Noname manuscript No.
(will be inserted by the editor)

Real-time model-based video stabilization for micro aerial vehicles

Wilbert G Aguilar · Cecilio Angulo

Received: date / Accepted: date

Abstract The emerging branch of Micro Aerial Vehicles (MAVs) has attracted a great interest for their indoor navigation capabilities, but they require a high quality video for tele-operated or autonomous tasks. A common problem of on-board video quality is the effect of undesired movements, so different approaches solve it with both mechanical stabilizers or video stabilizer software. Very few video stabilizer algorithms in the literature can be applied in real-time but they do not discriminate at all between intentional movements of the tele-operator and undesired ones. In this paper, a novel technique is introduced for real-time video stabilization with low computational cost, without generating false movements or decreasing the performance of the stabilized video sequence. Our proposal uses a combination of geometric transformations and outliers rejection to obtain a robust inter-frame motion estimation, and a Kalman filter based on an ANN learned model of the MAV that includes the control action for motion intention estimation.

Keywords Video stabilization · Micro aerial vehicles · Real-time · RANSAC · Modelling · Motion intention · Kalman filter

1 Introduction

Unmanned Aerial Vehicles (UAVs) are used in several applications like surveillance, mapping, transport or rescue, for their versatility. Micro aerial vehicles

W. G. Aguilar
Automatic Control Department (ESAII), Universitat Politècnica de Catalunya, Barcelona 08028, Catalonia, Spain
Tel.: +34-93-4016976
Fax: +34-93-4017045
E-mail: wilbert.aguilar@upc.edu

C. Angulo
Automatic Control Department (ESAII), Universitat Politècnica de Catalunya, Barcelona 08028, Catalonia, Spain

(MAVs), a class of UAVs, has gained prominence for its flight maneuverability in closed spaces, time and cost of manufacturing and maintenance, and safety for application in human robot interaction. Robust guidance, navigation, and control systems [21] are required for these platforms, and their performance depends on the input data of on-board sensors and cameras.

One common problem in video sequences captured on aerial vehicles is undesired motion generated during flight for their complex dynamic, and the effect is higher in MAVs. Annoying rotations and translations due to aerodynamic characteristics appear in the sequence of images, increasing the difficult of, usually remote, control of micro aerial vehicles.

1.1 Related work

There are multiple techniques in the literature designed to compensate undesired movements of the camera [7,19,30]. Recently, the video stabilization algorithm *L1 Optimal* provided by the YouTube editor was introduced in [16]. Another interesting just released proposal is the Parrot's *Director Mode*, implemented as an iOS application (iPhone Operative System) for post-processing of videos captured with Parrot's AR.Drones.

Most of the previous video stabilization techniques contain three phases:

- Inter-frame motion estimation
- Motion intention estimation
- Motion compensation

1.1.1 Inter-frame motion estimation

The approaches used to estimate the parameters relating two consecutive frames are: optical flow [10] or geometric transformation models [22,34,32].

In both approaches, feature points are detected and described for being used instead of all pixels from each image. A list of algorithms performing this challenge can be found in the literature [17,9,27], but Binary Robust Invariant Scalable Keypoints (BRISK) [23], Fast Retina Keypoint (FREAK) [5], Oriented FAST and Rotated BRIEF (ORB) [29], Scale Invariant Feature Transform [24] (SIFT) and Speed Up Robust Feature (SURF) [8,25] are common in the computer vision field.

The second phase is the matching of feature points between consecutive frames, and the successful of the motion estimation process depends on the correct pairing. An additional robust method is required to remove incorrectly matched points (outliers). RANdom SAmple Consensus (RANSAC) is a reliable iterative technique for outliers rejection on a mathematical model [14,31, 11].

1.1.2 Motion intention estimation

In order to obtain stable videos instead of static videos, the inter-frame motion parameters are accumulated throughout the whole sequence. This accumulative motion is composed by desired and undesired movements.

The intentional movements are estimated by suppressing high frequency jitters from the accumulative global motion. Several motion smoothing methods are available to find the motion intention, such as particle filter [34], Kalman filter [32], or Gaussian filter [26].

These video stabilization algorithms are focus on the tracking of feature points to compensate the movement respect to them. Based on this idea, the objective of motion smoothing is to obtain the intentional movement of the feature points to compute the inter-frame motion parameters after that.

An alternative option is [1], where we propose estimating the motion intention of the motion parameters instead of the feature points.

1.1.3 Motion compensation

Finally, the current frame is warped using parameters obtained from the motion intention to generate a stable video sequence.

1.2 Phantom movements

This phenomenon was presented in a previous article [1]. The phantom movements correspond mainly to false displacement generated by video stabilization algorithm in the scale and/or translation parameters due to the compensation of the eliminated high frequency movements in the motion smoothing process. Real movements can be removed and/or delays can be introduced, but both effects are defined as phantom movements.

Previous works on video stabilization achieve good results eliminating undesired movements in images captured with hand-held devices and complex systems, but generate phantom movements. There is no problem for post-processing applications, but for tele-operated system, phantom movements represent a dangerous issue.

In the same paper [1], we presented a proposal based on a low pass filter and the use of the control action as logical gate with hysteresis.

1.3 Our approach

In the present article, we propose a combination of the projective and affine model to obtain a reliable transformation (robustness) with a lower computational cost (fastness) and a lower deformation.

Additionally, we propose an algorithm based on the approach of [1] for estimating the intentional motion of parameters and no feature points. In

contradistinction to [1], we use Kalman filter and the model of the MAV, where the intentional control actions are uncoupled from unintentional motion.

The model of the MAV includes the control action, solving the issue of phantom movements and, at the same time, minimizing the number of previuos frame to one. The algorithm depends only on the last frame and can be applied in real time without delays or decreasing the performance

Our proposed technique can be applied in real-time for tele-operated micro aerial vehicles during indoor flights because the modeling was carried out with data of indoor tests. Outdoor flights implicate additional problems as turbulence out of the scope of this paper.

The rest of this paper is organized as follows: Our proposal for estimating inter-frame motion parameters based on a combined transformation model and an outliers rejection algorithm is explained in the next Section. In the Section 3, we introduce a novel technique of motion intention estimation based on the model of the MAV that includes the input control action. Experimental results and conclusions are finally in the Section 4 and 5. presented.

2 Proposal for inter-frame motion estimation

The geometric transformation [13,18,15] is used to describe the mathematical relationship between every two consecutives images in the video. One image is the reference and the other one is the frame to be processed. This mathematical relationship can be represented as below:

$$\mathbf{I}_{sp} = \mathbf{H}_t \cdot \mathbf{I}_t \tag{1}$$

where $\mathbf{I}_{sp} = [x_{sp}, y_{sp}, 1]^T$ and $\mathbf{I}_t = [x_t, y_t, 1]^T$ are the coordinates of the interest points at the reference image and the uncompensated image, respectively, and $\mathbf{H}_t$ is the geometric transformation matrix.

This matrix contains motion parameters that depend on the model used to represent the warping effect generated between two frames during the movement of the camera. Parametric motion models can be 2-D or 3-D. The 2-D models are widely used in video stabilization algorithms and the most common are: translation, affine, nonreflective similarity, and projective model. The last one is known as homography. We use a combination of projective and affine transformation to obtain a robust inter-frame motion.

2.1 Using the projective transformation

In the subsection 1.1, we cited several approaches for computing feature points. Our video stabilization algorithm uses SURF and each detected point has an associated 64-dimensional descriptor. The inter-frame geometric transformation is based on the computed interest points, represented in the 64-dimensional space of the SURF descriptors, for both frames. The points of one frame must

be matched with their correspondence in the other frame. The matching process searches the nearest neighbors, i.e. the pair of feature points with the minimum euclidean distance in this 64-dimensional space.

For images captured in uncontrolled conditions, the matched process generates false correspondences. One option to solve this issue is searching the nearest neighbors, i.e. the pair of feature points with the minimum euclidean distance between their descriptors, but this matched process is not reliable.

In [1], RANSAC is used for rejection of pair of points incorrectly matched. We propose a similar approach but using the projective transformation instead of the affine tranform as mathematical model of RANSAC. The affine transformation is used later.

The projective transformation, so-called homography, contains six parameters, three rotations and three translations. The matrix transformation is composed by eight linearly independent parameters.

$$\mathbf{H}_t = \begin{bmatrix} h_{11} & h_{12} & h_{13} \\ h_{21} & h_{22} & h_{23} \\ h_{31} & h_{32} & 1 \end{bmatrix} \tag{2}$$

The Algorithm RANSAC, showed in 1, is applied after to match feature points. In each iteration, the projective transformation $\mathbf{H}_j$ is estimated based on four pairs of matched points randomly selected and employed to warp the j_{th} frame. We are using, as cost function J_j, the gray level difference between the reference frame and the current frame warped by using $\mathbf{H}_j$.

Finally, we select the parameters from the projective transformation with the minimum cost function:

$$\arg \min_{(\phi,s,t_x,t_y)} \sum_j \left| \mathbf{Frame}'_j - \mathbf{Frame}_{sp} \right| \tag{3}$$

where $\mathbf{Frame}'_j$ and $\mathbf{Frame}_{sp}$ are the warped and reference frame.

1: **for** $j = 1$ to N **do**
2: j_{th} projective transformation estimation: $\mathbf{H}_j$
3: j_{th} warping of the i_{th} frame: $\mathbf{Frame}'_j$
4: j_{th} cost function computation: $J_j = \left| \mathbf{Frame}'_j - \mathbf{Frame}_{sp} \right|$
5: **end for**
6: Selection of parameters of $\mathbf{H}_{opt}$ for cost function minimization: $\arg\min_{(\phi,s,t_x,t_y)} \sum_j \left| \mathbf{Frame}'_j - \mathbf{Frame}_{sp} \right|$

Algorithm 1: RANSAC algorithm based on cost function $\sum_j J_j$.

2.2 Defining the reference frame

It is important to specify the frame to be compensated and the frame to be used as reference in the algorithm. The current frame will be warped by motion

compensation obtaining a stable sequence in the output video, but there are different alternatives for the reference.

An experimental comparative study has been carried out in [2] on three candidates to reference frame: the initial frame ($\mathbf{Frame}_{sp} = \mathbf{Frame}_0$), the previous frame ($\mathbf{Frame}_{sp} = \mathbf{Frame}_{i-1}$), and the compensated previous frame ($\mathbf{Frame}_{sp} = \mathbf{Frame}'_{i-1}$). The analysis of the three proposed approaches was conducted by using data obtained from an on-board camera of a micro aerial vehicle. Based on mean square error (MSE) between monochromatic images with size $M \cdot N$,

$$\mathrm{MSE(k)} = \frac{1}{\mathrm{M \cdot N}} \sum_{\mathrm{i=0}}^{\mathrm{M-1}} \sum_{\mathrm{j=0}}^{\mathrm{N-1}} \|\mathbf{I}_{\mathrm{k}}(\mathrm{i,j}) - \mathbf{I}_{\mathrm{k-1}}(\mathrm{i,j})\|^2 \tag{4}$$

the obtained results showed that the previous frame is the best candidate to reference.

2.3 Using the affine transformation

For hand-held cameras and on-board monocular vision devices, most of the undesired movements and parasitic vibrations in the image are considered significant only around the roll axis. The affine model is the selected geometric model of these movements for three reasons:

- The affine model represent the main undesired movements of cameras on micro aerial vehicles.
- In spite of the reliability of the projective transformation is higher as model of RANSAC, the deformation of the frame warped with the affine transformation is lower and the final video is more stable.
- Relevant motion parameters can be extracted directly from the transformation matrix. This parameters are essential for estimating the motion intention.

The motion parameters of the model are: two translations in the plane parallel to the image, roll rotation ϕ about the axis perpendicular to the xy plane, and scale s that is proportional to the motion in the roll axis orientation.

$$\mathbf{H}_t = \begin{bmatrix} s\cos(\phi) & -s\sin(\phi) & t_x \\ s\sin(\phi) & s\cos(\phi) & t_y \\ 0 & 0 & 1 \end{bmatrix} \tag{5}$$

In the affine model, there are two possible angles: $arctan(\frac{\mathbf{H}_t(2,1)}{\mathbf{H}_t(1,1)})$ and $arctan(-\frac{\mathbf{H}_t(1,2)}{\mathbf{H}_t(2,2)})$. We estimate the mean angle adjustable to these values. This model is called nonreflective similarity and is a particular case of the affine model.

2.4 Using a combination of transformations

Some techniques of fast video stabilization employ smoothing trajectories of interest points, however this approach requires a continuous 3D pose estimation. For real-time applications, this method is not recommendable due to the high computational cost required to estimate the 3D pose in each point.

Instead, our approach use the affine transformation based on the projective transformation.

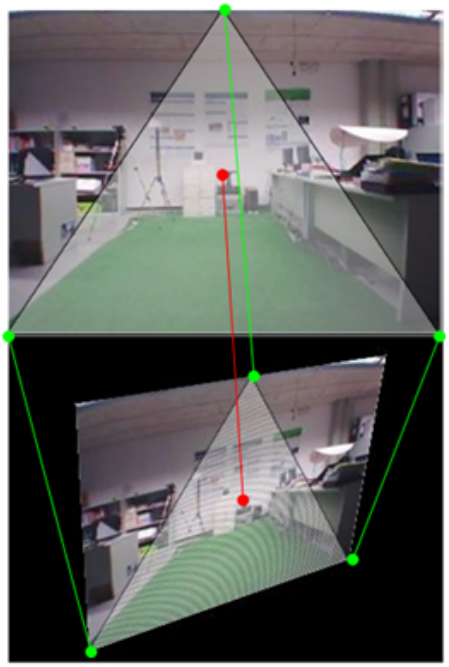

Fig. 1 Reference and warped points.

On the one hand, the homography is more reliable in RANSAC than the affine model, but, the ITF (Inter-frame fidelity) measured between consecutive frames stabilized using this homography is lower than using the affine model. The projective model contains three rotations that increases the image deformation.

Therefore, our technique obtains the best homography matrix using RANSAC as explained in the Section 2.1. Next, three reference points are selected to find the angle and scale, and one more to obtain the translation in the 2D axis. The projective transformation used to estimate the affine model lets compute a mean rotation angle of the image, and reduces the cumulative error.

Three reference points (points in the border of Figure 1) are situated in strategic locations in order to maximize the captured area. If the points are

near, the region enclosed by them is small, hence, it is desired the longest distance between them. There are several options to locate three points in a rectangular image with the maximum distance between them. Considering that we are using an on board camera of a micro aerial vehicle, the triangular option in Figure 2 has been chosen. The reason of this choice is that the on board camera of the MAV should mainly be focus on the lower front region of the scene when flying. Another option is to calculate the mean value between the angles of the affine transformation estimated with each distribution of interest points. However, depending on the sequence, it would generate a jitter effect in the output.

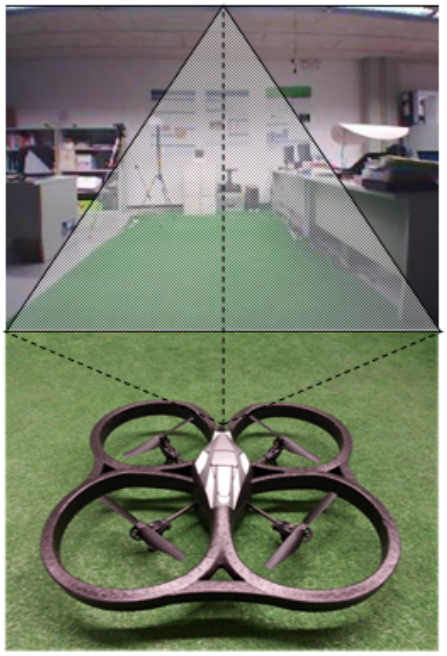

Fig. 2 Area of interest.

After computing the affine parameters without cumulative error using the homography as reference, we estimate the translation model from the fourth point located in the center of the image as reference and its correspondence estimated with the geometric transformation. The use of this fourth reference point guarantees stabilization respect to the center of the image. By joining the estimated transformations, the matrix of compensation is derived. When applied on the current frame, a compensated frame similar to the reference frame is obtained. Hence, inter-frame movements are minimized from the full video sequence to obtain a scene as similar as possible to the reference frame, compensating the undesired movement one frame at a time.

3 Model based motion intention estimation

Our approach based on a combination of geometric transformations obtains a reliable inter-frame motion estimation, and a high performance as video stabilizer in static scenes [2,3]. However, our objective is to achieve this robustness for real-time video stabilization in micro-aerial vehicles.

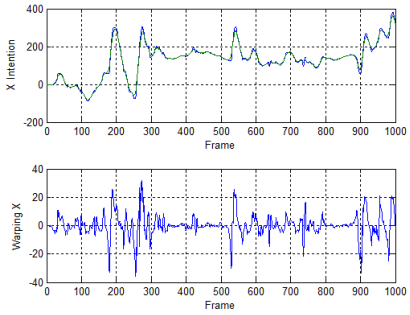

Fig. 3 Translation in the x-axis. Top: Motion intention signal estimated with the low-pass filter. Down: High frequency signal to be compensated

During the tele-operated flight, the visual field of on-board camera is moving continuously and some movements of the capture device should not be eliminated but softly compensated, generating a stable video instead of a static scene. The intentional movements of the camera must be estimated and removed from the cumulative motion parameters, obtaining a high frequency signal. We use this signal for simultaneously compensating vibrations and keeping intentional motion. The top plot in the Figure 3 shows the accumulative motion parameter (blue) and the motion intention (green). The difference between the signals (down plot in the the Figure 3) is utilized for warping the whole sequence.

There are several video stabilization algorithms, as mentioned in the Section 1, that use smoothing methods for the motion intention estimation. In a recent paper [1], a novel proposal for motion intention estimation has been introduced based on a second-order Butterworth filter [6]. This technique allows to compensate high frequency signals of the cumulative motion parameters without decreasing video quality nor generating phantom movements.

Despite of the significant advances presented by the algorithm as a real-time video stabilizer, the use of any filter always generate a delay in the output, and the second-order Butterworth filter is not the exception. In order to avoid the use of a motion smoothing technique and add an undesired delay in the video stabilization process, we propose in this work to consider the mathematical model that relates control actions and motion parameters. This model can be obtained off-line through experimentation with the MAV and its use in the real-time video stabilization architecture is straightforward.

Table 1 Intentional and undesired movements of parameters due to control actions

Control/Parameter	Angle	Scale	X	Y
Roll	undesired	undesired	INTENTIONAL	undesired
Pitch	undesired	INTENTIONAL	undesired	undesired
Yaw	undesired	undesired	INTENTIONAL	undesired
Altitude	undesired	undesired	undesired	INTENTIONAL

3.1 Model estimation of the MAV

The platform used in the experimentation is the AR.Drone 1.0, a low-cost MAV built by the French company Parrot. It has been selected for multiple reasons: low cost, low energy consumption, safe flying, and vehicle size. The AR.Drone can be controlled with hand-held devices as smartphones or tablets with operative system iOS or Android. Additionally, Parrot has opened the SDK (Software Develop Kit) for operating systems Linux and Windows, so it can be controlled with a laptop/desktop computer. The control system of the drone allows to manipulate four different control action: pitch, roll, yaw, and altitude.

Data has been collected from the IMU (inertial measurement unit) of the AR.Drone for several control actions for carrying out the aerial robot modeling. Then, the direct model has been estimated considering control actions as input, and MAV's position and velocities as outputs[1]. Finally, our interest is the estimation of the model relating the control action to the motion parameters in the image, based on the model in [4].

3.2 Hypothesis in the model

In order to solve some modeling concerns, two hypotheses based on the data have been considered [4]:

- The models for each motion parameter between frames are decoupled, and defined by the following relations: scale depends on pitch control, translation in the y-axis depends on altitude control, and translation in the x-axis depends on roll and yaw control.
- The relation between control action and motion parameters is a static nonlinear model combined with a linear model.

In the first hypothesis, the motion parameter angle is not considered because we are estimating the intentional motion. In the same way, there are movements in the y-axis that depends on pitch control, and in the x-axis depends on the roll control. But, in both cases, the movements are undesired. The table 1 shows the intentionality of movements of parameters depending on each control action.

[1] The complete experimentation and associated results can be checked in [4]

Considering the second hypothesis, the modeling process has been separated into two parts:

- Static nonlinear model estimation
- Dynamic linear model estimation

3.3 Neural network based static nonlinear model estimation

The nonlinearity is due to a saturation effect in the angle control system. For angles higher than the saturation limit, the acceleration is constant. One of the configuration parameters of the AR.Drone is the maximum angle for each rotation. For this reason, it is important to explain that the nonlinear model is necessary for application where the acceleration of the action control is not constant.

In [4], the nonlinear part of the model is estimated as an static system, using a fifth degree polynomial that relates the action control with the motion parameters.

$$P(q) = aq^5 + bq^4 + cq^3 + dq^2 + eq + f \tag{6}$$

For improving the input-output model of the MAV, we are using a feed-forward neural network that consists of one hidden layer with five neurons.

The neural network is trained a thousand times reducing the root-mean-square error (RMSE). We have obtained a RMSE = 0.0251 with the neural network, lower than using the approximating polynomial (RMSE = 0.2072).

In Figure 4, the polynomial and the approximate stationary values are plotted and compared.

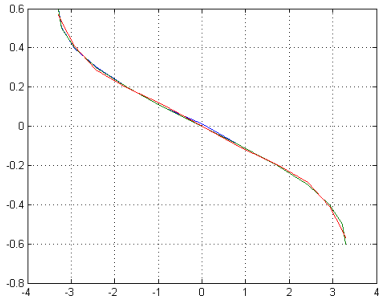

Fig. 4 Static nonlinear model. Real data (Green), Polinomy (Red) RMSE = 0.2072, Neural Network (Blue) RMSE = 0.0251

3.4 Identification of the dynamic linear model

Once the nonlinearity has been estimated, we can identify the model that relates the action control with the motion parameter for a constant acceleration. In the Figure 5, there is a graphic of control action data for roll (top) and the filtered motion parameter for x-translation (down).

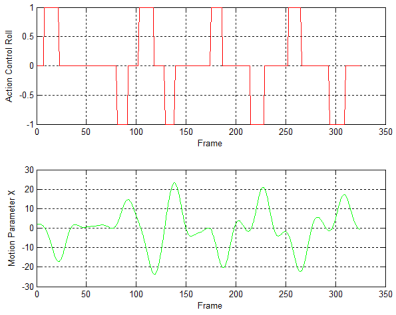

Fig. 5 Dynamics model estimation. Top: Input. Down: Output

Using a model identification tool for dynamics linear, we obtain the transfer function:

$$\mathrm{G(S)} = \frac{\mathrm{K_p}}{1 + \mathrm{T_p} * \mathrm{S}} * \exp^{(-\mathrm{T_d} * \mathrm{S})} \tag{7}$$

with a process gain K_p, process time constant T_p and a time delay T_d.

In Figure 6, the filtered motion parameter (Black) and the output of the estimated model (Blue) are shown. Compared to the real motion, the performance of the estimated model is better than the motion parameter estimated off-line with the filter. Our approach is focused on indoor flight application, but the model for outdoor flight is proposed as a future work.

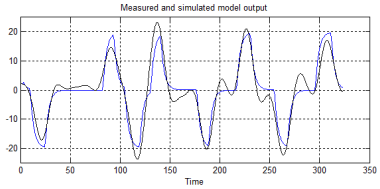

Fig. 6 Dynamics model estimation. Results

3.5 Kalman Filter

In literature we can find some algorithms that use Kalman filter for motion intention estimation. However, these algorithms employ Kalman filter as for feature point tracking previous to the motion parameters estimation.

Our approach uses the Kalman filter as a motion smoothing technique. The Kalman filter is applied after computing the motion parameters and is based on the mathematical model of the MAV. The model has two parts: A static nonlinear model and a Dynamic linear model.

Estimating the inverse of the static nonlinear model, and applying to the input, we obtain the input of the dynamic linear model. In the Kalman filter, we use the state space representation of the dynamic linear model. Therefore G(S) is representing as:

$$\begin{aligned} x_k &= Ax_{k-1} + Bu_{k-1} \\ z_k &= Cx_k + D \end{aligned}$$

Most of the video stabilization algorithms that use Kalman filter have not consider the input u_{k-1}. Our algorithm is based on the input for eliminating phantom movements.

One of advantages of the Kalman filter is that can be applied in real-time because depends only on the last frame. In Figure 7 we can see a comparison of the motion parameters, intentional motion based on low-pass filter (using 6 previous frames), and intentional motion based on Kalman filter.

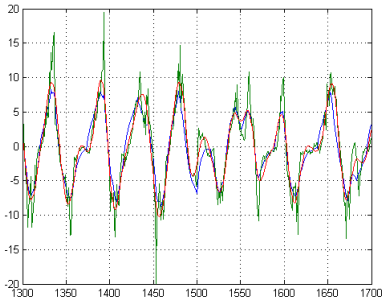

Fig. 7 Intentional motion. Motion parameters (Green), low-pass filter (Red), Kalman filter (Blue)

In Figure 8 we present our full video stabilization algorithm.

4 Results and discussion

We have carried out experiments using a MAV with an on-board camera in four different scenarios, all of them indoor. The employed MAV is the AR.Drone

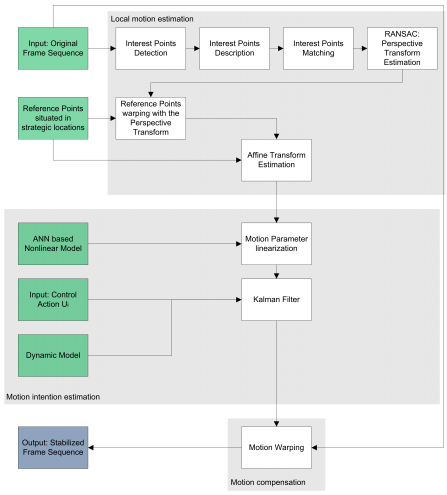

Fig. 8 Flow chart. Our proposal for video stabilization

2.0, described on the Section 4.1. Our video stabilization algorithm is implemented in a ground station, a laptop with a Processor Intel Core i7-2670QM 2.20GHz, Turbo Boost up to 3.1GHz and RAM 16.0 Gb. We use ROS (Robot Operative System) to comunicate the ground station with the MAV. The onboard camera is 720p (resolution = 1280x720) and records up 30 monocular frames per second (sample frequency = 25 Hz).

4.1 Metrics of evaluation

In the literature, the video stabilization algorithms use subjective (Mean Opinion Score [28]) and objective metrics (bounding boxes, referencing lines, and synthetic sequence [20]) to evaluate performance.

Focus on the quality of the final stable video, we use the Inter-frame Transformation Fidelity (ITF) [33], a widely used evaluation metric of effectiveness and performance,

$$\mathrm{ITF} = \frac{1}{\mathrm{N_f} - 1} \sum_{\mathrm{k}=1}^{\mathrm{N_f}-1} \mathrm{PSNR(k)} \tag{8}$$

where N_f is the number of video frames and

$$\mathrm{PSNR(k)} = 10 \log_{10} \frac{\mathrm{I}^2_{\mathrm{MAX}}}{\mathrm{MSE(k)}} \tag{9}$$

is the peak signal-to-noise ratio between two consecutive frames, with I_{MAX} being the maximum pixel intensity in the frame and MSE being the mean square error mentioned in subsection 2.2.

Focus on the motion realism of the final stable video, we use the root mean square error (RMSE) [12]. RMSE evaluates the difference between the estimated motion from the stabilized video and the real motion of the flight robot in the xy-plane. A lower RMSE means a estimated motion intention more similar to the real motion.

$$RMSE = \frac{1}{2F}\left(\sqrt{\sum_{j=0}^{F}(E_{x,j} - T_{x,j})^2} + \sqrt{\sum_{i=0}^{F}(E_{y,i} - T_{y,i})^2}\right) \tag{10}$$

where $E_{x,j}$, $E_{y,j}$ are the estimated, and $T_{x,j}$, $T_{y,j}$ are the observed motions in the axes for the jth frame. F denotes the number of frames in the sequence.

A tracker based on optical flow [35], and camera calibration for radial distortion [36] are used to computed the real motion from the video recorded with a zenith camera.

4.2 Comparison with other algorithms

Our approach has been compared with three algorithm from the literature:

- L1-Optimal off line method [16], applied in the YouTube Editor as a video stabilization option
- Our last algorithm based a low-pass filter as motion smoothing technique [1]
- Subspace video stabilization, utilized in the commercial software Adobe After Effects.

The performance evaluation of these video stabilization approaches, focus on ITF and RMSE, was carried out using each technique to stabilize different videos. We used the follow class of videos:

- Four videos without moving objects (30fps)
- Four videos without moving objects (10fps)
- Four videos with moving objects (30fps)

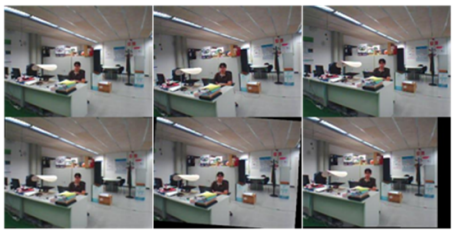

Fig. 9 Video 4. Top: Original video. Down: Stabilized video

Table 2 Evaluation Metrics: Videos without moving objects (30fps).

Algorithm	Evaluation Metric	Video 1	Video 2	Video 3	Video 4
Original	ITF(dB)	14.09	13.43	14.65	16.96
Our Approach	ITF(dB)	19.49	19.55	19.89	21.20
Our Approach	RMSE	0.021	0.018	0.024	0.015
L1-Optimal	ITF(dB)	19.62	19.57	20.16	20.24
L1-Optimal	RMSE	0.046	0.051	0.047	0.036
Motion Smoothing	ITF(dB)	19.48	19.52	19.89	21.12
Motion Smoothing	RMSE	0.028	0.023	0.029	0.017
Subspace	ITF(dB)	19.58	19.59	20.12	20.91
Subspace	RMSE	0.049	0.053	0.048	0.034

In Table 2, we present experimental result of four videos recorded in scenarios without moving objects, and stabilized with four different methods including our algorithm. In the Figure 9, we can see three frames from the original and stabilized video 1. Results show that our approach, applied in real-time, achieves ITF values as high as using other approach from the literature, applied off-line.

Additionally, the RMSE of our algorithm is lower because phantom movements are not generated, i.e., the motion of the video with our technique is more real without decreasing the video stability. [1] is also able to compensate the image without generating phantom movements, but required 6 last frames. On the other hand, our approach depends only on the last frame, and the computational time of the Table 2 means that there is no problem to apply our algorithm for stabilizing a 30fps video.

The effect of phantom movements is graphically compared between in Figure 10 to L1-Optimal and our approach. Our algorithm reduces the phantom movements as good as [1].

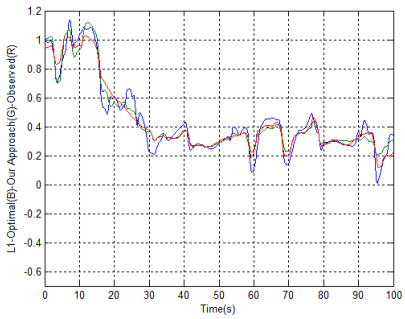

Fig. 10 Comparison of the Scale. L1-Optimal (Blue), Our Approach (Green), Observed (Red),

The approach presented in this paper is robust to the presence of moving objects, and low frequency videos.

Table 3 Evaluation Metrics: Videos without moving objects (10fps).

Algorithm	Evaluation Metric	Video 1	Video 2	Video 3	Video 4
Original	ITF(dB)	12.27	12.12	12.28	13.43
Our Approach	ITF(dB)	17.47	17.08	17.96	18.20
Our Approach	RMSE	0.022	0.020	0.023	0.019
L1-Optimal	ITF(dB)	17.64	17.42	17.15	18.22
L1-Optimal	RMSE	0.057	0.059	0.054	0.046
Motion Smoothing	ITF(dB)	17.40	17.07	17.92	18.14
Motion Smoothing	RMSE	0.023	0.020	0.025	0.019
Subspace	ITF(dB)	17.57	17.39	17.11	18.12
Subspace	RMSE	0.057	0.060	0.058	0.049

Fig. 11 Video 4. Top: Original video. Down: Stabilized video

Table 4 Evaluation Metrics: Videos with moving objects (30fps)

Algorithm	Evaluation Metric	Video 5	Video 6	Video 7	Video 8
Original	ITF(dB)	12.46	12.27	12.82	14.48
Our Approach	ITF(dB)	17.31	17.03	17.95	18.43
Our Approach	RMSE	0.026	0.022	0.025	0.018
L1-Optimal	ITF(dB)	17.49	17.44	17.94	18.47
L1-Optimal	RMSE	0.060	0.057	0.059	0.046
Motion Smoothing	ITF(dB)	17.21	16.96	17.90	18.34
Motion Smoothing	RMSE	0.024	0.022	0.027	0.018
Subspace	ITF(dB)	17.42	17.39	17.51	18.40
Subspace	RMSE	0.059	0.058	0.063	0.047

In the Table 3, we present experimental result of the four last videos recorded at 10fps.

The Table 4 corresponds to results obtained from videos recorded in scenarios with moving objects. In the Figure 11, we can see three frames from the original and stabilized video 5 with moving objects.

L1-Optimal and Subspace are two of the best video stabilization algorithms, and are applied off-line in two of the most famous video edition software. Our algorithm works in real-time but shows a robustness, to moving objects and low frequency (Table 3, 4), as good as L1-Optimal and Subspace.

5 Conclusions

In this paper, we have presented a novel video stabilization algorithm able to be applied in real-time, robust to scenes with moving objects and complex dynamic movements generating by on-board cameras of micro aerial vehicles. We can achieve a stable video sequence without generating phantom movements to compensate unintentional motion. In this way, our algorithm provides a reliable tool for tele-operated systems.

Our technique is based on the MAV model estimation including the control action, and the application of this model in the Kalman filter to smoothing motion without generating false movements. For post-processing applications the algorithm from the literature are sufficient, but our aim is the tele-operation and autonomous task of MAVs. In this sense, solving the issue of phantom movements could mean the difference that prevents an accident.

Our algorithm obtains a high performance for indoor flight. In the future, we plan to evaluate our video stabilization method in aggressive environments with turbulence and communication problems, as well as to apply it for increasing the performance of tracking algorithms.

Acknowledgements This work has been partially supported by the Spanish Ministry of Economy and Competitiveness, through the PATRICIA project (TIN 2012-38416-C03-01). The research fellow Wilbert G. Aguilar thanks the funding through a grant from the program "Convocatoria Abierta 2011" issued by the Secretary of Education, Science, Technology and Innovation SENESCYT of the Republic of Ecuador.

References

1. Aguilar, W.G., Angulo, C.: Real-time video stabilization without phantom movements for micro aerial vehicles. In: EURASIP Journal on Image and Video Processing. submitted (2014)
2. Aguilar, W.G., Angulo, C.: Optimization of robust video stabilization based on motion intention for micro aerial vehicles. In: Systems Signals and Devices (SSD), 2014 International Multi-Conference on. Spain (February 2014)
3. Aguilar, W.G., Angulo, C.: Estabilización robusta de vídeo basada en diferencia de nivel de gris. In: Proceedings of the 8th Congress of Science and Technology ESPE 2013. Ecuador (June 2013)
4. Aguilar, W.G., Angulo, C.: Control autónomo de cuadricopteros para seguimiento de trayectorias. In: Proceedings of the 9th Congress of Science and Technology ESPE 2014. submitted (May 2014)
5. Alahi, A., Ortiz, R., Vandergheynst, P.: Freak: Fast retina keypoint. In: Computer Vision and Pattern Recognition (CVPR), 2012 IEEE Conference on, pp. 510–517 (2012)
6. Bailey, S.W., Bodenheimer, B.: A comparison of motion capture data recorded from a vicon system and a microsoft kinect sensor. In: Proceedings of the ACM Symposium on Applied Perception, SAP '12, pp. 121–121. ACM, New York, NY, USA (2012)
7. Battiato, S., Gallo, G., Puglisi, G., Scellato, S.: SIFT features tracking for video stabilization. In: Image Analysis and Processing, 2007. ICIAP 2007. 14th International Conference on, pp. 825–830 (2007)
8. Bay, H., Tuytelaars, T., Gool, L.: SURF: Speeded up robust features. In: A. Leonardis, H. Bischof, A. Pinz (eds.) Computer Vision – ECCV 2006, *Lecture Notes in Computer Science*, vol. 3951, pp. 404–417. Springer Berlin Heidelberg, Berlin, Germany (2006)

9. Canny, J.: A computational approach to edge detection. Pattern Analysis and Machine Intelligence, IEEE Transactions on **PAMI-8**(6), 679–698 (1986)
10. Chang, H.C., Lai, S.H., Lu, K.R.: A robust and efficient video stabilization algorithm. In: Multimedia and Expo, 2004. ICME '04. 2004 IEEE International Conference on, vol. 1, pp. 29–32 Vol.1 (2004)
11. Derpanis, K.G.: Overview of the RANSAC algorithm. Tech. rep., Computer Science, York University (2010)
12. Fang, C.L., Tsai, T.H., Chang, C.H.: Video stabilization with local rotational motion model. In: Circuits and Systems (APCCAS), 2012 IEEE Asia Pacific Conference on, pp. 551–554 (2012)
13. Faugeras, O., Luong, Q.T., Papadopoulou, T.: The Geometry of Multiple Images: The Laws That Govern The Formation of Images of A Scene and Some of Their Applications. MIT Press, Cambridge, MA, USA (2001)
14. Fischler, M.A., Bolles, R.C.: Random sample consensus: a paradigm for model fitting with applications to image analysis and automated cartography. Commun. ACM **24**(6), 381–395 (1981). URL http://doi.acm.org/10.1145/358669.358692
15. Forsyth, D.A., Ponce, J.: Computer Vision: A Modern Approach. Prentice Hall Professional Technical Reference, New Jersey, USA (2002)
16. Grundmann, M., Kwatra, V., Essa, I.: Auto-directed video stabilization with robust l1 optimal camera paths. In: Computer Vision and Pattern Recognition (CVPR), 2011 IEEE Conference on, pp. 225–232 (2011)
17. Harris, C., Stephens, M.: A Combined Corner and Edge Detection. In: Proceedings of The Fourth Alvey Vision Conference, pp. 147–151 (1988). URL http://www.csse.uwa.edu.au/~pk/research/matlabfns/Spatial/Docs/ Harris/A_Combined_Corner_and_Edge_Detector.pdf
18. Hartley, R., Zisserman, A.: Multiple View Geometry in Computer Vision, 2 edn. Cambridge University Press, New York, NY, USA (2003)
19. Hsu, Y.F., Chou, C.C., Shih, M.Y.: Moving camera video stabilization using homography consistency. In: Image Processing (ICIP), 2012 19th IEEE International Conference on, pp. 2761–2764 (2012)
20. Kang, S.J., Wang, T.S., Kim, D.H., Morales, A., Ko, S.J.: Video stabilization based on motion segmentation. In: Consumer Electronics (ICCE), 2012 IEEE International Conference on, pp. 416–417 (2012)
21. Kendoul, F.: Survey of advances in guidance, navigation, and control of unmanned rotorcraft systems. Journal of Field Robotics **29**(2), 315–378 (2012)
22. Lee, K.Y., Chuang, Y.Y., Chen, B.Y., Ouhyoung, M.: Video stabilization using robust feature trajectories. In: Computer Vision, 2009 IEEE 12th International Conference on, pp. 1397–1404 (2009)
23. Leutenegger, S., Chli, M., Siegwart, R.: Brisk: Binary robust invariant scalable keypoints. In: Computer Vision (ICCV), 2011 IEEE International Conference on, pp. 2548–2555 (2011)
24. Lowe, D.: Object recognition from local scale-invariant features. In: Computer Vision, 1999. The Proceedings of the Seventh IEEE International Conference on, vol. 2, pp. 1150–1157 vol.2 (1999)
25. Luo, J., Oubong, G.: A Comparison of SIFT, PCA-SIFT and SURF. International Journal of Image Processing (IJIP) **3**(4), 143–152 (2009). URL http://www.cscjournals.org/csc/manuscript/Journals/IJIP/volume3/ Issue4/IJIP-51.pdf
26. Matsushita, Y., Ofek, E., Ge, W., Tang, X., Shum, H.Y.: Full-frame video stabilization with motion inpainting. Pattern Analysis and Machine Intelligence, IEEE Transactions on **28**(7), 1150–1163 (2006)
27. Mikolajczyk, K., Schmid, C.: Scale & affine invariant interest point detectors. International Journal of Computer Vision **60**(1), 63–86 (2004). URL http://dx.doi.org/10.1023/B%3AVISI.0000027790.02288.f2
28. Niskanen, M., Silven, O., Tico, M.: Video stabilization performance assessment. In: Multimedia and Expo, 2006 IEEE International Conference on, pp. 405–408 (2006)
29. Rublee, E., Rabaud, V., Konolige, K., Bradski, G.: Orb: An efficient alternative to sift or surf. In: Computer Vision (ICCV), 2011 IEEE International Conference on, pp. 2564–2571 (2011)

30. Song, C., Zhao, H., Jing, W., Zhu, H.: Robust video stabilization based on particle filtering with weighted feature points. Consumer Electronics, IEEE Transactions on **58**(2), 570–577 (2012)
31. Tordoff, B., Murray, D.: Guided sampling and consensus for motion estimation. In: A. Heyden, G. Sparr, M. Nielsen, P. Johansen (eds.) Computer Vision — ECCV 2002, *Lecture Notes in Computer Science*, vol. 2350, pp. 82–96. Springer Berlin Heidelberg, Berlin, Germany (2002)
32. Wang, C., Kim, J.H., Byun, K.Y., Ni, J., Ko, S.J.: Robust digital image stabilization using the kalman filter. Consumer Electronics, IEEE Transactions on **55**(1), 6–14 (2009)
33. Xu, J., Chang, H.W., Yang, S., Wang, M.: Fast feature-based video stabilization without accumulative global motion estimation. Consumer Electronics, IEEE Transactions on **58**(3), 993–999 (2012)
34. Yang, J., Schonfeld, D., Mohamed, M.: Robust video stabilization based on particle filter tracking of projected camera motion. Circuits and Systems for Video Technology, IEEE Transactions on **19**(7), 945–954 (2009)
35. Yilmaz, A., Javed, O., Shah, M.: Object tracking: A survey. ACM Comput. Surv. **38**(4) (2006)
36. Zhang, Z.: A flexible new technique for camera calibration. Pattern Analysis and Machine Intelligence, IEEE Transactions on **22**(11), 1330–1334 (2000)