ROBUST VIDEO REGISTRATION APPLIED TO FIELD-SPORTS VIDEO ANALYSIS

Bernard Ghanem, Tianzhu Zhang

Advanced Digital Sciences Center of Illinois
Singapore

Narendra Ahuja

University of Illinois at Urbana-Champaign
Electrical and Computer Engineering
Urbana, IL USA

ABSTRACT

Video (image-to-image) registration is a fundamental problem in computer vision. Registering video frames to the same coordinate system is necessary before meaningful inference can be made from a dynamic scene in the presence of camera motion. Standard registration techniques detect specific structures (e.g. points and lines), find potential correspondences, and use a random sampling method to choose inlier correspondences. Unlike these standards, we propose a parameter-free, robust registration method that avoids explicit structure matching by *matching* entire images or image patches. We frame the registration problem in a sparse representation setting, where outlier pixels are assumed to be sparse in an image. Here, robust video registration (RVR) becomes equivalent to solving a sequence of ℓ_1 minimization problems, each of which can be solved using the Inexact Augmented Lagrangian Method (IALM). Our RVR method is made efficient (sublinear complexity in the number of pixels) by exploiting a hybrid coarse-to-fine and random sampling strategy along with the temporal smoothness of camera motion. We showcase RVR in the domain of sports videos, specifically American football. Our experiments on real-world data show that RVR outperforms standard methods and is useful in several applications (e.g. automatic panoramic stitching and non-static background subtraction).

Index Terms— registration, homography, ℓ_1 minimization

1. INTRODUCTION

Video registration refers to the problem of spatially aligning video frames in the same absolute coordinate system determined by a reference image. The spatial transformation between the t^{th} video frame $\mathbf{I}_t$ and the reference image $\mathbf{I}_r$ governs the relative camera motion between these two images. Effectively estimating the mapping between coordinate systems is imperative for any computer vision application that makes use of object positions in a dynamic scene including object tracking and action recognition. These applications assume that pixel motion in video frames is only due to moving objects and not to *apparent motion* resulting from changes in camera parameters (e.g. pan, tilt, and/or zoom). This is why video registration in the presence of camera motion is a common problem in many computer vision related domains (e.g. augmented reality [1] and sports video analysis [2, 3]) and a fundamental pre-processing step that is necessary before meaningful higher level inference can be performed on dynamic scene content.

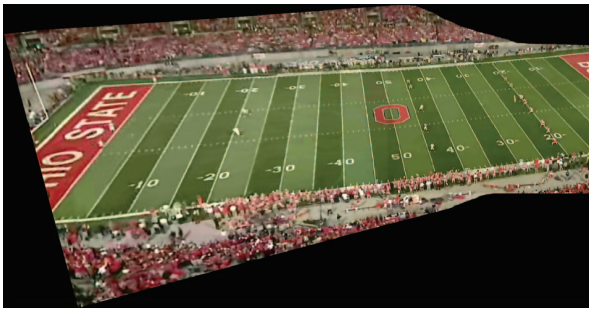

Fig. 1. Example of panoramic stitching using our robust registration method applied to an American football video

Let us consider the case of analyzing broadcast field sports video (specifically American football), which is the driving application considered in this paper. In sports video, the camera captures segments of the field at a time where important events (e.g. touchdown) occur. Since the entire field is never captured in the same frame, the camera undergoes pan-tilt-zoom (PTZ) transformations following events unfolding on the field. To reliably interpret these events, higher level inference (e.g. player tracking) is required. However, motion-based player tracking is rendered useless, if camera motion is not compensated for. For example, a player who is motionless on the field appears to be moving when the camera pans.

The most common spatial transformation between consecutive frames is the projective transform, known as homography. For a static scene, a non-translating camera undergoing PTZ transformations captures a sequence of video frames, where each consecutive pair of frames is related by a homography irrespective of scene geometry [4]. It is well known that only 4 point correspondences (or 4 line correspondences [5]) are needed to estimate the homography. The standard registration approach consists of detecting a set of distinctive feature points (e.g. SIFT), finding proper matches (e.g. between SIFT descriptors of SIFT features [6]), and estimating $\mathbf{H}_{t,t-1}$ using the Direct Linear Transformation (DLT) method [4]. However, since some feature points and/or some matches may be outliers to the underlying homography, the popular random

sampling method (RANSAC) is employed [6]. For example, in broadcast American football, outliers are pixels belonging to moving players, audience, or referees. In what follows, we refer to this standard technique as the Feature-Matching-RANSAC (FMR) method. Once homographies are computed, many registration-based applications can be handled including automatic panoramic stitching [6] (Figure 1), spatiotemporal trajectory alignment [7], and sports video analysis [2].

Despite the prevalence of the FMR method, its performance is dependent on the accuracy of feature detection and matching. In American football, the detected features concentrate on the painted digits and the logo in the middle of the field. In some cases (e.g. camera zoom), these distinctive features disappear altogether leading to degradation in performance. To improve stability, higher order features are detected and matched such as lines [5] or ellipses [8]. These structures are not easily detected in images (due to view change and occlusion) and are not generalizable to generic video. Moreover, potential feature matches are established independently and are usually based on a heuristic matching criterion, such as, comparing the translation of matched features to a threshold. Also, RANSAC performance is contingent on what its parameters are. For instance, one RANSAC parameter that highly impacts performance is the inlier probability, which assumes that the percentage of inliers is known apriori. This is a strong assumption in general, so this parameter is usually user-defined or application-dependent.

In this paper, we propose a robust video registration (RVR) framework that makes use of recent theoretical advances in sparse representation and compressive sampling. Unlike FMR methods, our RVR framework avoids the instability of detecting and matching points, lines, or other primitives structures by *matching* entire images or image patches. No explicit correspondence between features is required. The matching process computes an optimal homography that maps one image into the other by assuming that outlier pixels are sufficiently sparse in each image. No other prior information is assumed here. The underlying optimization problem is modeled as an ℓ_1 minimization problem that can be solved iteratively and efficiently using the Inexact Augmented Lagrangian Method (IALM). If point correspondences are available and reliable, they can be seamlessly incorporated into the RVR framework as additional linear constraints. RVR is parameter-free except for tolerance values (stopping criteria) that determine when convergence occurs. We take explicit measures to reduce the computational cost of solving the ℓ_1 problem. Spatially, we employ a coarse-to-fine strategy based on random subspace projection. Temporally, we exploit the smooth temporal variation of camera motion.

The paper is organized as follows. In Section 2, we give a mathematical formulation of the video registration problem embedded in a sparse representation framework. Section 3 describes our proposed method and how each homography is solved for. In Section 4, we validate our RVR framework by analyzing real-world American football videos and using it in popular registration applications (e.g. automatic panorama generation and non-static background subtraction), while also comparing it to standard FMR methods when necessary.

2. PROBLEM FORMULATION

Given F video frames (e.g. American football play), we aim to estimate a sequence of homographies that map consecutive video frames. Denote $\mathbf{I}_t \in \mathbb{R}^{M\times N}$ as the image at time t and $\vec{\mathbf{i}}_t$ as its vectorized version. We also denote the homography from $\vec{\mathbf{i}}_t$ to $\vec{\mathbf{i}}_{t+1}$ as $\vec{\mathbf{h}}_t$. Let $\tilde{\vec{\mathbf{i}}}_{t+1} = \vec{\mathbf{i}}_t \circ \vec{\mathbf{h}}_t$ be the result of spatially transforming image $\vec{\mathbf{i}}_t$ using $\vec{\mathbf{h}}_t$. We denote the error arising from outliers pixels (e.g. pixels belonging to moving players) as $\vec{\mathbf{e}}_t = \tilde{\vec{\mathbf{i}}}_{t+1} - \vec{\mathbf{i}}_{t+1}$. This error vector is assumed to be sufficiently sparse, which is valid for many dynamic scenes containing background. Here, we assume that the homographies are general (i.e. 8 general DOF). This can be changed to accommodate prior models on the nature of each homography (e.g. rotation and slight zoom). Note that this framework is also available for image patches, whereby multiple patches in one image jointly undergo the same homography. This will only add more linear equality constraints.

The RVR problem becomes equivalent to estimating the optimal sequence of homographies that map consecutive frames *and* render the sparsest error. This allows for reliable representation and robustness to outliers. In this paper, we do not explicitly model the temporal relationship between homographies. This decouples the problem into $F-1$ optimization problems. More elaborate temporal modeling may be added as constraints or regularization terms; however, such modeling is beyond the scope of this paper.

We give a mathematical description of our RVR framework next. Note that what follows is the basic formulation of the framework. It can be extended to more general cases, as we will mention later. For every $1 \le t \le F-1$, we ideally seek the sparsest solution (minimum ℓ_0 norm), but since this problem is NP-hard in general and is non-convex due to the nonlinear constraints, we replace the cost function with its convex envelope (ℓ_1 norm) in (1). This relaxed problem has been recently used in face recognition [9] and texture analysis [10].

$$\begin{aligned} &\min_{\vec{\mathbf{e}}_{t+1},\vec{\mathbf{h}}_t} \ \|\vec{\mathbf{e}}_{t+1}\|_1 \qquad (1)\\ &\text{subject to:} \ \ \vec{\mathbf{i}}_t \circ \vec{\mathbf{h}}_t = \vec{\mathbf{i}}_{t+1} + \vec{\mathbf{e}}_{t+1} \end{aligned}$$

Although the objective function is convex, the equality constraint is not. Similar to recent image alignment works [9], we linearize the constraint around a current estimate of the homography and solve the linearized convex problem iteratively. Therefore, at the $(k+1)^{\text{th}}$ iteration, we start with an estimate of each homography denoted as $\vec{\mathbf{h}}_t^{(k)}$. Here, the current estimate will be $\vec{\mathbf{h}}_t^{(k+1)} = \vec{\mathbf{h}}_t^{(k)} + \triangle\vec{\mathbf{h}}_t$. With this relaxation, (1) is relaxed to (2), where $\vec{\delta}_{t+1}^{(k)} = \vec{\mathbf{i}}_{t+1} - \vec{\mathbf{i}}_t \circ \vec{\mathbf{h}}_t^{(k)}$ represents the error incurred at iteration k and $\mathbf{J}_t^{(k)} \in \mathbb{R}^{MN\times 8}$ the Jacobian of $\vec{\mathbf{i}}_t \circ \vec{\mathbf{h}}_t$ with respect to $\vec{\mathbf{h}}_t$. Applying the chain rule, $\mathbf{J}_t^{(k)}$ can be written in terms of the spatial derivatives of $\vec{\mathbf{i}}_t$. The problem in (2) is a linear program and thus can be solved in polynomial time. Although this linearized version has not been shown to converge to a local minimum of (1), our empirical results (as

well as results in [9]) suggest that convergence is guaranteed especially when the solution is initialized in a neighborhood of a local minimum. We observed that only a small number of iterations (about 10) is required for convergence.

$$\min_{\triangle\vec{\mathbf{h}}_t, \vec{\mathbf{e}}_{t+1}} \|\vec{\mathbf{e}}_{t+1}\|_1 \quad (2)$$
$$\text{subject to: } \mathbf{J}_t^{(k)} \triangle \vec{\mathbf{h}}_t - \vec{\mathbf{e}}_{t+1} = \vec{\delta}_{t+1}^{(k)}$$

3. PROPOSED RVR FRAMEWORK

In this section, we describe how the k^{th} iteration of (2) is solved and how the sequence of homographies is computed. The optimization problem in (2) is convex yet non-smooth due to the ℓ_1 objective. In this paper, we solve (2) using the Inexact Augmented Lagrangian Method (IALM), which is an iterative method whose update rules are simple and closed form and whose convergence rate is linear [9, 10]. In IALM, constraints are added as penalty terms in the objective function with first order *and* second order Lagrangian multipliers. The augmented Lagrangian function for (2) is $\mathcal{L} = \|\vec{\mathbf{e}}_{t+1}\|_1 + \vec{\lambda}^T(\mathbf{J}_t^{(k)} \triangle \vec{\mathbf{h}}_t - \vec{\delta}_{t+1}^{(k)} - \vec{\mathbf{e}}_{t+1}) + \frac{\mu}{2}\|\mathbf{J}_t^{(k)} \triangle \vec{\mathbf{h}}_t - \vec{\delta}_{t+1}^{(k)} - \vec{\mathbf{e}}_{t+1}\|_2^2$. This unconstrained objective is minimized using alternating optimization steps, which lead to simple closed form update rules. Updating $\triangle\vec{\mathbf{h}}_t$ requires solving a least squares problem. Conversely, updating $\vec{\mathbf{e}}_{t+1}$ makes use of the well-known ℓ_1 soft-thresholding identity $\mathcal{S}_\lambda(\vec{\mathbf{a}}) = \arg\min(\lambda\|\vec{\mathbf{x}}\|_1 + \frac{1}{2}\|\vec{\mathbf{x}} - \vec{\mathbf{a}}\|_2^2)$, where $\mathcal{S}_\lambda(\mathbf{a}_i) = \max(0, |\mathbf{a}_i| - \lambda)$. For more details on IALM, refer to [9]. The overall RVR algorithm is summarized in Algorithm 1, where the inner *while* loop designates the IALM. In all our experiments, the stopping criterion compares successive changes in the solution to a threshold $\epsilon_{\text{IALM}} = 0.1$.

Algorithm 1: *Robust Video Registration (RVR)*

Input : $\vec{\mathbf{i}}_t\ \forall t, \vec{\mathbf{h}}_t^{(0)}\ \forall t, k = 0, \rho > 1$

while *not converged* **do**
  Compute $\vec{\delta}_t^{(k)}$ and $\mathbf{J}_t^{(k)}\ \forall t$
  $m = 0, \triangle\vec{\mathbf{h}}_t^{(m)} = \vec{\mathbf{0}}, \vec{\lambda}^{(m)} = \vec{\mathbf{0}}, \mu^{(m)} > 0$
  while *not converged* **do**
    $\vec{\mathbf{e}}_{t+1}^{(m+1)} = \mathcal{S}_{\frac{1}{\mu^{(m)}}}(\mathbf{J}_t^{(k)} \triangle \vec{\mathbf{h}}_t^{(m)} - \vec{\delta}_{t+1}^{(k)} + \frac{\mu^{(j)}\vec{\lambda}^{(m)}}{2})$
    $\triangle\vec{\mathbf{h}}_t^{(m+1)} = (\mathbf{J}_t^{(k)T}\mathbf{J}_t^{(k)})^{-1}\mathbf{J}_t^{(k)T}(\vec{\delta}_{t+1}^{(k)} + \vec{\mathbf{e}}_{t+1}^{(m+1)} - \frac{\vec{\lambda}^{(m)}}{\mu^{(m)}})$
    $\vec{\lambda}^{(m+1)} = \vec{\lambda}^{(m)} + \mu^{(m)}(\mathbf{J}_t^{(k)} \triangle \vec{\mathbf{h}}_t^{(m+1)} - \vec{\delta}_{t+1}^{(k)} - \vec{\mathbf{e}}_{t+1}^{(m+1)})$
    $\mu^{(m+1)} = \rho\mu^{(m)};\ \ m \longleftarrow m + 1$
  end
  $\vec{\mathbf{h}}_t^{(k+1)} = \vec{\mathbf{h}}_t^{(k)} + \triangle\vec{\mathbf{h}}_t^{(m)};\ \ \vec{\mathbf{e}}_{t+1}^{(k+1)} = \vec{\mathbf{e}}_{t+1}^{(m)}$
end

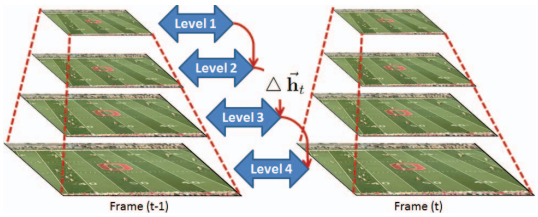

Fig. 2. Coarse-to-Fine processing

Efficient Implementation and Extensions: Although IALM is more efficient than first-order descent methods, it remains computationally expensive – linear in the number of pixels. So, we employ spatial and temporal strategies to improve efficiency. Temporally, camera motion varies smoothly, so we initialize $\mathbf{H}_{t+1}$ with $\mathbf{H}_t$. Spatially, we use a coarse-to-fine strategy, where the solution at a coarser level is the initialization for a finer one (refer to Figure 2). Also, to reduce the number of pixels processed per level, we randomly sample pixels to consider in the update equations above. This builds on theoretical guarantees in the compressive sensing community [11], whereby randomly projecting the original problem unto a subspace with lower dimensionality (d) will *almost always* yield the same solution if it is sufficiently sparse. If α_t denotes the ratio of nonzero elements in $\vec{\mathbf{e}}_t$ and d_{MIN} the minimum subspace dimensionality, then d_{MIN} is the smallest nonnegative scalar that satisfies $\log d_{\text{MIN}} + \frac{d_{\text{MIN}}}{2\alpha_t MN} \geq \log MN$. By setting $\alpha_t = \frac{\|\vec{\mathbf{e}}_{t-1}\|_1}{MN}$, we can adaptively select the random sampling rate. In many cases, α_t is small and only $15 - 20\%$ of pixels are sampled. These spatial and temporal strategies reduce runtime by as much as 200%. For example, when $M = 480$ and $N = 640$, computing a homography takes 3 seconds. Using parallelized batch processing, the per-frame runtime reduces to below 1 second, which is on-par with optimized FMR methods. Moreover, (2) is extendable to the case where auxiliary prior knowledge on outlier pixels is known. This prior is represented mathematically as a matrix $\mathbf{W}$ that pre-multiplies $\vec{\mathbf{e}}_{t+1}$ yielding a weighted version of (2). In the simplest case, $\mathbf{W} = diag(\vec{\mathbf{w}})$ where $\mathbf{w}_i$ is the probability that pixel i is an inlier. For example, if a human detector is available, $\mathbf{w}_i$ is inversely proportional to the detection score. If $\mathbf{W}$ is invertible, the same IALM can solve this problem using a simple change of variable $\tilde{\mathbf{e}}_{t+1} = \mathbf{W}\vec{\mathbf{e}}_{t+1}$.

4. EXPERIMENTAL RESULTS

Here, we validate the performance of our RVR method by applying it to real-world sports data from the American football domain. For this purpose, we compiled a dataset of 90 plays from online sources. Each play consists of $400 - 650$ frames each of 480×640 pixels. We intend to make this dataset publicly available in the future. First, we compare the performance of our RVR method to the standard FMR method (SIFT, DLT, and RANSAC). For fair comparison, we manually set the inlier probability of RANSAC to an appropriate value for every

play. Other parameters (e.g. thresholds for SIFT detection and matching) were set to their default values. We compute the average residual error ratio between the two methods. The error mask for each image is constructed by thresholding $\vec{\mathbf{e}}_t$. In Figure 3, we show the average error ratio and an example of the error produced by each method. Clearly, RVR outperforms FMR on average by at least 10%. The RVR error concentrates on the players only, while FMR leads to noticeable misregistration error, especially at field lines.

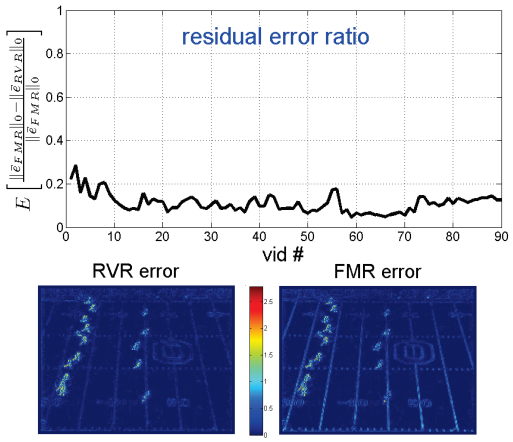

Fig. 3. Quantitative comparison between FMR and RVR

Next, we apply our RVR method to 3 applications that showcase how the homographies can be used for further analysis. Non-static background subtraction is the first application. Here, camera motion is compensated for to separate foreground from background. We detect foreground pixels by adaptively thresholding $\vec{\mathbf{e}}_t$ in each frame followed by morphological operations. Figure 4 shows an example of this detection. Clearly, foreground pixels are reliably detected with few false positives, which is very useful for multi-object tracking.

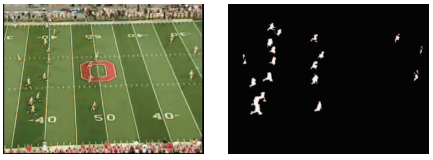

Fig. 4. Non-static background subtraction using RVR

Also, we automatically generate a panoramic stitch of the field as in Figure 1. We use multiband blending [6] and bilateral filtering to create the panoramic. In the third application, we decompose the estimated homography into its constituent camera motion parameters (i.e. zoom, pan, and tilt). If a PTZ camera is non-translational and f_t is its focal length at time t, then it is easy to show that $\sqrt{det(\mathbf{H}_t, 1)} = \frac{f_1}{f_t}$ (defines the zoom factor). In Figure 5, we plot this zoom factor over time for a single play. As expected, the plot is piecewise linear in broadcast video. We detect all piecewise linear intervals and map each interval into one of three zoom regions: "No Zoom", "Zoom In", and "Zoom Out". This mapping is done by simply setting cutoff values for each zoom region. This information is quite informative, since zoom in broadcast sports is indicative of the importance of a particular event within each play.

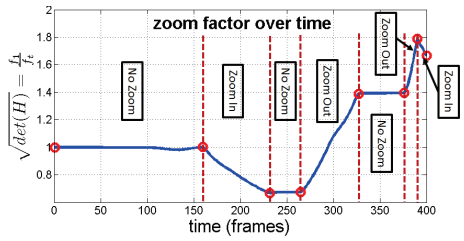

Fig. 5. Zoom factor (in blue) during an example play

5. CONCLUSION

In this paper, we propose a robust and efficient video registration framework that builds upon recent advancements in sparse representation and compressive sensing. Homographies are computed between image pairs by solving a sequence of relaxed ℓ_1 minimization problems using a hybrid coarse-to-fine and random sampling strategy. Experimental results on sports videos show that our method significantly outperforms standard registration methods.

Acknowledgements: This work is supported by the research grant for the Human Sixth Sense Programme at the Advanced Digital Sciences Center from Singapore's Agency for Science, Technology and Research (A*STAR).

6. REFERENCES

[1] G. Simon, A.W. Fitzgibbon, and A. Zisserman, "Markerless tracking using planar structures in the scene," in *IEEE and ACM International Symposium on Augmented Reality*. 2000, pp. 120–128, IEEE.

[2] Robin Hess and Alan Fern, "Improved Video Registration using Non-Distinctive Local Image Features," *CVPR*, 2007.

[3] Kenji Okuma, James J. Little, and David G. Lowe, "Automatic rectification of long image sequences," in *ACCV*, 2004, pp. 1–6.

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

[5] J. J. Guerrero and C. Sagues, "From lines to homographies between uncalibrated images," *Pattern Recognition and Image Analysis*, 2001.

[6] Matthew Brown and David G. Lowe, "Automatic Panoramic Image Stitching using Invariant Features," *IJCV*, vol. 74, pp. 59–73, 2006.

[7] Y. Caspi and M. Irani, "Spatio-temporal alignment of sequences," *IEEE TPAMI*, vol. 24, no. 11, pp. 1409–1424, Nov. 2002.

[8] Ankur Gupta and JJ Little, "Using Line and Ellipse Features for Rectification of Broadcast Hockey Video," in *CVR*, 2011.

[9] Andrew Wagner, John Wright, Arvind Ganesh, Zihan Zhou, Hossein Mobahi, and Yi Ma, "Towards a Practical Face Recognition System: Robust Alignment and Illumination by Sparse Representation.," *IEEE TPAMI*, May 2011.

[10] Zhengdong Zhang, Xiao Liang, Arvind Ganesh, and Yi Ma, "TILT: transform invariant low-rank textures," *ACCV*, pp. 314–328, 2011.

[11] Emmanuel J. Candes and Terence Tao, "Near-Optimal Signal Recovery From Random Projections: Universal Encoding Strategies?," *IEEE Transactions on Information Theory*, vol. 52, pp. 5406–5425, 2006.