

PROGRESS ON FLIGHT VIDEO DATA ANALYSES FOR ASSESSMENT OF PVFR ROUTES AND SNI OPERATIONS FOR ROTORCRAFT

Jeffrey B. Mulligan
NASA Ames Research Center, MS 262-2, Moffett Field, CA 94035

Background: In the fall of 2003, a series of flight tests were performed in the Tullahoma, Tennessee area to assess the ability of non-instrument rated helicopter pilots to fly precision routes with the aid of a Global Positioning System (GPS) receiver. During each flight, recordings were made from four video cameras, two of which were attached to a goggle frame worn by the pilot. This paper describes the processing methodologies developed for these data.

INTRODUCTION

This research project seeks to determine the extent to which a cockpit GPS receiver can enable VFR pilots to adhere to precision routes, allowing Simultaneous Non-Interfering (SNI) operations in conjunction with fixed-wing traffic. To this end, a series of flight tests were flown in October 2003, in which pilots flew a route specified by 21 waypoints, some of which corresponded to visible landmarks, and others which were specified only by their latitude and longitude. Complete details of the route and flight test protocol can be found in Hickok & McConkey (2003).

Video data was collected during each flight using the Ames portable eye-tracking system, described in Darken et al. (2003). This system recorded four video streams onto a single 8mm videocassette. Additionally, two audio channels were recorded, one of which consisted of the cockpit audio, while the other channel was used to record video time code and GPS data. The remainder of this paper describes the processing applied to the video data and the current state of the analyses.

VIDEO PROCESSING

Before any processing could be done, the data first had to be transferred from the tapes to a computer. This was done at the University of Tennessee Space Institute (UTSI) campus, using a computer workstation equipped with an analog frame grabber (Matrox Meteor 1). Specially developed software allowed real-time digitization to a pair of dedicated hard disk drives with a capacity of approximately 30 minutes. As each recording had a duration of approximately 1 hour, each recording had to be digitized in two sections. After digitization, the "raw" images were converted to JPEG sequences, and moved to a conventional file system. The audio and GPS recordings were digitized along with the video. Following this procedure, the files

were transferred from UTSI to NASA Ames over the internet.

Camera Demultiplexing

Figure 1 shows a typical video field. Each field consisted of four quadrants, each of which corresponded to one of the four cameras. Camera demultiplexing refers to the process of taking a single movie consisting of the composite frames, and creating four movies corresponding to the individual camera streams. This was accomplished by a straightforward selection of the spatial subregions corresponding to each camera's image. The process was complicated, however, by the fact that the quad processor (which combined the four camera signals into a single signal) sampled the camera signals asynchronously; in other words, each frame put out by the quad processor and captured on tape consisted not of a complete frame from each camera, but was generally made up of portions of two consecutive camera frames. When the objects viewed by the camera were stationary, this could be ignored, but when the objects moved the result was a "tearing" of the frame (see figure 2). Because each of the four cameras had its own clock, the frame rates were all slightly different, and the tearing artifact occurred at a different position within each subimage.

This artifact was eliminated by first locating the occurrence of the tearing artifact, by looking for image discontinuities between pairs of adjacent scan lines. The vertical position within the frame containing the maximum discontinuity was determined, and plotted as a function of time. Because the artifact was produced by the difference in clock frequencies between the two devices, the discontinuities corresponding to the artifact fall on a function which is linear in time, resembling a "sawtooth." We fit a model to the observed data to reject outliers generated by vertical discontinuities in the image not related to the artifact.



Figure 1: Typical raw video field showing images from the four cameras; upper left: over-the-shoulder view from fixed camera; upper right: head-mounted eye camera; lower left: forward-looking head-mounted scene camera; lower right: view of pilot's head from camera fixed to instrument panel.



Figure 2: Recorded eye camera image showing "tearing" artifact resulting from rapid motion of the eye interacting with temporal resampling done by the quad processor.

An additional complication arises from the fact that the quad processor uses the interlaced format for its output signal. To reconstruct a camera field, we must "deinterlace" the recorded video from the quad processor. When the tearing artifact is present, it is only visible in one of the two fields output from the quad processor. Depending on whether it is the first or the second field, we must go forward or backward in

time to recover the missing parts of the frame.

Vibration Compensation

In viewing the recordings from the "face camera" mounted on the instrument panel, non-rigid distortions of the image were observed, which were presumed to result from vibration of the camera. These distortions were corrected as follows: first, a few prominent stationary features (parts of the vehicle visible to the sides of the pilot) were identified and tracked over the entire sequence. The motion of the camera in time was recovered from these displacements by remembering that the video lines are scanned sequentially in time; thus, the time at which a feature was imaged was proportional to the vertical position within the frame. After assigning the proper time to each observation, the motion was well-fit by a simple sinusoid. Using the inferred motion of the camera, the images were then warped to produce a relatively undistorted sequence.

Eye Camera Video

Our initial analysis of the eye camera images consisted of localization of the pupil (inner boundary of the iris) and the corneal reflection of the infrared LED's used to illuminate the eye. (In the day flights, the illumination provided by the LED's was generally much less than the ambient illumination, but the reflections of the LEDs themselves were still visible.)

For the night flights, we obtained images similar to those we routinely gather in the laboratory. The images from the day flights, however, posed some new challenges. Because the ambient daylight illumination was much stronger than that provided by the LED illuminators, these sequences are rife with illumination variations, as the vehicle changed its attitude relative to the sun. Another source of illumination variations was the vehicle rotor: because the clear windshield extended back over the pilot's head, a shadow was cast as the rotor passed overhead. Because this was a brief event, it only affected a few video scan lines, producing a dark band in the image (see Figure 3). The band appears vertical in figure 3 because the image has been rotated to put the eye in the proper orientation.



Figure 3: Day flight eye image showing dark band caused by rotor shadow, and partial occlusion of the eye by the upper lashes.

In addition to the illumination variations, there are a number of other features of the daylight eye images which have made robust tracking problematic. Because of the high ambient light levels, the pupil tends to be constricted, making it a smaller target. Similarly, the resting pose of the eyelids tends to be more closed, as if the pilots were squinting. This is problematic for two reasons: first, the eyelids hide more of the eye when they are partially closed; second, the upper eyelashes move in front of the pupil as the lid is closed, obscuring the features we are trying to detect.

Because of all these factors, our initial efforts to track the eye in the daylight videos have been only partially successful, with estimates obtained for only about 40% of the frames in the two flights processed. To overcome this shortcoming, we plan to redo the analysis, introducing a number of new techniques. In frames where the eye is visible, we will track the limbus (outer margin of the iris) in addition to the pupil. In addition to providing an additional feature, localization of the limbus will also provide a check on the pupil localization, because these two features should share a common center. (Refraction by the cornea makes them have slightly different centers for eccentric gaze directions, but this can be taken into account.)

We also plan to introduce methods to estimate gaze direction when the eye itself is hidden by the upper eyelid. We expect that the vertical component of gaze will be especially easy to recover, because the lid

moves with the eye, and therefore the vertical position of the lid is monotonically related to the vertical component of gaze. The horizontal component may be more difficult to extract, but we note that because of the fact that the cornea is a small dome rising out of the roughly spherical eyeball, its lateral motion causes a change in the shape of the covering eyelid, and in particular the form of the margin of the lid. Accuracy using this technique may suffer for two reasons: first, the measure itself is likely to be less sensitive than direct measurement of the pupil position; and second, we may not have calibration data for the extreme down-gaze positions for which the lid entirely hides the eye. However, these gaze directions do not correspond to those of most interest for this study (i.e., the GPS receiver and out-the-window landmarks), but rather correspond to the instruments at the bottom of the panel, and charts in the pilot's lap. Therefore, we believe that degraded accuracy for these gaze targets will be acceptable.

Face Camera Processing

We obtain an estimate of the pose of the pilot's head by analysis of the images from the fixed camera mounted on the instrument panel. Our procedure is a mix of automatic and manual procedures. First, a set of conspicuous features on the head are selected, such as the headset earphones, the microphone, etc. Next, a training set of 150 frames is selected. For each frame in the training set, an operator manually indicates the position of each feature using the mouse. At this point, we have 150 views of each feature, stored as small subimages. The various appearances of a feature can be efficiently described using a small number of parameters by applying a Principal Components Analysis (PCA) to the set of feature appearances, a technique first applied to entire face images by Turk and Pentland (1991).

We next obtain an approximate 3-dimensional configuration of the features from a pair of "mug-shot" views, that is by picking a view which is close to frontal and another which is nearly profile. The positions of the features in the frontal view give us the approximate x (side-to-side) and y (vertical) coordinates of the features, while the profile view provides approximate z (fore-and-aft) and y . We then refine the the estimates by alternately optimizing the structure and pose parameters over all 150 training images. This procedure stabilizes after 2 or 3 iterations, at which point we have estimates of both the 3-D structure of the features, and the pose of the head in each of the training images.

The next step is to derive the relationship between the pose and the appearance of each of the features (as described by the eigen-feature coefficients). For each

training frame, we have a set of pose parameters (angles) and a set of coefficients describing the appearance of the features. We derive an algebraic relation between the pose angles and each of the coefficients, which allows us to predict the appearance of each feature for an arbitrary pose – including poses which we may not have seen before.

We are now ready to describe the pose estimation process for an arbitrary new frame: we first make a guess about the pose, either recycling the final pose estimate from the previous frame, or assuming a frontal view for the first frame. Using this guess, we predict the corresponding appearance of each of the features. Using the expected feature appearances, we then search for each of the features in the image using cross-correlation. From the locations of the features, we estimate the pose. If the new estimate of the pose differs from our initial guess, we recompute the appearance of the features using the new pose estimate, and repeat the process until the estimate is stable (usually 2-3 iterations). Typical results are shown in figure 5.



Figure 5: Face camera image with line overlaid rendering of 3-D line segment model linking feature points.

Scene Camera Processing

The images gathered by the head-mounted scene camera provide a second source of information about the position and orientation of the head. *Structure-from-motion* refers to a technique by which both the camera pose and the 3-D locations of scene features can be computed from a series of images. While we ultimately hope to apply this technique, here we present a simpler method in which we approximate the camera motion by a pure rotation about the camera's optical nodal point. This simplification affords two advantages: first, we do not have to solve for the 3-D structure (or construct an accurate model of the cockpit interior); second, instead of identifying and tracking individual features, we can simply solve for the camera

pose parameters which provide the best overall registration of the image with the previous image or a template formed by mosaicing a set of images.

To register images related by large rotations, we must take into account the effect of the perspective projection performed by the camera-lens system. Because the camera-lens system projects the sphere of viewing directions onto a flat image plane, it is necessary to apply a complex non-rigid warp to bring two images into correspondence. We address this problem by adopting a cylindrical coordinate system to which we transform all the images.

To derive the transformation from the image sensor coordinates to the global coordinate system, we assume a generic pinhole camera model. But this is a poor approximation to our actual camera, which has a short focal length wide-angle lens which introduces considerable lens distortion. This distortion is evident in the appearance of the horizon, which usually appears curved in the raw video images. We apply an approximate correction for lens distortion by assuming a generic lens distortion model, and adjusting its single parameter to produce a linear horizon in a small number of representative frames.

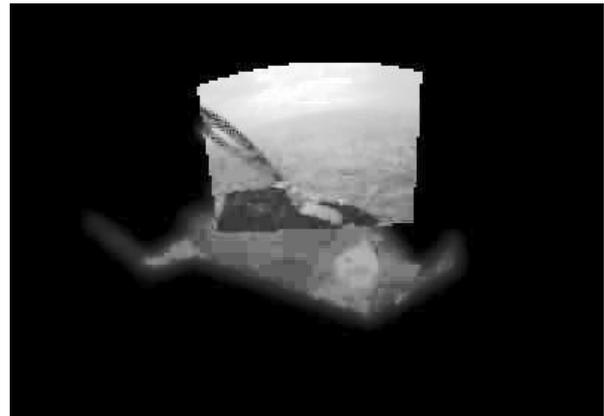


Figure 6: Composite image showing image from scene camera warped to cylindrical coordinate space, and overlaid on mosaic of instrument panel.

After correcting the raw video for lens distortion, we proceed to construct a mosaic of the cockpit as follows: we initialize the mosaic using an image filled by the instrument panel. Successive images are processed by first making an initial guess concerning the camera orientation (usually the orientation estimated for the previous frame). We then use the estimated orientation to warp the image to the common image space. The quality of the resulting registration is assessed by computing the normalized cross-correlation. The STEPIT optimization routine (Chandler, 1969) is used

to adjust the rotation parameters to optimize the fit.

Typical images from the scene camera contain both fixed features of the cockpit, and moving terrain features seen out the window. Because of the motion of the aircraft, these terrain features are not useful in determining the pose of the head, and we therefore wish to exclude them from the registration process. This is done by hand-construction of a mask which selects the portion of the mosaic image corresponding to the vehicle instrument panel and frame. Figure 6 shows the masked mosaic, with an input frame registered and overlaid.

SUMMARY

We have described a number of image processing procedures which have been applied to video data collected in the 2003 Tullahoma data collection flights. Our most reliable data has been obtained from the face-camera-based head pose estimation, with estimates obtained for approximately 85% of all frames, while the least reliable has been the day flight eye camera measurements, with estimates obtained for only 40% of all frames. We hope to improve the reliability and accuracy of all measures in the coming year.

REFERENCES

Chandler, J. P. (1969). "Subroutine STEPIT – Finds local minima of a smooth function of several parameters," *Behavioral Science*, v. 14, pp. 81-82.

Darken, R. P., Sullivan, J. A., and Mulligan, J. B. (2003). "Progress on the simulator and eye-tracker for assessment of PVFR routes and SNI operations for rotorcraft," FAA FY03 program review, Human Factors Vertical Flight.

Hickok, S. M., and McConkey, E. D. (2003). "Flight test plan to assess of PVFR routes and SNI operations for rotorcraft," FAA FY03 program review, Human Factors Vertical Flight.

Turk, M., and Pentland, A. (1991). "Eigenfaces for recognition," *Journal of Cognitive Neuroscience*, v. 3(1), pp. 71-86.