1 Introduction

As the usage of handheld digital camcorders increase amongst the general population, the problem of undesired motions being introduced into captured video sequences will become ever more prevalent. Even the slightest movements on the part of the user will introduce these undesired motions, and unfortunately users will find it difficult to remain absolutely stable for any amount of time. These undesired motions detract from the intended subject of the video sequence, and may potentially ruin the entire video sequence. Undoubtedly, the user will want these types motions removed while preserving desired motions. Luckily, the camcorder’s digital format lends itself very nicely to a post-processing corrective step to remove such undesired motions.

We introduce a method for removing such unwanted motions by characterizing desired motions as smooth transitions with gradual starts and stops that occur over the span of some time period, and undesired motions as anything that detract from this ideal. Video sequences can then be corrected by "smoothing" the motion over some span of time. For "smoothing" to be performed, some representation of motion is needed and since planar homographies map points from one video frame to another, it essentially describes the transformation/motion between frames. Homographies are calculated for each sequential pair of frames, and by accumulating and manipulating each of these entries, this creates a set of ideal homographies representing only the desired motions. We then force each pair of frames to respect the corresponding ideal homography. The resulting frames are subsequently recombined to create a corrected video sequence containing only desired motions.

The problem of defining exactly how gradual the starts and stops, and how smooth the transitions should be to classify a motion as desired remains. Unfortunately, the ideal definitions of these parameters are entirely subjective and dependent on each unique user. We made the decision that any motion lasting less than 4 seconds is unwanted, and any motion that last for more than 4 seconds is desired. However, the method as described will be general enough for other possible definitions of what motions is desired or undesired, so long as the definitions can be represented as a set of ideal homographies.
2 Initial Motion Estimation

The first step in our algorithm estimates a homography between each pair of successive frames in the video sequence. A homography $H$ between frames $k$ and $k+1$ is defined as

$$ H^i = \begin{bmatrix} h_{i1}^1 & h_{i2}^1 & h_{i3}^1 \\ h_{i1}^2 & h_{i2}^2 & h_{i3}^2 \\ h_{i1}^3 & h_{i2}^3 & h_{i3}^3 \end{bmatrix} $$

and given frames $f^1, f^2, \ldots, f^k, \ldots, f^N$, $N$ defining the total number of frames in each sequence, Forstner features [1] are detected for frames $f^1, f^2, \ldots, f^{N-1}$. For each set of features in frame $f^k$, a set of point of correspondences is found in frame $f^{k+1}$ via normalized correlation. Hartley normalization is performed on each set of matching points, and a Direct Linear Transform aided by RANSAC [3] is used to calculate the homography, $H^k$, between each pair of frames, $f^k, f^{k+1}$.

This approach in calculating homographies implicitly makes a few assumptions on the captured scene structures. The scene must remain relatively static, as any objects moving independently of the camera will throw off the homography estimations. Motion blur will also reduce the accuracy of the point correspondences, which in turn will also reduce the accuracy of the homography estimations.

Once the homographies have been estimated, each entry $h_{i1}^k, \ldots, h_{i3}^k$ for a pair frames $f^k, f^{k+1}$, provides some information on what transformation the camera undertook between the two frames. To examine entries from different homographies together, we must first normalize each matrix to some type of common scaling factor and we can do this by multiplying each matrix by the $1/h_{33}^k$. Once the homography matrices have been normalized to some common scaling factor with respect to each other, we can examine the set of homography matrices as a whole by placing each entry from all the homography matrices into a single vector,

$$ v_{11} = [h_{11}^1, h_{21}^1, \ldots, h_{N-1}^{N-1}] $$
$$ v_{12} = [h_{12}^1, h_{22}^1, \ldots, h_{N-1}^{N-1}] $$
$$ \vdots $$
$$ v_{32} = [h_{13}^1, h_{23}^1, \ldots, h_{N-1}^{N-1}] $$

thus creating a set of eight vectors of the eight entries that represent the camera motion over the entire video sequence.

3 Motion Correction

Once the camera motion has been estimated, a decision must be made to determine exactly what camera motions we want to keep and what camera motions we want to correct for. We take a heuristic approach and assume that any camera motion lasting less than 4 seconds is unwanted motion. To remove these unwanted motions, each of the eight $v_{11}, v_{12}, \ldots, v_{32}$ are convolved with a Gaussian kernel with a window size equal to 2 times the frames per second rate of the motion sequence. We define the resulting convolved vectors as $\eta_{11}, \eta_{12}, \ldots, \eta_{32}$. The effect of the Gaussian filter on the set of vectors is similar to what would happen if a Gaussian filter is applied to a 2D image; high frequency noise is removed, while dominant image features are "smoothed" out over the entire image. The smoothing effect that a Gaussian filter provides is exactly the characteristic that we want our resulting video sequence to possess. After convolution, the set of convolved vectors are then recombined into an ideal set of homography matrices

$$ \mathbb{H}^k = \begin{bmatrix} \eta_{11}(k) & \eta_{12}(k) & \eta_{13}(k) \\ \eta_{21}(k) & \eta_{22}(k) & \eta_{23}(k) \\ \eta_{31}(k) & \eta_{32}(k) & \eta_{33}(k) \end{bmatrix} $$
that we would like each pair of frames, $f^k, f^{k+1}$ to respect.

After the ideal set of homography matrices has been determined, the existing frames of the video sequence must be transformed in such a way so that each pair of transformed frames respect the corresponding ideal homography. For this problem, we first start from the very first frame of the movie sequence and observe that the second frame can be transformed to respect the ideal homography matrix. Given an initial set of points, $pts^k = [(x^k_1, y^k_1), (x^k_2, y^k_2), (x^k_3, y^k_3), (x^k_4, y^k_4)]$, from the first frame, this initial set of points can be mapped both onto the second frame and the second ideal frame.

$$f^1 \rightarrow f^2 : pts^2 = H^1 \cdot pts^1$$
$$f^1 \rightarrow f'^2 : pts^2_{\text{ideal}} = \hat{H}^1 \cdot pts^1$$

Once these two sets of points have been transformed with respect to the second frame and second frame ideal frame, these transformed points can be used to calculate the homography, $\hat{H}^2$, between the second frame and second ideal frames themselves.

$$f^2 \rightarrow f'^2 : pts^2_{\text{ideal}} = \hat{H}^2 \cdot pts^2$$

$\hat{H}^2$ can then be used to transform the entire second frame so that the ideal homography matrix now defines the relationship between the first frame and the ideal second frame.

Calculating $\hat{H}^3$ is similar except now, $\hat{H}^2$ is used. This is more clearly explained in Fig. 1. The diagram shows a set of initial points, $pts^2$, from Frame 2, $f^2$, being mapped onto Frame 3, $f^3$ via the estimated homography matrix, $H^2$. This same set of initial points then undergoes two homography transformations, one to map the points onto the ideal frame 2 and a third to map those just transformed points onto ideal frame 3.

$$f^2 \rightarrow f^3 : pts^3 = H^2 \cdot pts^2$$
$$f^2 \rightarrow f'^3 : pts^3_{\text{ideal}} = \hat{H}^2 \cdot pts^2$$

A homography is then calculated for the two set of points as calculated,

$$f^3 \rightarrow f'^3 : pts^3_{\text{ideal}} = \hat{H}^3 \cdot pts^3$$

and a homography now exists to map Frame 3 onto Ideal Frame 3.

For all subsequent $\hat{H}_k$ a similar process can be applied.

$$f^{k-1} \rightarrow f^k : pts^k = H^{k-1} \cdot pts^{k-1}$$
$$f^{k-1} \rightarrow f'^k : pts^k_{\text{ideal}} = \hat{H}^{k-1} \cdot pts^{k-1}_{\text{ideal}}$$
$$f^k \rightarrow f'^k : pts^k_{\text{ideal}} = \hat{H}^k \cdot pts^k$$

This means that each subsequent frame transformation is dependent on all the frame transformations that have been calculated before that. Intuitively, you would want the current image to be aligned to minimize unwanted camera motion with respect to a previous image that has been aligned as well, but this creates a problem in the implementation. In particular, the initial homography estimations must be precise, otherwise should the estimation of an earlier homography matrix be incorrect, all subsequent homography matrices will also be incorrect.

4 Results

The video sequences were taken with a handheld digital camera capturing images at a rate of 15 fps. Since we wanted to remove motions that last less than 4 seconds, a Gaussian
blur of an appropriate size was applied to the translations. The datasets were taken with a subject capturing a video sequence while walking down a hall, or around a particular object. Each video sequence has a significant amount of unwanted motion due to each step the subject takes and from the subject’s inability to keep the camera absolutely stable. To better visualize the effects of convolving a vector of entries from the homography matrix, Fig 4. compares the measured values for $h_{13}$ of the homography matrix (black) and the smoothed out values (red). The motions incurred from each step of walking is clearly depicted in the sinusoidal form that each of the two graph takes.

The effects of the smoothing can also be seen in Fig. 2 and 3. The figures were created by taking 20 frames from both the corrected and uncorrected video sequences, and taking the average across the frames. These images illustrate the motions that were present before and after motion correction was applied, and the corrected averages display a distinct contrast to the uncorrected averages. Each of the corrected averages convey a camera motion that is much more distinct than the uncorrected averages which instead showed lots of spurious motions. For Fig. 2, pay attention to shelves, as the uncorrected average show more erratic blurring as compared with the corrected average.

5 Conclusions

Our camera stabilization routine was able to correct for many types of camera instability. The algorithm relied on very few assumptions except for those inherent in calculating homographies and the one we made regarding unwanted motion can easily be replaced with another equally valid one.

However, since the correct homography matrix of one frame depended on all those that preceded it, should one homography matrix be incorrectly estimated, the errors will be propagated to all the others after it. Unfortunately, the serial nature of these calculations cannot be helped, as the homography matrix of one frame is inherently dependent on the one previous to it. This makes our algorithm highly dependent on the accuracy of the initial homography estimations.

6 Future Work

In order for our camera stabilization algorithm to be applied to real world applications, object motion must be first be detected and and then separated from scene motion. Currently,
Figure 2: Walking down a hall. Both sequences uncorrected (left) and corrected (right) averaged over 25 frames

Figure 3: Moving across a table. Both sequences uncorrected (left) and corrected (right) averaged over 20 frames

Figure 4: The measured and smoothed vector containing values of the 7th entry of all the homography matrices
point correspondences are determined via normalized correlation, and should the brightness of the scene suddenly change (a cloud moving over the sun for example), we need to use a point correspondence algorithm that is not dependent on the pixel intensities alone.

The homography matrices contain structure that we are completely neglecting by simply blurring the entries of the matrices directly. If we were to instead, recover the actual relative rotations and translations of the camera from the homography matrices, the manipulations would be much more intuitive.

7 References

