Proceedings of the 28th IEEE
EMBS Annual International Conference
New York City, USA, Aug 30-Sept 3, 2006

Compensating for Camera Translation in Video Eye Movement Recordings by Tracking a Landmark Selected Automatically by a Genetic Algorithm

Faisal Karmali and Mark Shelhamer, *Member, IEEE*

***Abstract*—We develop a method for accurately estimating the motion of a camera relative to a highly deformable surface, specifically the movement of a camera relative to the eye. A small rectangular landmark is selected and tracked throughout a set of video frames as a measure of vertical camera translation. The specific goal is to present a process based on a genetic algorithm that selects a suitable landmark. We find that *co-correlation*, a statistic relating the time series of a large population of landmarks, is a robust predictor of the accuracy of the landmarks. This statistic is used to iteratively select the best landmark from the population. At each iteration new landmarks are created that inherit properties of the previous population of landmarks. We show that the algorithm can select a landmark that will estimate camera translation with an accuracy of 1.8 pixels, which means that the direction the eye is looking can be determined with an accuracy of better than 0.6°.**

I. Introduction

Eye movement studies are important for understanding the properties of neural control systems and the arrangement of the brain. They provide an almost direct measure of vestibular (balance) function and provide insight into many general forms of neural processing, such as sensorimotor integration and prediction. Video eye monitoring, using high-speed cameras and appropriate image processing, has the potential to provide non-invasive measurement of eye position with high spatial accuracy and temporal resolution.

We recently conducted a study of eye movements in different gravity levels (g) aboard a NASA KC-135 aircraft, which flies a parabolic trajectory to provide alternating levels of reduced (~0 g) and enhanced (~1.8 g) g levels. Subjects noted that a point target viewed binocularly in darkness seemed to split into two vertically separated targets. This suggested a vertical misalignment of the eyes, which was confirmed with binocular video eye movement recordings using a video system consisting of two infrared cameras rigidly attached to a "headset" [1]. The recordings were analyzed to produce time series of eye position (determined using pupil position), which confirmed that the eyes vertically diverged as a function of g level from 1° to 3° [4]. One possible objection to this result is that the cameras could be moving relative to the head, which would create an artifactual difference in vertical eye position. We are convinced that this result is not an artifact because it is corroborated by the subjective reports, and through careful qualitative analysis of video. However, we would like to develop a quantitative approach to confirm our belief.

Attempts have been made to solve the problem of measuring camera motion relative to the eye using image processing of video eye recordings. The general approach is to find and track a landmark in the image that moves when the camera moves, but not during other events, such as eye or eyelid movement. In the corneal reflection technique, infrared LEDs fixed to the camera produce bright reflection on the cornea, whose location in the image is used to compute camera translation. In practice the aspherical shape of the cornea makes precise calibration necessary. A second technique is to mark the skin around the eye and assume that any translation of these landmarks is due only to movement of the cameras. However, this would require marking the skin before the experiment, and our data has already been gathered at great expense, and thus this method cannot be applied. A third technique is to detect and track movement of the medial canthus, the place where the upper and lower lids converge next to the nose. Unfortunately it is not visible in many of our images due to a small camera field, and thus cannot be tracked.

Another landmark that can be tracked is the upper eyelid. We previously developed an automatic image-processing technique to measure the positions of the eyelids, which are then used to make corrections to the pupil location so that eye position can be determined with less than 1° of error [5]. However, eyelid position is affected by blinks, and thus is valid for this purpose only when the difference between the two eyes is required, since the eyelids move by the same amount. We explored techniques where the en-block movement of each video frame was estimated using cross-correlations, but this technique has a high error, which is expected because eye and eyelid movement interfere with the cross-correlation. Computing optical flow [3] to estimate camera translation resulted in some success, but optical flow computation is difficult with the repetitive local textural features of the skin. None of the existing solutions are an accurate and practical means for camera movement estimation.

Our goal is to develop an automatic algorithm to estimate the amount of camera translation relative to a deformable physiological image. In this paper, we constrain the problem

Manuscript received April 3, 2006. This work was supported by NIH DC006090, NASA/NSBRI, NSERC.

Faisal Karmali is with the Johns Hopkins University School of Medicine, Baltimore, MD 21218 USA (phone: 410-218-7614; fax: 410-614-1746; e-mail: faisal@ jhu.edu).

Mark Shelhamer is with the Departments of Otolaryngology and Biomedical Engineering, the Johns Hopkins University School of Medicine (e-mail: mjs@dizzy.med.jhu.edu).

to vertical translations of a camera relative to the eye and surrounding tissue. We are constrained by the small field of view of the video images, since they were acquired at great expense and it is not possible to repeat the experiment (Fig. 1).

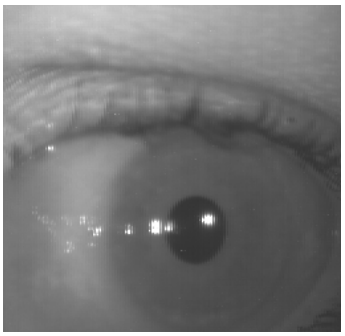

Fig. 1. An example of the field of view in video eye tracking

The general approach developed here is to select an arbitrary rectangular landmark from a reference frame and track that landmark in each frame of the video. Tracking can be performed using a cross-correlation [6]. Using a smaller landmark allows us to assume rigidity over a small portion of the video image. A large population of potential landmarks exists and can be generated. The question is how to select the appropriate landmark.

In this paper we develop a method that automatically selects an appropriate landmark using a genetic algorithm (Fig. 2). A genetic algorithm is an optimization procedure that searches the solution space using techniques inspired by evolutionary biology, such as inheritance and recombination. They are typically implemented by creating a population of candidate solutions that evolve to become better solutions. In each iteration, the fitness of the population is calculated, and a new population is created from the best members, which become the current population in the subsequent iteration [2]. Our implementation begins by creating a large population of potential landmarks. A metric, *co-correlation*, is presented which predicts tracking accuracy of a particular landmark, and is used as an objective function when searching the space of all possible landmarks. Optimization efficiency is improved by iteratively eliminating, early in processing, solutions that are not "*worthy;"* those which are unlikely to succeed because they have a low *co-correlation*. After the best landmarks are found, they undergo *recombination* to produce a new population of landmarks, to which the process is applied again. This iterative procedure successfully creates and selects landmarks which enable us to measure camera translation.

II. Methods

A. Landmark Tracking

A landmark is a small rectangular section of the video image. It is extracted from a reference video frame, and then identified in each video frame. The difference in the landmark's position in each frame relative to the reference frame is an estimate of camera translation. The landmark is found using a two-dimensional cross-correlation [6]. This produces a 2-D probability distribution, and the location of the maximum value in the distribution corresponds to the horizontal and vertical location (in pixels) of the landmark in the image. The cross-correlation is normalized by the number of overlapping pixels so that landmarks that overlap the edge of the image are not penalized. Edge enhancement to accentuate the difference between dark and light areas is applied to both the landmark and the video frame, using the linear combination of the original image and the image filtered with a first-order filter.

This algorithm computes only translation, and not tilt, of the camera. Because of the way the headset sits, any rotation will be about the center of the head, and will manifest mostly as a translation of the two cameras, in opposite directions. Empirically, headset tilts of 2° are typical, but if we assume a tilt of 5°, which is negligible in image processing, and a distance between the two eyes of 65 mm, and given our system's ratio of 4.5º/mm, a maximum artifact of $\tan^{-1}(5°)$x65 mm = 5.66 mm = 1.26° is introduced.

B. Automatic Evaluation of Landmark Accuracy

A measure termed *co-correlation* was developed as a way to estimate the tracking accuracy of a landmark. The process begins by creating a large population of landmarks (typically 100), either random rectangles or a grid of overlapping rectangles that span the image. Landmarks close to the pupil are omitted from analysis, since they are more strongly influenced by eye movements than camera motion. The pupil is detected using an intensity threshold, since it appears much darker than other objects in infrared.

Each landmark is found in each video frame as described above, producing a time series of vertical translation estimates for each landmark. The correlation coefficient of each time series with every other time series is calculated, to produce a correlation matrix that estimates the level of correlation in the movement of each landmark relative to each of the others. Finally the *co-correlation*, the sum of each row of the correlation matrix, is computed, which is a measure of how much mutual movement a landmark has relative to all other landmarks. The landmarks with the highest *co-correlation* are labeled *worthy*.

To improve computational efficiency, an iterative reduction approach is taken to compute *co-correlation*. Rather than apply each landmark to every video frame, a subset of video frames is used (for example, every 20^{th} frame), and an interim subpopulation of *worthy* landmarks is selected consisting of those landmarks with the highest *co-correlations*. In the next iteration, a larger subset of video frames is used, and *co-correlation* is recomputed for the smaller subpopulation of landmarks (Fig. 2).

C. Creating New Landmarks

After the best landmarks have been found, the *recombination* step creates new landmarks with similar properties, in an attempt to find new landmarks that further reduce error. Fig. 3 explains the rules used to create the new landmarks; in brief it works by combining the geometric properties of two landmarks to create two new landmarks.

Typically out of 100 landmarks, 10 *worthy* landmarks will be recombined. Each landmark will be recombined with each other, including itself, to create 100 new landmarks. The process of determining the best landmarks is then repeated with this new population.

D. Landmark Stability

Since we are only interested in estimating the motion of the imaging cameras, it is important that error introduced by other sources of motion is either eliminated or characterized. In this application, these are potential sources of changes in pixel intensity in the image:

1. The eyeball can rotate about three axes, which manifests mostly as a horizontal and vertical movement of the pupil and iris.
2. The camera can translate and rotate relative to the eye. Because of the way the camera is secured to the head, most of this movement manifests as vertical and horizontal translation of the image.
3. The eyelids move due to blinks and fatigue, which manifests as mostly vertical translation and stretching of skin above and below the eyes.
4. The skin around the eye can move due to external forces; however we believe this movement is small.
5. Movement of background light reflections can selectively illuminate one part of the image. This is not predicable, but is also not common.

Movement of the eyeball is ignored by detecting the pupil and ignoring landmarks near it. Lighting changes and mechanical deformation of the skin are believed to be small. However, deformation of the soft tissue around the skin due to blinks and fatigue may not negligible. To address this issue, landmark tracking accuracy during intentional soft tissue movement was studied.

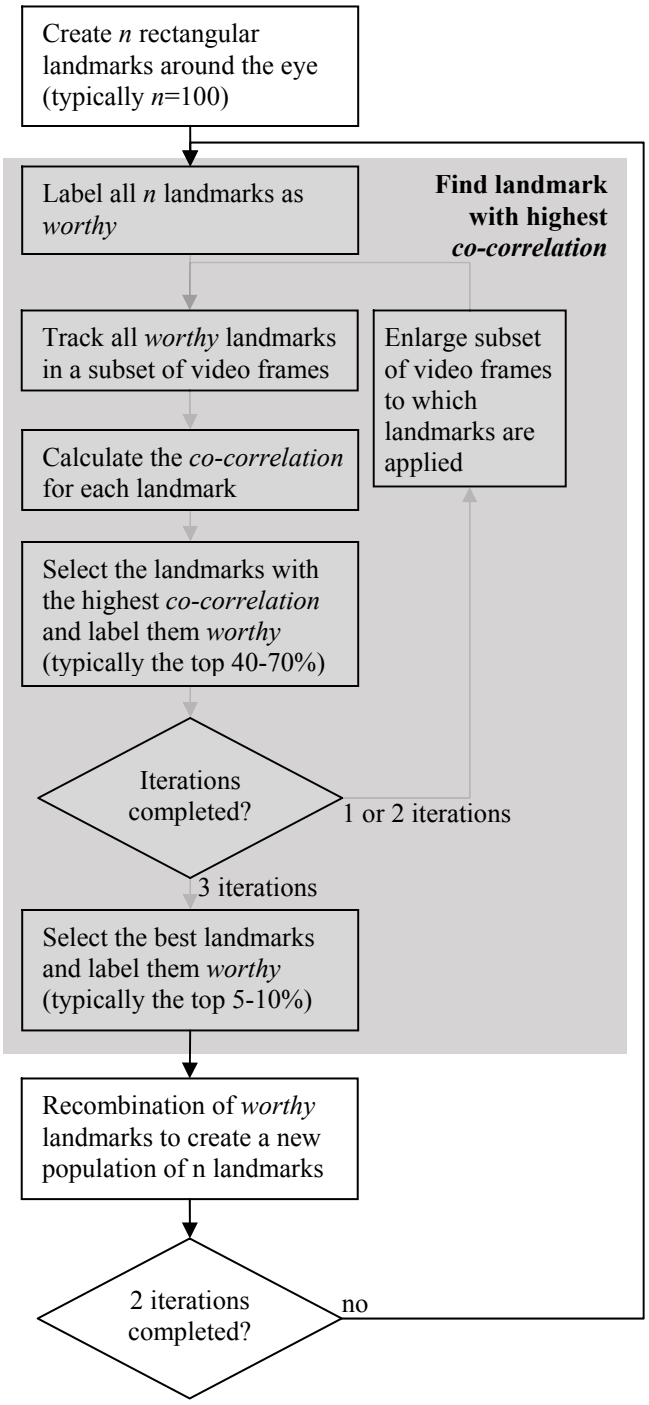

Fig. 2. Flowchart giving an overview of the algorithm used to create landmarks and determine the best landmark using *co-correlation*

E. Implementation and Evaluation

To evaluate accuracy, benchmark datasets were created by an experimenter who interactively located a physiological marker in each video frame, usually a crease in the skin far below the lower eyelid. In some cases benchmarks were created by placing an eye patch with an artificial pupil on one eye, while the experimenter moved the headset and the subject moved their eyes and eyelids. Accuracy is defined as the RMS difference between the benchmark and the time series resulting from landmark tracking. It is computed in pixels and converted to degrees to give the accuracy in determining the direction the eye is looking after pupil

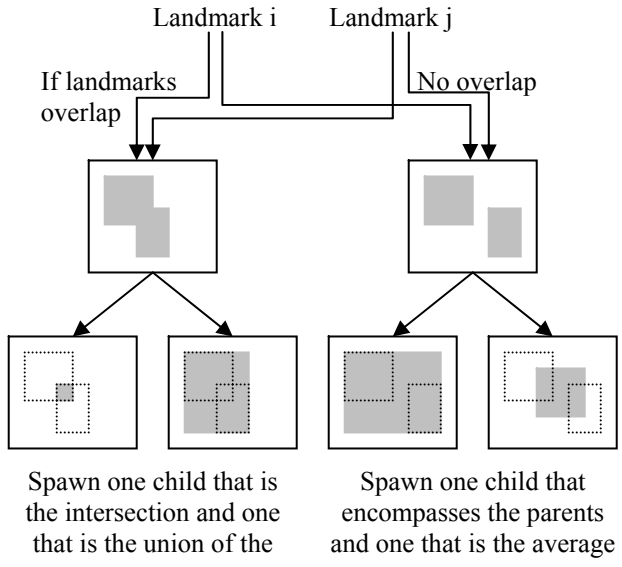

Fig. 3. Recombination of two landmarks to create two new landmarks. The way they are combined differs depending on whether they overlap

position is compensated for landmark position.

Co-correlation and accuracy of landmark tracking compared to a benchmark were computed for several hundred landmarks in four video recordings from four subjects. Recording were made at 50 Hz and consisted of left and right eye images, each 256x256 pixels with 256 levels of gray, and were imported into Matlab and analyzed using the Image Processing Toolbox.

III. Results

An important result is that *co-correlation* predicts how accurately a particular landmark estimates camera translation. Each dot in Fig. 4 represents one landmark, with a low mean error indicating the most accurate landmarks. The gray dots show a trend that as *co-correlation* increases, the mean error in tracking the landmark position decreases. This shows that *co-correlation* can be used as an objective function when searching the solution space. The gray plus signs indicate the *worthy* landmarks: those with the highest *co-correlation*, with accuracies ranging from 0.4° to 1.4°. These are very low; however, note that the landmark with the three highest *co-correlation* have an error larger than the best landmark. Thus, although the *worthy* landmarks have a much lower error than the rest of the landmarks, picking only the landmark with the highest *co-correlation* does not yield the best solution.

Another significant result is that *recombination* improves accuracy. The black dots in Fig. 4 are the results for the landmarks created by *recombination*. The black plus signs indicate *worthy* landmarks for the second iteration. The errors for these landmarks are between 0.4° and 0.6°. These errors are similar to the best landmarks in the first iteration, but the range is much smaller. In this case, any one of the *worthy* landmarks can be chosen as an accurate landmark. Thus, while in the first iteration the worst-case accuracy from amongst the *worthy* landmarks was 1.4°, in the second iteration it is 0.6°.

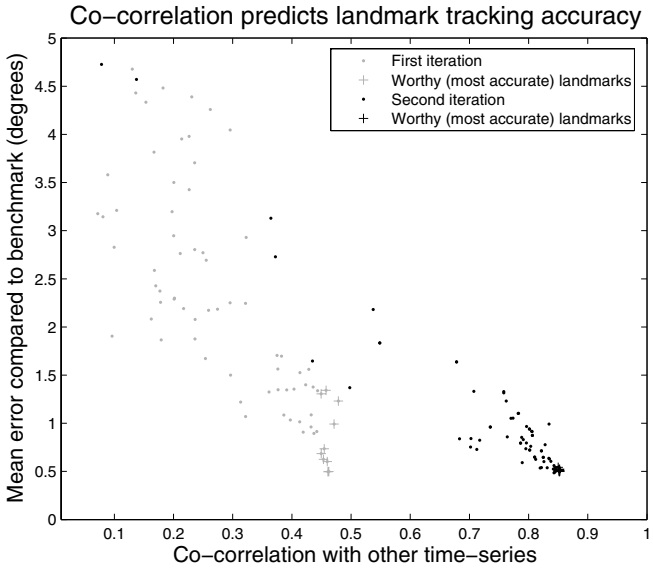

Fig. 4. *Co-correlation* predicts tracking accuracy, as determined by user-defined benchmark data. The benchmark data is never used by the image processing algorithms, so this is a very robust result. The landmarks used in the second iteration (black dots) are created by recombination of the *worthy* landmarks from the first iteration (gray plus signs). The selected landmarks (black plus signs) have errors less than 0.6°.

To see if deformation of the soft tissue around the skin affected the results, an artificial pupil on a patch over one eye was tracked as a head-fixed benchmark, while the subject moved their eyes and eyelids and the experimenter moved the headset. The accuracy for the *worthy* landmarks ranged from 0.6° to 1.2° (results not shown).

IV. Conclusions

The genetic algorithm presented here can accurately estimate vertical camera translation in eye movement video images. The *co-correlation* measure is a robust predictor of tracking accuracy for a given landmark, and can be used to select the best landmark from a population. With two iterations it is possible to find a landmark that will result in an accuracy of better than 0.6° in determining the direction the eye is looking.

Although this accuracy is sufficient for most eye movement studies, some studies, such as those of static eye alignment, require better accuracy. Better landmarks would be afforded by a camera lens with a larger field of view.

While this technique was developed specifically to compensate for unwanted camera movement relative to the head, it could also be expanded to track intentional movement of the head when the camera is fixed in space. This would allow tracking eye and head position. This is important for applications where free head movement without the impediment of a heavy head-mounted video system is desired, such as studies of eye movements in pilots.

Acknowledgment

We thank Ms. Tiffany L. Chen for assistance with software and creating benchmark datasets and Dr. Jerry Prince for advice.

References

[1] A. H. Clarke, J. Ditterich, K. Drüen, U. Schönfeld and C Steineke, "Using high frame rate CMOS sensors for three-dimensional eye tracking," *Behav Res Meth Inst Comp*, vol. 34, pp. 549-560, 2002.

[2] D. E. Goldberg, David E, *Genetic Algorithms in Search, Optimization and Machine Learning*, Boston, MA: Kluwer Academic Publishers, 1989.

[3] B. K. P. Horn and B. G. Schunck, "Determining Optical Flow," *Artificial Intelligence*, vol. 185 pp. 185-203, 1981.

[4] F. Karmali, S. Ramat, and M. Shelhamer "Vertical Skew due to Varying Gravitoinertial Forces: A Possible Consequence of Otolith Asymmetry," *Abstracts of Sixth Symposium on the role of the Vestibular Organs in Space Exploration*, Portland, OR, 2002.

[5] F. Karmali and M. Shelhamer, "Automatic detection of camera translation in eye video recordings using multiple methods," *Ann.N.Y.Acad.Sci.*, vol. 1039, pp. 470-476, Apr. 2005.

[6] J. C. Russ, *The Image Processing Handbook, Fourth Edition*. Boca Raton, FL: CRC Press, 2002.