Robust and Real-Time Image Stabilization and Rectification

Dan Koppel[1], Yuan-Fang Wang[1], and Hua Lee[2]
[1]Department of Computer Science
[2]Department of Electrical and Computer Engineering
University of California, Santa Barbara, CA 93106

Abstract

This paper presents a unified framework for achieving robust and real-time image stabilization and rectification. While compensating for a small amount of image jitter due to platform vibration and hand tremble is not a very difficult task, canceling a large amount of image jitter, due to significant, long-range, and purposeful camera motion (such as panning, zooming, and rotation), is much more challenging. Our framework selectively compensates for unwanted camera motion to maintain a stable view of the scene. The rectified display has the same information content, but is shown in a much more operator-friendly way. Our contribution is three-fold: (1) proposing a unified image rectification algorithm to cancel large and purposeful image motion to achieve a stable display that is applicable for both far-field and near-field image conditions, (2) improving the robustness and real-time performance of these algorithms with extensive validation on real images, and (3) illustrating the potential of these algorithms by applying them to real-world problems in diverse application domains.

1. Introduction

Image stabilization and rectification algorithms are of many practical uses. They are often needed to cancel unwanted image jitter due to the operator/platform vibration or flutter to arrive at a steady display. For examples, video shot by a hand-held camcorder often exhibits noticeable image jitter due to the operator's hand tremble, and so is that shot by the camera mounted on a mobile platform that is subject to random mechanical vibration and ground disturbance.

Canceling a small amount of image jitter due to platform vibration and hand tremble is not a very difficult task. For instance, many modern camcorders come with a "steady-shot" feature that uses simple image correlation (often implemented in hardware) to cancel out image jitter. However, canceling a *large* amount of image jitter, due to *significant, long-range,* and *purposeful* camera motion (such as panning, zooming, and rotation), is a much difficult task. Image jitter in this case cannot be considered random in nature and an image stabilization algorithm needs to employ a much more sophisticated formulation to maintain a steady view.

While the terms "*significant, long-range,* and *purposeful*" may imply that we should not cancel this motion, it should be remembered that in many real-world imaging systems there may be multiple objectives with conflicting solutions. By this we mean that while significant and purposeful camera manipulation is needed to explore new perspectives and acquire novel views, such manipulation often times causes difficulty for the human operator in image interpretation. Hence, an image stabilization algorithm should be designed to allow significant freedom in image *acquisition* while alleviating difficulty in image *interpretation*.

We mention here two practical problems in diverse application domains that can make use of such an image stabilization and rectification algorithm. The first application is in rectifying the video display in video-endoscopy. Endoscopes procedures are minimally invasive surgical procedures where several small incisions are made on the patient to accommodate surgical instruments such as scalpels, scissors, staple guns, and an endoscope. The scope acquires images of the bodily cavity that are displayed in real time on a monitor to provide the visual feedback to the surgeon to perform surgery.

In order to view the anatomy in a highly constrictive body cavity (e.g., nasal passage in rhinoscopy and inner ear cavity in otoscopy) and subject to the entry point constraint, the surgeon often manipulates the scope with large panning and rotation motion to eliminate blind spots. The views acquired can be highly non-intuitive, e.g., the anatomy can appear with large perspective distortion, sideways, or even upside down. Hence, while this type of manipulation is necessary to reveal anatomical details, it does cause significant difficulty in image interpretation.

The second application is in rectifying the video display in an unmanned aerial vehicle (UAV). Under the control of a ground operator, an UAV may purposefully pitch, roll, and rotate to maneuver into certain positions or to evade ground fire. Executing such maneuvers severely alters the capture angle of the on-board camera. Again, the banking action is purposefully aiding in the flight but hindering the intuitive nature of the viewed video.

As should become clear from the preceding discussion, in both applications, the operator manipulates the camera using large panning, zooming, and rotating actions to obtain better views of the subjects (i.e., organs and ground vehicles). The views thus displayed can be highly non-intuitive, may have large perspective or other types of distortion, and may even be upside down. The freedom in such manipulation is absolutely necessary and should not be restricted solely for easing the difficulty in image interpretation. Instead, the goal in designing image rectification algorithms for such applications should be to maintain some consistency and uniformity in the display, while allowing the operator to survey the scene as before.

One important point to note is that traditional jitter cancellation algorithms are not applicable in the above surveyed scenarios where camera motion is large and purposeful. This is because traditional algorithms are designed based on the assumption that undesirable noise (e.g., hand tremble and platform vibration) causes image jitter. In the absence of the noise, images should be stationary (as in a hand-held camcorder) or stably moving (as in the camera on a mobile platform). Hence, these algorithms do not distinguish between different degrees of freedom in camera motion (such as panning and rotation), nor are they aware of the fact that some are desirable (such as panning) and others are not (such as body rotation). An intelligent image stabilization and rectification algorithm must (1) recovers and distinguishes multiple degrees of freedom in camera movement and (2) selectively compensates for some but not others.

In more details, given a video sequence taken with significant, long-range, and purposeful camera motion, our algorithm deduces the amount of unwanted camera movement (e.g., body rotation relative to a specified reference frame). The images are then rectified and displayed in such a way as if that particular degree of freedom never changed (other degrees of freedom, such as panning and translation, are preserved). The rectification algorithm can be decomposed into three stages: (1) tracking 2D features within an image sequence, (2) relating 2D feature movement to 3D camera motion (and hence, recovering the unwanted rotation angle), and (3) image rectification.

Finally, this paper presents our continuing research into image stabilization and rectification. Our previous research in image rectification was in the domain of medicine and was concerned with maintaining a constant “head-up” view in closed encoscopic surgery to alleviate the “dis-orientation” problem in image visualization.

In this paper, we extended our rectification framework to the domain of autonomous aerial vehicle navigation. The image rectification problem in UAV differs significantly from that in endoscopy due to their diametric imaging conditions. While endoscopy deals with “near-field” visualization of highly-curved surfaces with large perspective distortion, UAV application is concerned with “far-field” visualization of mostly planar surfaces with approximately parallel or orthographical projection. However, this does not imply that UAV application is simpler. On the contrary, traditional mathematic frameworks for depth inference and recovery, such as the 8-point algorithm, produce numerically unstable results for planar surfaces. In fact, it was shown that a numerically degenerate configuration results when the 3D points lie in a plane. Hence, while the philosophy of recovering 3D structure and motion from video to properly compensate for the unwanted image motion is the same for both image stabilization problems in endoscopy and UAV, the mathematic frameworks are quite different.

Our contribution is three-fold: (1) proposing a unified image rectification algorithm to cancel large and purposeful image motion to achieve stable display that is applicable for both far-field and near-field image conditions, (2) improving the robustness and real-time performance of these algorithms with extensive validation on real images, and (3) illustrating the potential of these algorithms by applying them to real-world problems in diverse application domains.

2. Related Work

Theoretically, we formulate our image stabilization and rectification framework as one that infers the camera’s degrees of freedom in movement and then selectively compensates for the unwanted degrees of freedom for image rectification. Motion-parameter estimation is a topic that has been researched extensively in computer vision. Many techniques exist that recover 2D and 3D motion based on a variety of information sources and mathematic formulations (e.g., color correlation, optical flow, flexible shape templates, and discrete point correspondence, to name a few).

As our goal is to recover 3D camera motion parameters, we adopt the popular shape-from-motion formulation based on point correspondences. These types of techniques track discrete points in successive frames and use rigidity constraints on the 3D point configuration for inferring depth and motion. Further discussion on these methods can be found in [2,3,4,5,9]. The popular 8-point algorithm requires a minimum of 8 point correspondences in a non-degenerate configuration for the solution. The algorithm is readily applicable for image rectification in video-endoscopy. However, for UAV application, the background is largely planar (or appears to be so because of the large ratio of the apparent depth to the variation in the depth of the 3D structure). This actually represents a degenerate configuration for the 8-point algorithm. Thus, another method is required for this situation.

When the set of points is coplanar, a similar algorithm [8], requiring four point correspondences, is available. This "four-point" algorithm recovers both motion and structure (i.e., the plane). Previously, this algorithm has been used in [6,7]. However, unlike our application, in [6,7] some a priori knowledge was used in the processing.

3. Mathematical Formulation

The formulations of motion-parameter recovery for the planar (UAV) and non-planar (video-endoscopy) cases are similar in spirit but differ in key ways. The non-planar case makes use of a linear algorithm that requires 8 point correspondences as a minimum. By contrast, the planar case needs only 4 point correspondences. Since the 8-point algorithm is well-known, only its result is presented here.

For a single point correspondence we have the following: Define $\mathbf{x} = (x, y, z)^T$ and $\mathbf{X} = (x/z, y/z, 1)^T$ as the 3D coordinates and image coordinates, respectively, of a point as seen in the camera's reference frame. Similarly the same quantities as seen in a second video frame is denoted by primed variables **x**' and **X**'. We have what is known as the epipolar constraint:

$\mathbf{X'^T}\,\mathbf{FX} = 0$ (a scalar equation)

where **F** is known as the fundamental matrix, and $\mathbf{F} = \mathbf{TR}$. **T** is the (anti-symmetric) matrix formed from the translation vector **t** defined from $\mathbf{x'} = \mathbf{Rx} + \mathbf{t}$, which also defines **R**, the rotation matrix.

For the coplanar case, a linear equation, relating the primed and unprimed coordinates, can also be obtained in the following manner: Let the plane on which the points reside be defined by $\hat{\mathbf{N}}^\mathbf{T}\mathbf{x} = d$, where "^" denotes a unit-vector. By absorbing d into $\hat{\mathbf{N}}^\mathbf{T}$, we obtain a normalized plane equation $\mathbf{N}^\mathbf{T}\mathbf{x} = 1$, where $\mathbf{N} = \hat{\mathbf{N}} / d$. Starting with the same motion equation $\mathbf{x'} = \mathbf{Rx} + \mathbf{t}$, we obtain:

$\mathbf{x'} = \mathbf{Rx} + \mathbf{tN^Tx} = (\mathbf{R} + \mathbf{tN^T})\mathbf{x} = \mathbf{Ax}$, and

$z'\mathbf{X'} = z\mathbf{AX}$

Using the basis vector $\mathbf{e}_3 = (0,0,1)^T$, we have the following identity:

$\mathbf{e_3^T X'} = 1$

Next, we multiply both sides by **X**':

$\mathbf{X'e_3^T X'} = \mathbf{X'}$

and use our previous result to get:

$(z/z')\mathbf{X'e_3^T AX} = (z/z')\mathbf{AX}$

This yields the following (vector) equation:

$(\mathbf{I} - \mathbf{X'e_3^T})\mathbf{AX} = 0$

Not that the third component of this vector equation is simply an identity that provides no useful constraint.

In the non-planar case, epipolar constraint gives rise to the fundamental matrix equation, and the solution of which (using at least 8 points) allows us to infer the motion parameters **T** and **R**. Similar to the epipolar constraint, the above derivation gives rise to a linear, homogeneous equation (or the fundamental matrix equation for the planar case), and the solution of which (using at least 4 points) provides an intermediate quantity (the fundamental matrix **A** for the planar case). The matrix **A** can then be factored into the parameters of interest. Because there are two equations per point correspondence, fewer (four) correspondences are needed to obtain a solution. The actual recovery of **R**, **t** and **N** from **A** is a bit involved, but a clear discussion of this is found in [8].

It is interesting to note that, while factoring the non-planar fundamental matrix into **R** and **T** is not a one-to-one mapping (since the number of degrees of freedom of the factors is less than **F**'s), for the coplanar algorithm it is in fact a one-to-one mapping. This is due to the extra three degrees of freedom provided by the recovery of the structure parameters. As a result, no extra constraint needs to be imposed in the factoring process, as is the case for the non-planar case.

One last point needs to be mentioned. For the coplanar method, two solutions are actually generated between two adjacent frames used in the computation. One of them represents the physical solution and the other one can be considered a spurious solution. While both are consistent with the two frames used to perform the computation, only one is consistent with the entire video sequence. While [8] suggests using structural consistency across three frames to resolve the ambiguity, we have found that the idea can be extrapolated to using an arbitrary number of frames. For our application, we use six frames with good results.

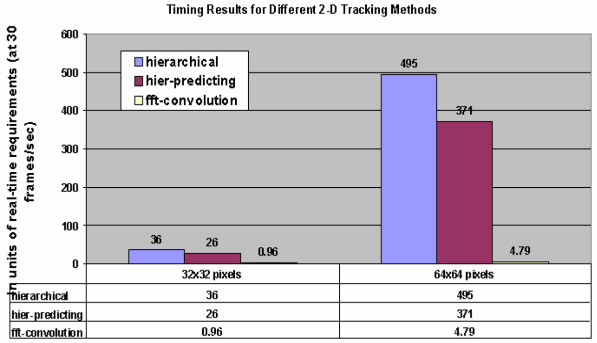
4. Experimental Results

We have used the ideas outlined above to design and implement a system for image stabilization and rectification that is applicable in a wide variety of settings. The adaptability of the framework to near-field (or non-planar) or far-field (or planar or nearly-planar) situations has been tested by running our algorithm on endoscopy video as well as that obtained from a UAV-mounted camera. In both cases, the unwanted degree of freedom is the camera's body rotation, i.e., a rotation about the camera's optical axis or the local z-axis. In video-endoscopy, such a rotation is commonly used in generating novel views and eliminating blind spots. In

UAV application, even the camera might be rigidly mounted under the belly, a UAV can execute banking and rolling actions that drastically change the orientation of the camera's coordinate frame relative to the ground coordinate frame.

Our framework allows the 3D camera motion (rotation and translation) to be inferred. The rotation about the optical axis can then be compensated for (by rotating the display the other way to counter the body rotation). The result is that the effects of purposeful, long-range camera body rotation are cancelled out successfully. This will maintain a consistent "head-up" direction in the visual display while allowing the same freedom in surveying the scene.

We present samples of video for both applications as well as timing and accuracy results below. Figure 3 shows video captured from a simulated endoscopy procedure running the non-planar formulation. The scene is the inside of a knee mockup. The camera is looking toward the upper leg and the end of the femur bone in the knee cavity. The camera executes large panning and rotation motion and is allowed to approach the cavity wall significantly, thus simulating real surgical situations. For Figure 3 (and all other similar display), video seuquences are shown from left to right and top to bottom. The display is grouped by showing the original, uncertified images on top and the corresponding rectified images immediately on the bottom. As can be seen that even with large panning, zooming and rotation (1st and 3rd rows), our algorithm is able to maintain the orientation of the femur bone in the rectified images (2nd and 4th rows).



	32x32 pixels	64x64 pixels
hierarchical	36	495
hier-predicting	26	371
fft-convolution	0.96	4.79

Figure 1: Timing statistics for feature tracking (used for getting point correspondences)

In Figure 4, a sequence of an actual endoscopy procedure on real tissue is shown. This represents a scenario where the surgeon rotates the camera to survey the cavity. As can be seen from the display, the original images (1st and 3rd rows) rotate almost 180 degrees (the position of the instrument rotates from the top right to bottom left). However, the rectified images (2nd and 4th row) show the same information but maintain the instrument at roughly the same location. In these scenarios, large variations in depth mean that good performance can be obtained by the non-planar formulation.

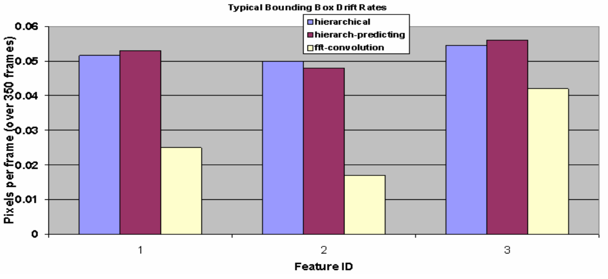


Figure 2: Accuracy statistics for feature tracking

In Figure 5 the results of running the application on real UAV video are displayed. Again, the original video is shown on top, with the corrected frames below. Since the ground scenery is flat with little deviation, the "far-field" or coplanar formulation is applied with good results.

Figure 1 and Figure 2 show the timing and accuracy results. We implemented a 2-D tracking algorithm using the FFT-convolution method. We compare the new FFT method with the traditional methods using hierarchical coarse-to-fine search in the spatial domain by minimizing the sum-of-square (SSD) error of intensity profiles. We implemented two versions of the hierarchical search methods: For regular hierarchical, we initially center the search window at the location predicted by the coarser solution. For hierarchical with prediction, we consider the estimated velocity of the bounding box and center the window where we think the new bounding box will be located. This allows us to employ a smaller initial search window size and thus gain in processing speed. In Figure 1 and Figure 2, the timing and accuracy data are presented in groups of three bars where the left bars represent the hierarchical method, the middle bars the hierarchical with prediction method, and the right bars the FFT convolution method

Figure 2 shows the accuracy results over three typical video sequences. The average drift per frame in tracked feature position over time is shown. Figure 1 shows the speed results. As processing speed is not a function of image contents, the comparison is in terms of the feature size. As can be seen from these, our method achieves better accuracy over traditional methods, and is significantly faster. For feature size of 64 by 64, our method is over one hundred times faster than the hierarchical method. This improvement is achieved without any special hardware (e.g., DSP) acceleration.

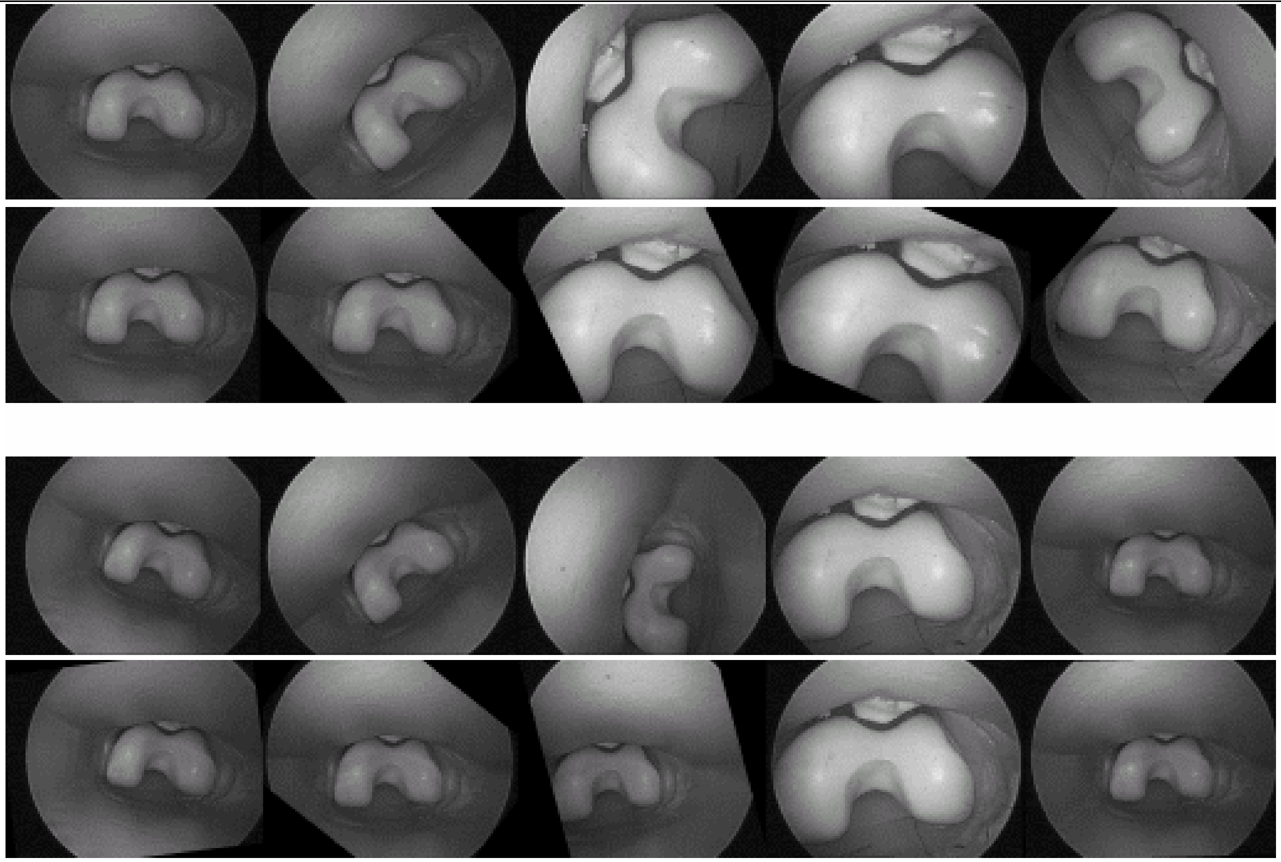


Figure 3: Endoscopic video using a knee mockup (top is original, bottom is compensated)

5. Concluding Remarks

We have demonstrated that the concepts of image stabilization and rectification are valid for situations involving more than just camera tremor and that viable stabilization methods can be applied to ease image interpretation. Our framework selectively compensates for unwanted camera motion to maintain a stable view of the scene. The rectified display has the same information content, but is shown in a much more operator-friendly way. A general framework that allows variation in scene geometry has been presented. By leveraging algorithms of complementary domains within the same overall methodology, real-world applications as diverse as endoscopy and UAV imaging have been shown to be amenable to the technique of motion understanding and compensation.

Acknowledgment

This research was supported in part by a contract from Karl-Storz Imaging, Inc and University of California Micro Program.

References

1. O.D. Faugeras and F. Lustman, "Motion and Structure from Motion in a Piecewise Planar Environment", International Journal of Pattern Recognition and Artificial Intelligence, 2(3):485-508, 1988.
2. O.D. Faugeras, F. Lustman, and G. Toscani, "Motion and Structure from Motion from Point and Line Matches", IEEE International Conference on Computer Vision, 1987.
3. R. I. Hartley, “In Defense of the Eight-Point Algorithm”, IEEE Transactions on Pattern Analysis and Machine INtelligence, Vol. 19, No. 6, June 1997
4. H.C. Longuet-Higgins, "A Computer Algorithm for Reconstructing a Scene from Two Projections," Nature, vol. 293, pp. 133-135, Sept 1981.
5. Q.-T. Luong and O.D. Faugeras, “The Fundamental Matrix: Theory, Algorithms, and Stability Analysis”, International Journal of Computer Vision, 1995, pp. 43-75.
6. O. Shakernia, Y. Ma, T.J. Koo, S. Sastry, "Landing an Unmanned Air Vehicle: Vision Based Motion Estimation and Nonlinear Control", Asian Journal of Control, pp. 128- 145, Vol. 1, No. 3, September 1999
7. O. Shakernia, R. Vidal, C.S. Sharp, Y. Ma, S. Sastry, "Multiple View Motion Estimation and Control for Landing an Unmanned Aerial Vehicle", IEEE International Conference on Robotics and Automation, Washington DC, May 2002
8. J. Weng, N. Ahuja, and T. S. Huang, "Motion and Structure from Point Correspondences: Planar Surfaces", IEEE Transactions on Signal Processing, vol. 39, no. 12, pp. 2691-2717, Dec. 1991.

9. G. Xu and Z. Zhang, “Epipolar Geometry in Stereo, Motion and Object Recognition”, Kluwer Academic Publishers, 1996

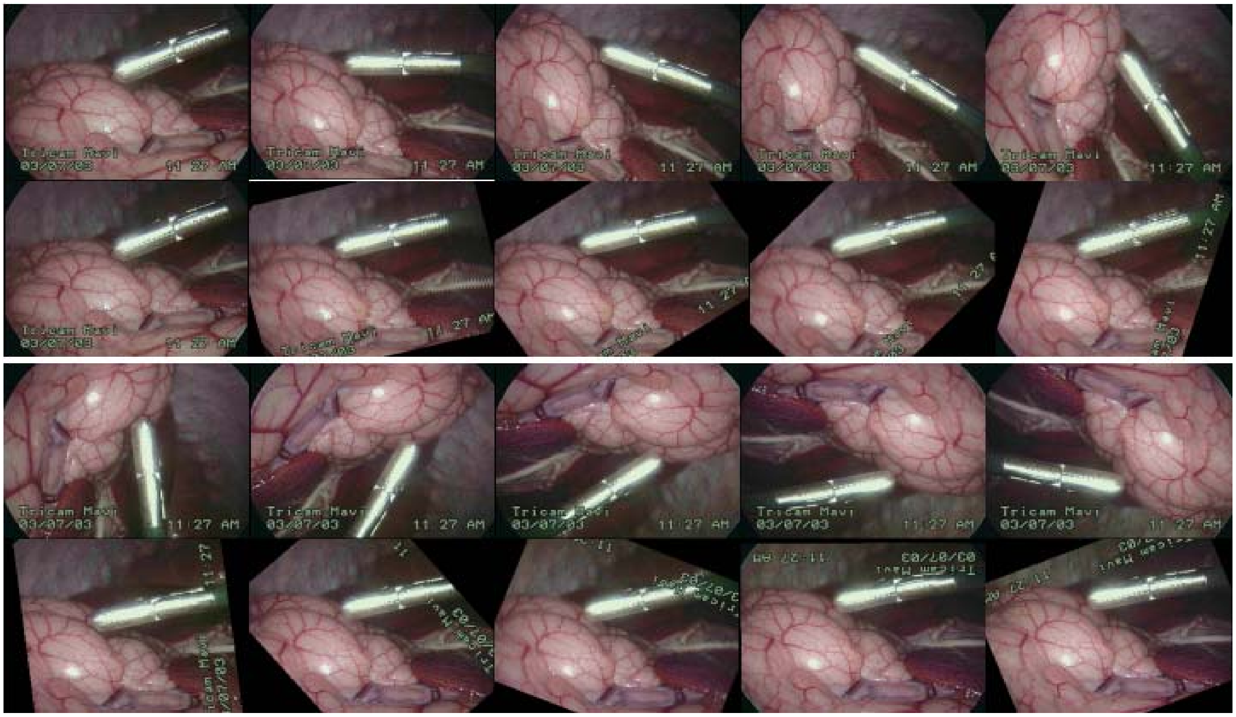


Figure 4: Actual surgical procedure (top is original, bottom is compensated)

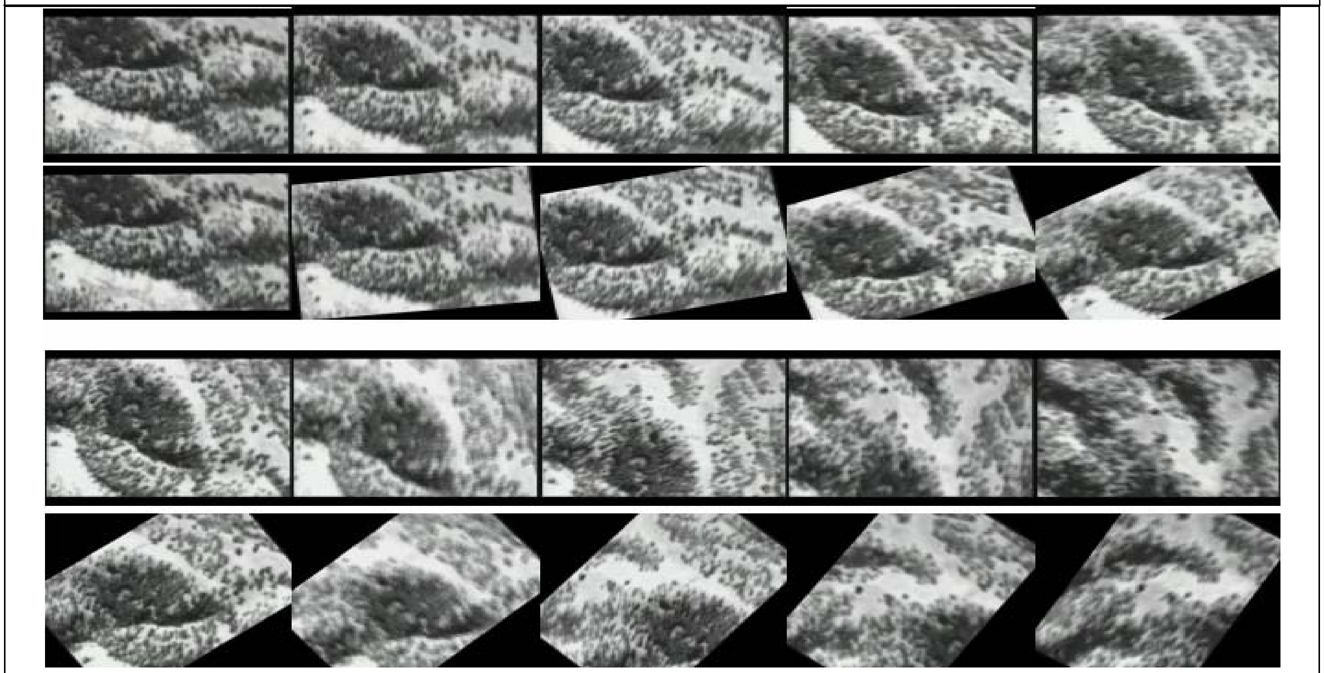


Figure 5: UAV video (please see *uploaded file*)