e^{-{d_{ma}^i}^2/(2\mathbf{R}_k)} \\
&p_{med}^i = \frac{1}{\sqrt{2\pi\mathbf{R}_k}} \cdot e^{-{d_{med}^i}^2/(2\mathbf{R}_k)}
\end{aligned} \tag{16}$$

where:

$$\begin{aligned}
&d_{ma}^i = \widehat{\mathbf{x}}_k - \text{mean}\{\mathbf{z}_h\} \\
&d_{med}^i = \widehat{\mathbf{x}}_k - \text{median}\{\mathbf{z}_h\} \\
&h = (k - N_t + 1), ..., k
\end{aligned} \tag{17}$$

The MAP estimation for each affine parameter is used to form the intentional motion affine transformation $\mathbf{T}_k^{IM}$, which is used in the next time step to stabilize I_{k+1}, as shown in:

$$I_{k+1}^{IM}(\mathbf{c}) = I_{k+1}(\mathbf{T}_k^{IM}\mathbf{T}_{(1:k+1)}^{c}\mathbf{c}) \tag{18}$$

As a result, $\{I_k^{IM}, k = \mathbb{N}\}$ is the sequence of stabilized images, in which the undesirable motion has been removed while keeping unmodified the intentional motion.

6 Results

The performance of the global inter-frame motion estimation has been tested through the PSNR (Peak Signal to Noise Ratio) [6], which measures the similarity between two images. The higher the PSNR value is, the more similar are the two images. Therefore, the computation of the PSNR measure for an inter-frame compensated sequence should be higher and more stable along the time than the sequence without compensation, providing that the computed global motion estimation is accurate and robust to errors. The video sequence 'corridor' (Fig. 6) has been used to compute the PSNR measure. This sequence has been acquired by a person that firstly stands on a fixed position in a corridor, and later walks towards the end of it. This induces a lower level of shaking in the first part of the sequence (until the frame 50) than in the second one. In addition, a zoom effect appears from the frame 50 as a consequence of the walking. Figure 1 shows that the PSNR measures related to the compensated sequence (dotted line) are higher and more stable than the sequence without compensation (solid line), demonstrating the good performance of the global inter-frame motion estimation technique.

The particle filter framework has been tested with synthetic data to simulate two different situations. The first one is an abrupt change in whatever of the affine parameters related to the intentional motion, which is modeled by a step

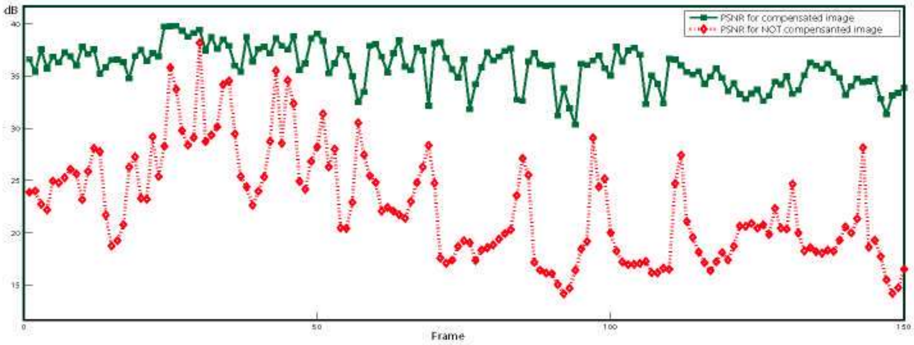

Fig. 1. PSNR measures for the video sequence 'corridor' without stabilization and with inter-frame stabilization, respectively represented by the dotted and solid line

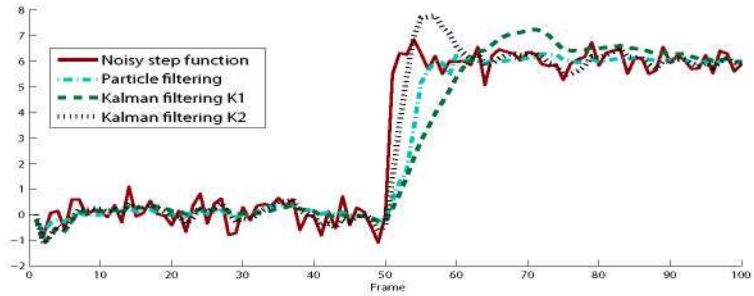

Fig. 2. Particle filter and Kalman filter responses for a step function, that simulates an abrupt change in the intentional motion

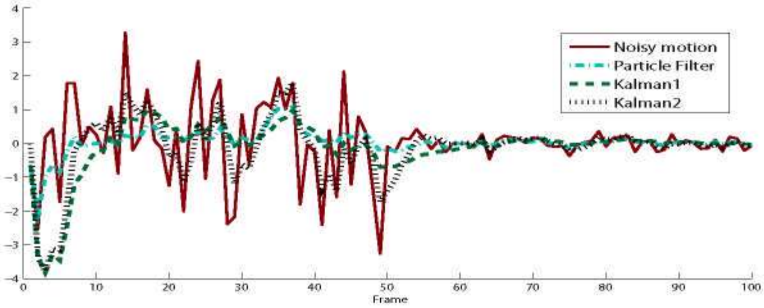

Fig. 3. Particle filter and Kalman filter responses for a function with variable level of noise, that simulates a variable level of shaking

function. In addition, a Gaussian noise of mean zero and standard deviation 0.5 has been added to it to model the camera shaking. Figure 2 shows the results of applying the proposed particle filter with $N_w = 10$ and $N_p = 20$, and two Kalman filters K1 and K2 (use for comparative purposes) with the same $\mathbf{R}$ (noise covariance matrix related to the measure) used in the particle filter, and with $\mathbf{Q}_{K1} = \mathbf{R} \cdot 10^{-4}$ and $\mathbf{Q_{K2}} = \mathbf{R} \cdot 10^{-2}$ (noise covariance matrices related to the system dynamics) respectively for each one. The particle filter achieves the best response to the noisy step function. This is verified through the computation of the MSE between each filtered result and the step function without the Gaussian noise, obtaining $MSE_{PF} = 29.3$ for the particle filter, the minimum error, and $MSE_{K1} = 60.1$ and $MSE_{K2} = 38.6$ for the Kalman filters K1 and K2 respectively.

In the second situation a person is walking (higher shaking), and suddenly stops, keeping standing on a fix position (lower shaking). The walking is model by a Gaussian distribution with mean zero and variance 1.5 (first 50 time steps), while the standing is model by a Gaussian distribution of mean zero and variance 0.25 (the 50 last ones). Figure 3 shows the results computed with the same parameters used in the previous situation. Again, the particle filter has a superior

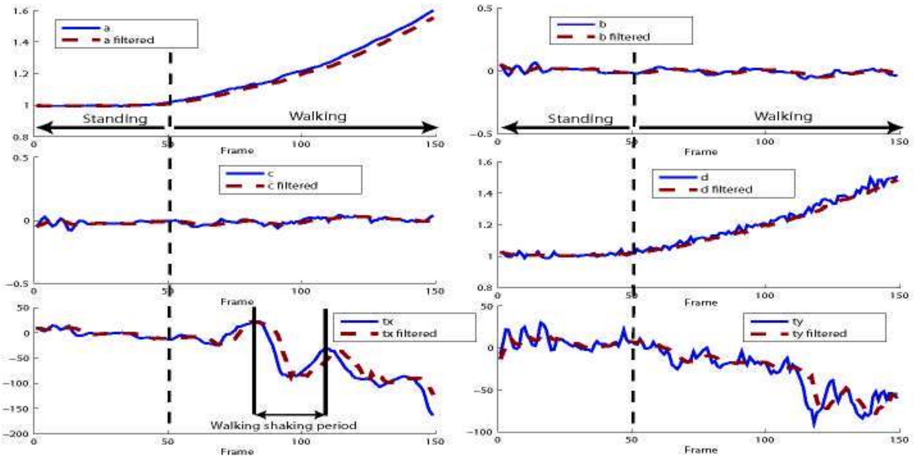

Fig. 4. Variations of the affine parameters a,b,c,d,tx and ty, related to the camera motion along the time, and their result after the particle filtering with $N_w = 10$

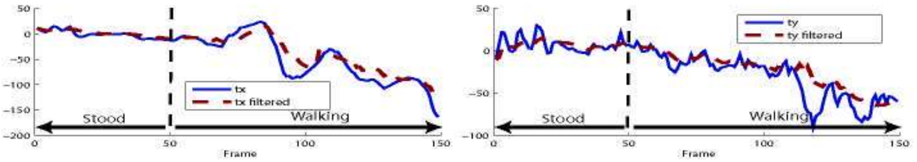

Fig. 5. Variations of the affine parameters tx and ty related to the camera motion along the time, and their result after the particle filtering with $N_w = 35$

performance, that is confirmed with the MSE measures: $MSE_{PF} = 19.8$ for the particle filter and $MSE_{K1} = 42$ and $MSE_{K2} = 53.1$ for the Kalman filters K1 and K2 respectively.

The overall video stabilization algorithm is tested with the previous sequence 'corridor', showing the filtering capability of the particle filter and the dependence on the parameter N_w. In the Fig. 4 is shown the evolution of the affine matrix parameters (a,b,c,d,tx and ty) related to the global camera motion and the intentional motion (i.e. after the particle filter processing) for $N_w = 10$ and $N_p = 20$. The particle filter properly smooths the camera shaking for the parameters a, b, c and d. However, it is not able to smooth enough the long-time shaking in tx and ty from the frame 50, that is when the person that carries the camera is walking. This is due to the value of N_w is less than the period of the walking shaking, which is approximately 35 as shown the Fig. 4. Therefore, setting N_w to a value equal o greater than 35 the filtering of the shake is improved, as shown in Fig. 5 for $N_w = 35$, where only the results concerning to the parameters tx and ty are depicted. Finally, Fig. 6 shows a selection of frames (respectively 16, 88 and 120) of the sequence 'corridor' before (upper row) and after (bottom row) the video stabilization with $N_w = 35$.

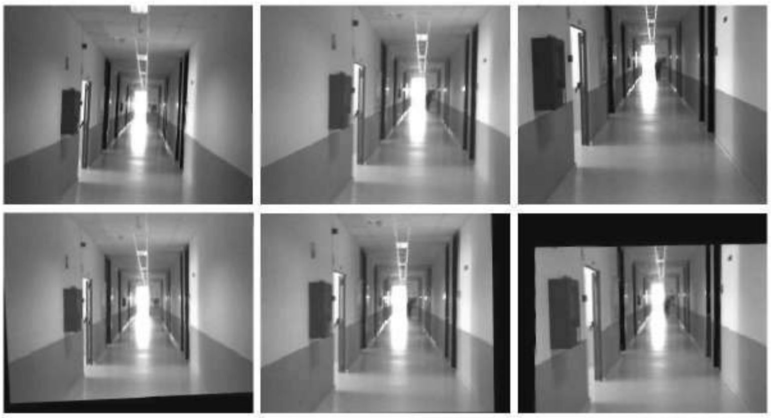

Fig. 6. A selection of frames (respectively 16, 88 and 120) of the sequence 'corridor' before (upper row) and after (bottom row) the video stabilization with $N_w = 35$

7 Conclusions

A novel video stabilization technique has been presented that satisfactorily preserves complex intentional motions, that usually are incorrectly filtered as camera shakings. This has been accomplished by a Particle Filter framework that automatically analyzes the video sequence and generates hypothesis about the most probable intentional motions. The stabilization quality has been ensured by an accurate global motion estimation, where the image motion is computed through a fast feature-based technique that uses the scale and the orientation information to correctly match features of similar objects. Additionally, the proposed RANSAC framework has allowed to robustly compute the global camera motion, despite independent moving objects and deficient image motion estimations. The obtained results corroborate the efficiency of the proposed technique for stabilizing sequences with complex intentional motions.

Acknowledgements

This work has been partially supported by the Ministerio de Ciencia e Innovación of the Spanish Government under project TEC2007-67764 (SmartVision) and by the Comunidad de Madrid under project S0505/TIC-0223 (Pro-Multidis).

References

1. Auberger, S., Miro, C.: Digital Video Stabilization Arquitecture for Low Cost Devices. In: Proc. ISPA, pp. 474–479 (2005)
2. Vella, F., Castorina, A., Mancuso, M., Messina, G.: Digital Image Stabilization by Adaptive Block Motion Vector Filtering. Trans. IEEE on Consumer Electronics 48(3), 796–801 (2002)

3. Wechsler, H., Duric, Z., Fayin, L., Cherkassky, V.: Motion estimation using statistical learning theory. IEEE Trans. on Pattern Analysis and Machine Intelligence 26(4), 466–478 (2004)
4. Rong, H., Rongjie, S., I-fan, S., Wenbin, C.: Video Stabilization Using Scale-Invariant Features. In: Proc. ICIV, pp. 871–877 (2007)
5. Yang, Y., Schonfeld, D., Chen, C., Mohamed, M.: Online Video Stabilization Based on Particle Filters. In: Proc. ICIP, pp. 1545–1548 (2006)
6. Battiato, S., Gallo, G., Puglisi, G., Scellato, S.: SIFT Features Tracking for Video Stabilization. In: Proc. ICIAP, pp. 825–830 (2007)
7. Litvin, A., Konrad, J., Karl, W.C., Mohamed, M.: Probabilistic video stabilization using Kalman filtering and mosaicking. In: Proc. SPIE, pp. 1545–1548 (2003)
8. del Blanco, C.R., Jaureguizar, F., Salgado, L.: Motion estimation through efficient matching of a reduced number of reliable singular points. In: Proc. SPIE, vol. 6811, p. 68110N(1–12) (2008)
9. Hartley, R.I., Zisserman, A.: Multiple View Geometry in Computer Vision, 2nd edn., pp. 153–156. Cambridge University Press, Cambridge (2004)
10. Stewart, C.V.: Robust parameter estimation in computer vision. SIAM Reviews 41(3), 513–537 (1999)
11. Meer, P., Stewart, C.V., Tyler, D.: Robust computer vision: an interdisciplinary challenge. Computer Vision and Image Understanding 78(1), 1–7 (2000)
12. Wang, Y., Ostermann, J., Zhang, Y.: Video Processing and Method per Communication. Prentice-Hall, Englewood Cliffs (2002)
13. Arulampalam, S., Maskell, S., Gordon, N.J., Clapp, T.: A Tutorial on Particle Filters for On-line Non-linear/Non-Gaussian Bayesian Tracking. Trans. IEEE on Signal Processing 50(2), 174–188 (2002)