Improved Video Registration using Non-Distinctive Local Image Features

Robin Hess and Alan Fern
School of Electrical Engineering and Computer Science
Oregon State University
Corvallis, OR 97331
{hess,afern}@eecs.oregonstate.edu

Abstract

The task of registering video frames with a static model is a common problem in many computer vision domains. The standard approach to registration involves finding point correspondences between the video and the model and using those correspondences to numerically determine registration transforms. Current methods locate video-to-model point correspondences by assembling a set of reference images to represent the model and then detecting and matching invariant local image features between the video frames and the set of reference images. These methods work well when all video frames can be guaranteed to contain a sufficient number of distinctive visual features. However, as we demonstrate, these methods are prone to severe misregistration errors in domains where many video frames lack distinctive image features. To overcome these errors, we introduce a concept of local distinctiveness which allows us to find model matches for nearly all video features, regardless of their distinctiveness on a global scale. We present results from the American football domain—where many video frames lack distinctive image features—which show a drastic improvement in registration accuracy over current methods. In addition, we introduce a simple, empirical stability test that allows our method to be fully automated. Finally, we present a registration dataset from the American football domain we hope can be used as a benchmarking tool for registration methods.

1. Introduction

Registering video frames with a static model is a common problem in many computer vision domains, including robot localization [13, 7], augmented reality [3, 14], sports analysis [5, 12], and others [1]. In general, video registration is required whenever we need to know what part of an object or scene a video frame depicts or where an object in that frame is located relative to a fixed coordinate system.

Consider, for example, our motivating problem of computing a high-level description of an American football play from video. A great deal of information about a particular play can be ascertained from the trajectories of the players. However, because the camera rapidly pans and zooms to follow the play's action, causing even a physically stationary player to appear to be moving as the video progresses, raw player trajectories in the video are meaningless from an interpretation standpoint. Prior to any interpretation step, therefore, player trajectories must be determined within the static football field coordinate system, where they are much more meaningful. This can be achieved by registering the football video with a model of the football field.

The standard approach to the registration problem is to compute, for each frame in the video sequence, a set of point correspondences between that frame and the model. These correspondences are then used to numerically determine a registration transform that maps the video frame to the model. The problem of finding such sets of correspondences was investigated specifically within the American football domain by Intille [5], who hypothesized that since the football field is (approximately) planar, the registration of football video with a 2-D football field model can be achieved by computing a planar homography mapping the video field surface to the model. A planar homography, which maps one plane to another, is a linear transform with eight degrees of freedom and can be computed from four or more 2D point correspondences [4]. Intille's approach to finding these correspondences involved locating, classifying and tracking line intersections on the field. Unfortunately, this method lacks generality, since many domains do not have such a precisely structured set of high level features as the lines on a football field. More importantly, because of the difficulty inherent in consistently detecting such high-level features, Intille's method proved to be unreliable and was abandoned in later work [6] in favor of tedious manual registration. In a set of informal experiments, we also found Intille's method to be ineffective, and we are unaware of any other successful demonstrations of

1-4244-1180-7/07/$25.00 ©2007 IEEE

robust registration of American football video.

Modern approaches to registration have taken advantage of recent breakthroughs in the detection [11] and description [10] of transform-invariant, local image features which are designed to facilitate consistent detection and easy, efficient matching between images. Using local feature techniques, the registration problem can be solved by assembling a set of reference images to represent the model and then detecting and matching local features between the model images and the video. [3], [13], and [7] are all examples of this type of approach.

Compared to Intille's method, local feature-based registration is attractive because of its generality and the proven robustness of finding reliable matches between distinctive local image features. As such, current local feature-based registration methods work well in domains with an ample supply of distinctive local features. It is important to note, though, that most current local feature-based methods rely *solely* on the presence of distinctive visual features for registration. Unfortunately, many domains can produce long segments of video without enough distinctive local features to robustly compute registration transforms, though there may still be many informative but non-distinctive features. In these domains, as we demonstrate in Section 3, relying completely on the presence of distinctive visual features for registration can result in crippling inaccuracy.

Again, the American football domain is a prime example of one in which total reliance on the presence of *distinctive* visual features can prove to be disastrous. There, important, distinctive visual features can be found at certain locations, such as within logos and around numbers on the field, but large regions of the field also exist that contain either no distinctive visual features at all or only a very small number of them. Often, video frames from the football domain depict only these latter regions of the field, making registration via distinctive feature matching either impossible or extremely unreliable. However, in video from the football domain, we can almost always guarantee the presence of *some* visual features, though they might be *non*-distinctive ones. For example, sets of identical hash marks, depicted in Figure 1, span the length of the football field, spaced one every yard. Such non-distinctive features convey a great deal of information about location on the field, and the ability to correctly match them to their corresponding model features would allow for robust computation of registration transforms. However, because these features are identical in appearance, they cannot be matched using common distinctive feature matching techniques.

The main contribution of this paper is to develop a generic registration approach that can leverage modern invariant feature techniques in domains like American football, where distinctive image features are often scarce but non-distinctive features are plentiful. Our method, which

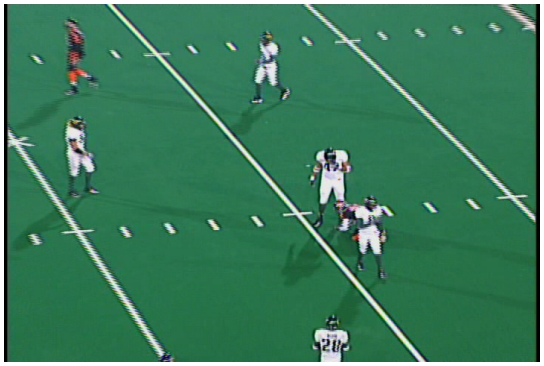


Figure 1. Some domains, such as the American football domain, shown here, produce images without distinctive visual features that can be easily matched. However, in these domains, the presence of *non*-distinctive visual features, such as the hash marks in the image above, can almost always be guaranteed.

is discussed in detail in Section 2, takes advantage of such non-distinctive visual features to register video, even in the absence of distinctive features. Specifically, we introduce a concept of *local* distinctiveness that enables us to find model matches for nearly all visual features in every video frame. In addition, we present a simple, empirical stability test that allows us to find a stable set of distinctive features with which to initialize the registration process, resulting in fully automatic registration.

Our approach is most similar in spirit to recent work by Okuma *et al*. [12]. Their approach avoids relying on distinctive features by utilizing generic point correspondences computed using the Kanade-Lucas-Tomasi tracking equation [15] along with edge-based model fitting. One major drawback of this approach is that it requires manual initialization for every video sequence to be registered. This can be cumbersome if a large collection of video must be processed, as is the case in our application domain. While Okuma *et al*.'s method is conceptually similar to the method we propose in this paper, our use of invariant image features, in conjunction with our initialization technique, allows for fully automatic operation.

Our empirical evaluation in Section 3, shows that, compared to distinctive feature-based approaches, our method is very effective in the challenging American football domain.

We also note that a secondary contribution of this work is to make available our substantial ground truth video dataset, which we hope can be used as a standard benchmarking tool for video registration methods.

2. Method

Our method registers a video sequence with a predefined, static model by finding point correspondences between the video and the model and using them to compute a registration transform for each frame. Under our method, video-

to-model point correspondences are found by matching invariant image features in the video to a set of features assembled from reference images to represent the model. In what follows, we describe how invariant image features are detected in the video and reference images; how the set of model features is assembled; how distinctive image features are matched to form video-to-model point correspondences; how these correspondences are used to compute registration transforms; and how accurate transforms can be computed, even in the absence of distinctive image features, by finding matches between *non*-distinctive image features using a concept of local distinctiveness.

2.1. Detecting Invariant Image Features

In this work, we use the Harris-affine detector [9] and SIFT descriptor [8] to detect and describe image features. Given an image as input, the Harris-affine/SIFT operator computes a set of feature points, each represented by a set of parameters describing the affine region surrounding the feature as well as a 128-dimensional descriptor vector. These features are invariant in that, in theory, the same ones will be detected in each of two images of the same object related by a reasonable degree of affine transformation, including translation, scale, in-plane rotation, and, to a limited extent, out-of-plane rotation. In addition, corresponding features in the two images will have very similar descriptors.

2.2. Assembling a Set of Model Features

There are several possible ways to form the set model features, denoted below as Π. Our goal is to do so in such a way that the model coordinates of the features in Π are known, thereby allowing us to determine the model coordinates of video features via feature matching.

In some domains, where the locations of model features are only important in relation to each other, Π can be formed simply and automatically by iteratively registering a set of reference images to each other [1, 3]. In other domains, however, the locations of features in Π must be known in reference to a specific global coordinate frame, such as a particular view of the model. In the American football domain, for example, we want to know the field coordinates (*e.g.* bottom hash on the home 35-yard line) of each model feature so that video frames can be localized on the field and not just registered to an arbitrary view of it. This is achieved by registering the set of reference images to a known view of the field, as depicted in Figure 2.

It is, in general, difficult to automatically register a set of reference images with specific coordinate frame in a domain-independent manner. Fortunately, the amount of manual work required to do so is minimal. For each reference image in the football domain, for example, we must specify just one set of four point correspondences to compute a planar homography mapping that image to the desired view of the field, as was done to generate the model in Figure 2. It is also possible to automatically register the set of reference images with each other, as when we are not aligning them with a specific coordinate frame, and thus to reduce the number of manually specified point correspondences to a single set for all reference images instead of one set per reference image.

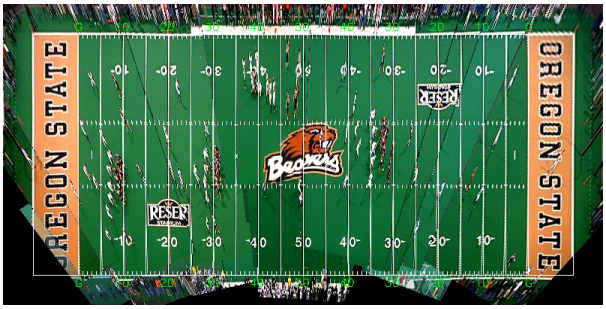


Figure 2. A set of reference images from the American football domain is registered with an overhead view of the field. Registering the set of reference images in this manner allows us to know the field coordinates of image features in the reference images and so to localize video frames on the field via feature matching.

2.3. Matching Image Features

In practice, it is not always the case that two images of the same object will produce all of the same image features, nor is it true that two descriptors belonging to matching features will be identical. Therefore, it is necessary to have some way to compute feature matches in which we can be highly confident. To do so, we make use of the 2NN heuristic proposed by Lowe in [8]. Given a feature X from video, we find from the set Π of model features X's two nearest neighbors, $\pi_1(X)$ and $\pi_2(X)$, with respect to the Euclidean distance between descriptor vectors. The 2NN heuristic considers X and $\pi_1(X)$ to be a distinctive match if, for a fixed threshold $\rho \in [0, 1]$,

$$\frac{\|d(X) - d(\pi_1(X))\|}{\|d(X) - d(\pi_2(X))\|} < \rho, \tag{1}$$

where $d(X)$ is the descriptor vector of feature X and $\|\cdot\|$ is the Euclidean norm. If (1) is not satisfied, X remains unmatched, even if X and $\pi_1(X)$ are, in fact, matching features. Figure 3 shows the results of matching features from a frame of football video to part of the model in Figure 2 using the 2NN heuristic.

By choosing ρ appropriately (we use $\rho = 0.6$), the 2NN heuristic yields a very small number of false positive matches. An unfortunate side-effect of this heuristic, however, is that it only finds matches between features whose descriptors are very different from those of the rest of the set of potential matching features. We call these features *globally distinctive*. The 2NN heuristic is generally incapable

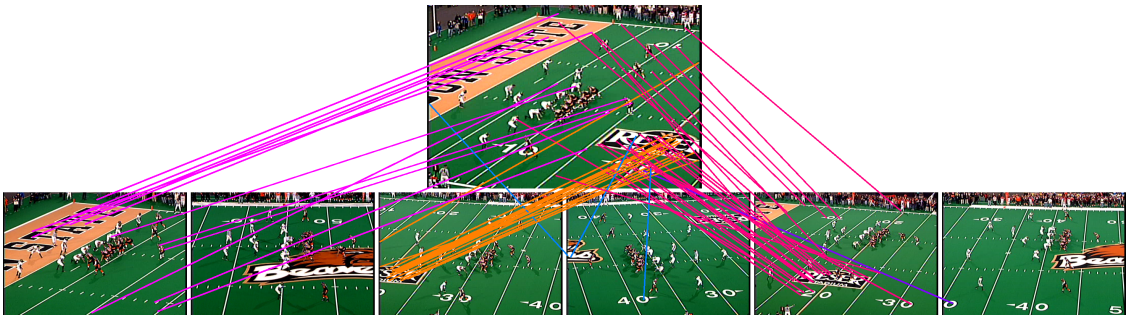


Figure 3. Image features from a video frame in the American football domain are matched using the 2NN heuristic to a portion of the model shown in Figure 2. There are very few false positive and many true positive ones. Most true positive matches, however are between distinctive image features, such as the field logo, with few correct matches between non-distinctive ones, such as the hash marks.

of establishing correspondences for features whose correct match in Π has a descriptor that is similar to many others in Π. Indeed, this is the case with any heuristic that attempts to minimize false positive matches while using only local context information for each feature. Unfortunately, as discussed above, relying solely on correspondences from globally distinctive features can impair our ability to successfully register video from some domains. In Section 2.5, we discuss a method to overcome this difficulty, but first we describe briefly how registration transforms are computed from video-to-model correspondences.

2.4. Computing Registration Transforms from Sets of Point Correspondences

Having constructed a set of model features and found a set of correspondences between each video frame and the model, it is possible to compute a registration transform for each frame. In the football domain, we can analytically compute homographies from four or more correspondences via least squares. Specifically, given a set of $n \geq 4$ video-to-model correspondences $\left((x_v^i, y_v^i), (x_m^i, y_m^i)\right)_{i=1}^n$, where the (x_v^i, y_v^i) are image coordinates in the video frame and the (x_m^i, y_m^i) are the corresponding model coordinates, a least squares planar homography can be computed by forming and solving the following linear system:

$$\begin{bmatrix} x_v^1 & y_v^1 & 1 & 0 & 0 & 0 & -x_m^1 x_v^1 & -x_m^1 y_v^1 \\ 0 & 0 & 0 & x_v^1 & y_v^1 & 1 & -y_m^1 x_v^1 & -y_m^1 y_v^1 \\ \vdots & & & \vdots & & & \vdots & \\ x_v^n & y_v^n & 1 & 0 & 0 & 0 & -x_m^n x_v^n & -x_m^n y_v^n \\ 0 & 0 & 0 & x_v^n & y_v^n & 1 & -y_m^n x_v^n & -y_m^n y_v^n \end{bmatrix} \begin{bmatrix} h_{11} \\ h_{12} \\ h_{13} \\ h_{21} \\ h_{22} \\ h_{23} \\ h_{31} \\ h_{32} \end{bmatrix} = \begin{bmatrix} x_m^1 \\ y_m^1 \\ \vdots \\ x_m^n \\ y_m^n \end{bmatrix}. \tag{2}$$

Here, the h_{ij} are the entries of the homography matrix. Because the homography is defined up to a scale factor, we may choose $h_{33} = 1$ [4]. We note that an alternative method for homography computation is the direct linear transform (DLT), which can handle the special case where $h_{33} = 0$. See [4] for a complete discussion of this topic.

Unfortunately, both least squares and DLT are very sensitive to outliers in the set of correspondences. In order to cope with the unavoidable presence of false correspondences, we use RANSAC [2] in conjunction with least squares to find a consistent set of inlier correspondences and a corresponding registration transform for each frame. We refer to the set of inliers as a frame's "core set", since it is comprised of correspondences in whose verity we are highly confident.

2.5. Using Local Distinctiveness to Find Additional Correspondences

The approach taken by current registration methods is to compute registration transforms using the procedure described between sections 2.1 and 2.4 (or some slight variation of it), using only correspondences between globally distinctive features found with the 2NN heuristic. As discussed above, and as demonstrated in section 3, this approach fairs poorly in domains where many frames lack globally distinctive image features. Our method attempts to maintain registration through a sequence of these frames by inducing correspondences between globally non-distinctive features using a concept of local distinctiveness. Specifically, we say that a feature X is *locally distinctive* relative to a spatial region R in a model or image, if it passes a spatially restricted 2NN test,

$$\frac{\|d(X) - d(\pi_1^R(X))\|}{\|d(X) - d(\pi_2^R(X))\|} < \rho, \tag{3}$$

where $\pi_i^R(X)$ is the i^{th} nearest neighbor of feature X within region R of the model or image. Note that even if X is not globally distinctive it can be locally distinctive relative to a particular R. If the region R can be selected so that it is likely to contain a correct match for X, then a locally distinctive match is likely to be a correct one.

Matching via local distinctiveness plays two roles in our registration method. The first is to track image features between frames. Because video is sampled at very high rates—typically around 30 frames per second—the amount of change between any two consecutive frames is

very small, and a given image feature is likely to move at most only a few pixels between those frames. We can take advantage of this fact by searching for a feature's locally distinctive match in the next frame relative to a small region R around the feature's spatial location in the current frame. By doing so, we essentially ensure our ability to find a correct match for that feature. The utility here lies in the fact that if a feature tracked to the current frame has previously been matched to the model, then that model match can be effectively carried over to the current frame. If the current frame does not have a sufficient set of globally distinctive matches, then these additional tracked matches, many of which may not be globally distinctive, can help produce a stable registration transform.

The second role of matching via local distinctiveness is to find new model matches for non-distinctive features. Specifically, if we can assume that we have found a core set of feature correspondences for the current frame that is sufficient for computing an accurate registration transform, we can use that transform to search for a locally distinctive model match for any unmatched feature X relative to a small region R around X's predicted model location. This is useful because it allows us to compute model correspondences for non-globally distinctive features whenever they appear in a video. These new matches can then be propagated through the video using the above tracking approach.

Our overall registration procedure uses the above two applications of matching via local distinctiveness as follows.

1. **Initialize.** Mark all frames as unprocessed and uninitialized. Select a frame for which the set of correspondences from globally distinctive features results in the "most stable" registration transform (see next section) after applying RANSAC. Initialize the core set of this frame to be the set of globally distinctive matches, and mark this frame as initialized.
2. **Include Globally Distinctive Features.** Select an unprocessed, initialized frame. Add all correspondences from globally distinctive features in the frame to its initial core set and use RANSAC on the entire set to compute a new expanded core set and its associated registration transform. Note that this step will not affect the core set of the initial frame selected in step 1.
3. **Include Unmatched Features.** For the selected frame, compute, as discussed above, a set of correspondences from features that are locally distinctive relative to the frame's registration transform. Union these correspondences with the current core set and use RANSAC to compute a final core set and the associated final registration transform for the frame. Mark this frame as processed.
4. **Model Match Propagation.** For each neighboring frame of the selected frame that has not been processed (either 1 or 2 frames), use the approach described above to attempt to track each of the features in the core set to the neighbor, and propagate forward the model matches of successfully tracked features. Initialize the neighbor's core set to the set of correspondences determined by the propagated matches, and mark the neighbor as initialized.
5. **Loop.** If unprocessed frames remain, go to step 2.

This approach allows us to maintain a large core set of video-to-model correspondences and to add new correspondences to that set as new features appear in the video frame. In this way, as long as there are enough good features—either globally or locally distinctive—in the video frame to produce an accurate registration transform, the core set, once formed, is self-sustaining, since an accurate registration transform allows us to find model correspondences for all video features matchable via local distinctiveness.

All that remains then is to determine, in step 1, which frame to select as the initial frame to process. This should be a frame for which we are most certain that the set of correspondences from globally distinctive features is sufficient for computing an accurate registration transform. It is, in general, dangerous to assume that the first frame of video will always be such a frame. We therefore make the assumption that at least one such frame exists in the video and develop a test for finding one of them. This test is presented in the next section. If the single-good-frame assumption does not hold, we may revert to manual initialization.

2.6. Stability Test for Core Set Initialization

The stability of a set of correspondences can be computed analytically by forming the least squares system in (2) and measuring its conditioning using techniques from existing theory on the conditioning of least squares problems. This theory, discussed at length in [16], provides bounds on the error amplification factor that ensues from small perturbances in the input. Unfortunately, we have found that, in practice, this analytical approach often yields a poor choice of initial core set. In turn, we have developed an empirical stability test, which is outlined in Algorithm 1.

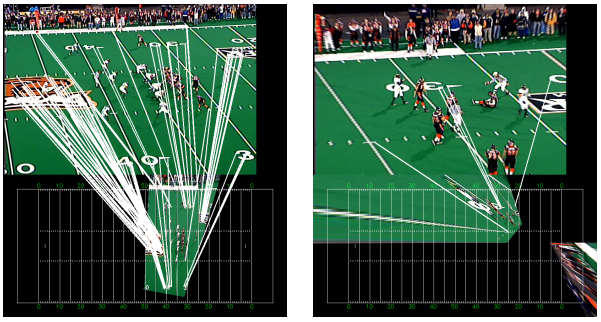
In general, sets of correspondences that produce stable transforms are large and widely distributed spatially. Sets of low cardinality whose correspondences are not well distributed, on the other hand, are very sensitive to small amounts of noise. Figure 5 helps to elucidate the difference between these two types of sets. Intuitively, our stability test identifies those sets of correspondences that are the most invulnerable to small amounts of noise.

The value **S** in Algorithm 1 represents a measure of how drastically predicted model locations change with slight perturbations in the input set of correspondences. Sets that result in very low values of **S** are least sensitive to measurement noise and produce the most stable transforms. We

Algorithm 1 Stability Test for Core Set Initialization

C: Potential initial core set
K: User defined number of iterations
S: Output stability

1: $\mathbf{T} \leftarrow$ Registration transform from $\mathbf{C}$
2: $\mathbf{L} \leftarrow$ Set of randomly sampled image locations
3: $\mathbf{L_T} \leftarrow$ Model coordinates of $\mathbf{L}$ via $\mathbf{T}$
4: $\mathbf{S} \leftarrow 0$
5: **for** $i \leftarrow 1..\mathbf{K}$ **do**
6: $\quad \widehat{\mathbf{C}} \leftarrow \mathbf{C}$ perturbed with $\mathcal{N}(0, \epsilon)$ noise
7: $\quad \widehat{\mathbf{T}} \leftarrow$ Registration transform from $\widehat{\mathbf{C}}$
8: $\quad \mathbf{L}_{\widehat{\mathbf{T}}} \leftarrow$ Model coordinates of $\mathbf{L}$ via $\widehat{\mathbf{T}}$
9: $\quad \mathbf{S} \leftarrow \mathbf{S} + \text{ERROR}(\mathbf{L}_{\widehat{\mathbf{T}}}, \mathbf{L_T})$
10: **end for**



(a) $\mathbf{S} = 1.12 \times 10^2$ (b) $\mathbf{S} = 1.60 \times 10^7$

Figure 5. The large and widely distributed set of correspondences in (a) yields an acceptable registration transform, but the small, isolated set in (b) yields one that is worthless. The quality of each set as a core set initializer is reflected by the value **S** returned by the stability test outlined in Algorithm 1.

initialize the registration process using the frame whose set of correspondences from globally distinctive features yields the lowest value of **S**.

3. Experiments

Our method was tested on a set of 25 video sequences from the American football domain[1], each between 280 and 500 frames in length. Sequences were selected from two different games to cover as much of the field surface as possible. Most of the sequences contain a significant number of frames without distinctive field features. However, almost all frames contain *some* field features, such as the hash marks depicted in Figure 1, for which a model match exists. Every tenth frame of every video has an associated set of hand-labeled ground truth video-to-model point correspondences. There are between 300 and 700 such hand-labeled correspondences for each video, for a total of about 12,000.

Our model was constructed from a set of 23 reference images as described in section 2.2 and as depicted in Figure 2. Reference images were selected from video not included in the dataset unless achieving total field cover in the reference set required us to select a frame from the dataset.

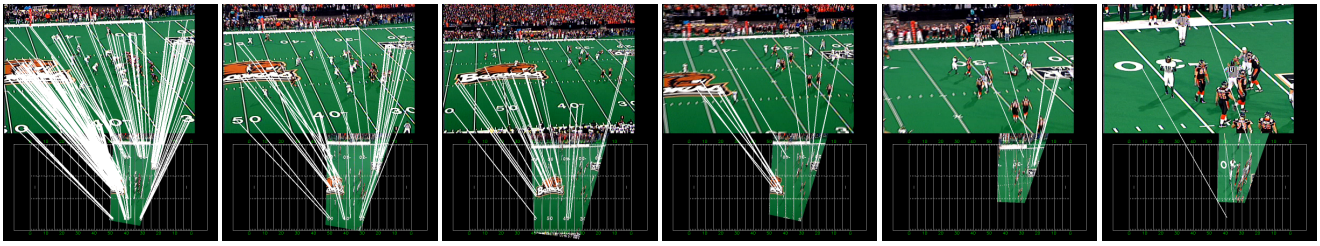
For comparison, we tested two other methods using the same dataset and model. The first, which we call *naïve registration* (NR), uses only correspondences between globally distinctive features found using the 2NN heuristic. The second, *registration with uninitialized matching via local distinctiveness* (RUMLD), uses matching via local distinctiveness as described in Section 2.5 but always initializes the core set using the first frame instead of using the initialization technique described in Section 2.6. Our complete method is referred to below as *registration with initialized matching via local distinctiveness* (RIMLD). All three methods include a sanity check that reverts to the last good registration transform if the current frame's transform becomes grossly unacceptable.

Figure 4 illustrates some registration results from NR and RIMLD that are representative for the American football domain. As is typically the case, both methods are accurate at the beginning of the video, where there is generally a large set of globally distinctive image features in the frame. However, as the video progresses, the set of correspondences used by NR to compute registration transforms gradually dwindles to the point of instability. At this point transforms computed by NR fail the sanity check, and NR reverts to the last known good transform, resulting in registration error that snowballs as the video proceeds to the end. RIMLD, on the other hand, maintains a large, stable set of correspondences throughout the length of the video, and registration is accurate to the end.

Registration accuracy for all three methods was quantified by computing the mean registration error for every tenth frame of every video in the dataset using the set of hand-labeled ground truth correspondences described above. Results were normalized to equal length by partitioning them into twenty quantums, and both the mean and maximum errors were computed for each quantum.

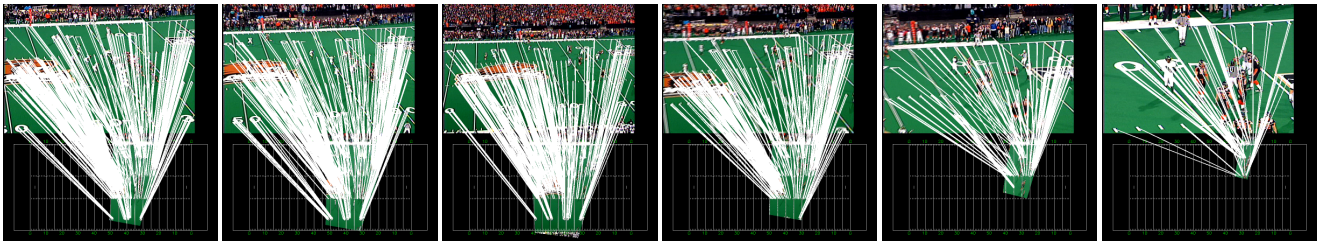
Interestingly, the results for RIMLD and RUMLD differ significantly on only a single video sequence. For this video, as might be expected, RUMLD's error rate is quite high at the beginning of the video, but it quickly reduces to nearly equal that of RIMLD after the core set has been expanded through a combination of locally and globally distinctive features. The similar performance of RIMLD and RUMLD can be explained by the fact that, with this one exception, the first frame of every video sequence in the dataset contains a large portion of one or more of the field logos or the end zones, where there are many distinctive features. Accordingly, these frames produce a stable enough

[1]Our registration dataset is publicly available online at `http://eecs.oregonstate.edu/football/registration/dataset`.



Frame 1	Frame 150	Frame 250	Frame 300	Frame 325	Frame 407
232 Correspondences	50 Correspondences	31 Correspondences	15 Correspondences	10 Correspondences	2 Correspondences

(a) Registration with NR



Frame 1	Frame 150	Frame 250	Frame 300	Frame 325	Frame 407
525 Correspondences	480 Correspondences	390 Correspondences	333 Correspondences	210 Correspondences	125 Correspondences

(b) Registration with RIMLD

Figure 4. The sequences above illustrate typical registration results on a 407 frame video from the American football domain. For this video, NR maintains accuracy as long as the frame contains a stable set of correspondences from globally distinctive features. As early as frame 250, this set loses stability, and registration becomes slightly inaccurate. By frame 325, the set of correspondences, down to 10 and concentrated within a small region of the image, yields a transform that fails the sanity check, and registration reverts to the last good transform. On the same video, RIMLD maintains a stable set of at least 100 video-to-model correspondences throughout the run, and registration is accurate until the end of the video.

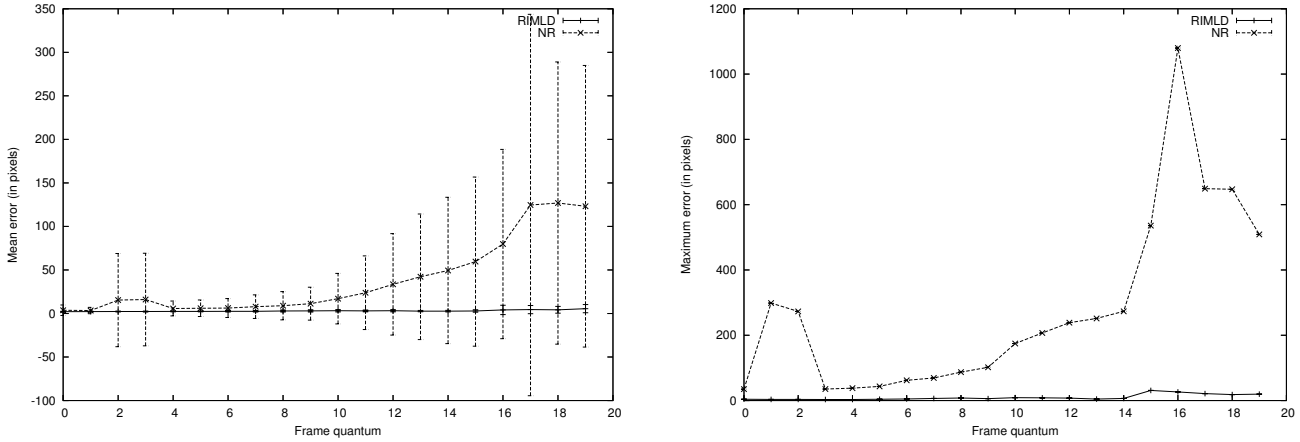
set of correspondences with which to initialize registration.

The results for RIMLD and NR, however, *do* differ significantly on nearly all videos in the dataset. The error for these two methods over the entire dataset is summarized in Figure 6. To put these error rates into perspective, we note that six pixels in our model are equal to one yard on the football field. This means that, even during important parts of the football play, NR's average error rate approaches 10 yards—quite significant if these results are to be used in an interpretation system—while RIMLD maintains an average error of around one half yard.

The reason RIMLD is so much more accurate than NR is because RIMLD is able to maintain a model match for nearly every visual feature in the video frame for which a model match exists. Even at the end of a video sequence, the core set often contains on the order of one hundred or more well distributed correspondences. By comparison, the set of correspondences determined by NR often shrinks to between 25 and 50, or even less, as early as halfway through the video, dwindling soon afterwards to ten or less. If such a small set of correspondences is not widely distributed, a small amount of error in either the video or model coordinates of the image features can become dramatically exaggerated in the resulting registration transform.

Besides raw registration error, another important gauge of registration quality is "smoothness." Because the appearance of a physical feature changes slightly in the video as time progresses, its feature descriptor changes also, and correspondences with a 2NN heuristic value near the threshold ρ may step back and forth over that threshold between frames. As the composition of the set of correspondences from globally distinctive features changes thus from frame to frame, registration with NR is prone to jitter. Because of the nature of the matching process in RIMLD, on the other hand, the composition of the core set of correspondences changes only slightly as the video moves from frame to frame, and registration with RIMLD is thus much smoother.

Of course, RIMLD is not perfect. As can be seen in Figure 6, RIMLD's error rate does also increase slightly towards the end of the video, reaching a maximum of around 30 pixels, or 5 yards in our model. The explanation for this slight increase is that, many times, in the last frames of the video sequence, the camera is zoomed so far in that there simply are not enough features in the frame, globally distinctive or otherwise, to robustly compute a registration transform from feature correspondences. This is an issue future registration methods—especially those that use only local image features—may need to address. However, in



(a) Mean registration error per frame quantum

(b) Maximum registration error per frame quantum

Figure 6. The above two plots depict (a) the mean and (b) the maximum registration error per frame quantum for RIMLD and NR over the entire dataset of 25 video sequences. The error bars in (a) indicate one standard deviation. By both measures, RIMLD is remarkably more accurate than NR, especially later in the video, when many frames contain few or no globally distinctive features. Note that six pixels in our model are equal to one yard on the football field.

American football video, the important action in the play is usually over by the time RIMLD begins to show signs of inaccuracy, so we do not concern ourselves with this matter.

4. Conclusions

In this paper, we introduced a method for video registration that uses invariant local image features in conjunction with a matching technique based on a concept of local distinctiveness that finds video-to-model correspondences between non-distinctive features. In addition, we presented a simple empirical stability test that provides a means by which our registration method can be fully automated under the assumption that at least one frame in the video—not necessarily the first—contains a stable set of correspondences from globally distinctive features. Our technique was shown to yield significantly more accurate registration results in the challenging American football domain than methods that rely only on the presence of globally distinctive features for registration. Finally, we offered our significant ground truth video dataset to the community for use as a benchmarking tool for video registration methods.

Acknowledgments

This work is supported in part by NSF grant IIS-0307592. In addition to NSF, we would also like to thank the coaches and staff of the Oregon State University football team for providing us with our video data.

References

[1] M. Brown and D. G. Lowe. Recognising panoramas. In *Proc. Intl. Conference on Computer Vision*, pages 1218–1225, 2003.

[2] M. A. Fischler and R. C. Bolles. Random sample consensus: a paradigm for model fitting with applications to image analysis and automated cartography. *Communications of the ACM*, 24(6):381–395, 1981.

[3] I. Gordon and D. G. Lowe. Scene modelling, recognition and tracking with invariant image features. In *Proc. Intl. Symposium on Mixed and Augmented Reality*, pages 110–119, 2004.

[4] R. Hartley and A. Zisserman. *Multiple View Geometry in Computer Vision*. Cambridge University Press, 2nd edition, 2004.

[5] S. Intille. Tracking using a local closed-world assumption: Tracking in the football domain. Technical Report 296, MIT Media Laboratory Perceptual Computing Section, 1994.

[6] S. Intille. *Visual Recognition of Multi-Agent Action*. PhD thesis, Massachusetts Institute of Technology, 1999.

[7] J. Košecká and X. Yang. Location recognition and global localization based on scale invariand keypoints. In *Proc. Workshop on Statistical Learning in Computer Vision, ECCV*, 2004.

[8] D. G. Lowe. Distinctive image features from scale-invariant keypoints. *Intl. Journal of Computer Vision*, 60(2):91–110, 2004.

[9] K. Mikolajczyk and C. Schmid. Scale & affine invariant interest point detectors. *Intl. Journal of Computer Vision*, 60(1):63–86, 2004.

[10] K. Mikolajczyk and C. Schmid. A performance evaluation of local descriptors. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 27(10):1615–1630, 2005.

[11] K. Mikolajczyk, T. Tuytelaars, C. Schmid, A. Zisserman, J. Matas, F. Schaffalitzky, T. Kadir, and L. van Gool. A comparison of affine region detectors. *Intlernational Journal of Computer Vision*, 65(1-2):43–72, 2005.

[12] K. Okuma, J. Little, and D. Lowe. Automatic rectification of long image sequences. In *Proc. Asian Conf. on Computer Vision*, 2004.

[13] S. Se, D. G. Lowe, and J. Little. Global localization using distinctive image features. In *Proc. Intl. Conference on Intelligent Robots and Systems*, pages 226–231, 2002.

[14] G. Simon, A. Fitzgibbon, and A. Zisserman. Markerless tracking using planar structures in the scene. In *Proc. Intl. Symposium on Augmented Reality*, pages 120–128, 2000.

[15] C. Tomasi and K. Takeo. Detection and tracking of point features. Technical Report CMU-CS-91-132, Carnegie Mellon University, 1991.

[16] L. N. Trefethen and D. Bau III. *Numerical Linear Algebra*. Society for Industrial and Applied Mathematics, 1997.