Researchers at the Advanced Digital Sciences Center (ADSC) have achieved a major advance in the state of the art for video registration of dynamic scenes, and demonstrated this advance in applications drawn from several different domains. The new technique, called sparse registration, typically increases the accuracy of video registration by 20% in consecutive frames of video footage of dynamic scenes, making it much easier to track objects in these scenes and understand the activities taking place in them, even if the video is from cameras that pan, tilt, and zoom. As this ability is needed for many potential applications in medicine, sports, surveillance, and beyond, we expect sparse registration to replace the conventional video registration approach that has been the de facto standard for quite some time.

Sparse registration needs no tuning of parameters to fit the application domain. Exploiting theoretical guarantees of convergence, sparse registration computes the optimal transformation to map one video frame to another, even if many pixels correspond to moving objects in the scene. ADSC’s current multi-scale implementation of sparse registration has running time similar to that of today’s standard method for video registration. A short video about sparse registration is accessible from http://adsc.illinois.edu.

This work is part of ADSC’s broader program for recognizing, understanding, and analyzing the actions of people as they go about their daily activities, which is a grand challenge in itself.

1 Image and Video Registration

Put simply, image registration is the process of determining exactly where images overlap with each other. For example, consider the three photos in Figure 1, taken from the top of a playground slide, which we registered by hand to give the integrated view shown in Figure 2. When a toddler sits at the top of the slide, she instantly and effortlessly understands how everything she sees on the playground fits together. But registration is very hard to do on a computer, especially if the camera moves. Registration must ignore temporary parts of the scene, such as the parent walking across the playground, or the shadows temporarily cast by moving clouds or people. But it is very difficult for a computer algorithm to determine which aspects of the scene are temporary.

Further, the computer must guess whether and how the camera is panning, tilting, zooming, and tracking, while the toddler knows how her body and eyes are moving. The toddler’s built-in understanding of her own “camera motion” tells her which apparent motion in the scene before her must be due to the movements of objects in the scene, rather than to her own motion. In contrast, the computer might conclude that a playmate looks bigger because the camera zoomed or moved forward, rather than because the playmate ran toward the camera.

More technically, registration refers to the problem of spatially aligning images in the same absolute coordinate system determined by a reference image. The spatial transformation from any image to the reference image describes the relative camera motion between these two images.
Video is made up of a sequence of still images, so video registration is the process of determining how successive video images fit together into a bigger scene. Good registration is vital for analyzing dynamic scenes. We cannot correctly track objects as they move in a video unless we understand how each video frame (image) fits into the overall scene. Without good tracking of object positions and motion, we cannot determine what actions and activities are taking place in the video. Thus the video must be registered before we can begin other analysis tasks, so that these tasks can assume that the pixel motion in video frames is only due to moving objects, and not to the apparent motion resulting from camera changes such as panning, tilting, and zooming. In sum, a preprocessing step of video registration is required for almost any sort of higher-level analysis of dynamic scenes that include camera motion.

2 Feature Matching RANSAC (FMR): The State of the Art in Video Registration

The Feature Matching RANSAC (FMR) method has been the de facto standard for image registration for at least five years now. Popular versions of FMR employ five basic steps. (1) Use Scale Invariant Feature Transform (SIFT) to find distinctive feature points in the images to be registered. SIFT’s points of interest are typically corners with sharp changes in appearance, as indicated by the green circles in Figure 3. (2) Find potential matches between SIFT points in the two images. Points match if their descriptors (histograms of oriented gradients around each SIFT point) are similar, according to a heuristic measure. (3) In theory, just 4-8 matches are needed to register two images. But FMR uses the popular random sampling method RANSAC to ensure that inappropriate matches, such as matches on moving objects, do not ruin the registration.

To use SIFT for video registration, we can consider each pair of successive video frames as two images to be registered. By observing how SIFT’s points of interest move relative to one another in the two frames, we can estimate the spatial transformation between the two frames, that is, the mapping from each pixel in one frame to a pixel or pixels in the other. This mapping is called a homography when the camera’s motion is limited to panning, tilting, and zooming. If the scene is static, such as the view from a panning camera on top of a mountain, then the motion in the scene is attributable to the camera, and FMR tends to register the frames properly. Popular commercial products that use FMR include Autostitch, which pieces together a panoramic image from a series of photos taken from a single viewpoint. Microsoft’s Photosynth uses FMR to support 3D visualization of a set of 2D photographs of a scene.

On the other hand, if the scene contains moving objects, then FMR methods face a bigger challenge. Intuitively, the dynamic aspects of a scene – the outliers – should not register, because their pixels do not conform to the homography between consecutive frames. FMR users must specify the fraction of
pixels that are outliers, which is scene-dependent. If FMR is not well tuned, then the registration can be poor, and tuning is hard when the scene is dynamic. For example, video of kindergarteners running around the playground in their school uniforms presents the worst possible challenge for FMR video registration: frequent occlusion as children pass by one another, many moving objects with similar appearance, and rapid scale change as children run toward and away from the camera. In this case, FMR’s points of interest are as likely to be located on the moving objects as on the static part of the scene that FMR is trying to register.

As another example, consider footage of a parent walking across a playground, shot at 30 frames per second. The green circles and blue lines in Figure 4 show FMR’s top choices and matches for points of interest in two successive frames. Most of these points lie on the moving parent rather than the static scene. Although RANSAC is responsible for removing these outliers, as the fraction of outliers grows, so does the probability that these outliers will be used to compute the final homography. In this particular video, the strolling parent will cause FMR’s homography for the playground to be off by 3-6 pixels for each consecutive pair of frames. This relatively small frame-to-frame misalignment adds up to significant registration errors once each video frame is registered to the reference frame.

If we use FMR to register the 11-second playground video from page 1, we get the result in Figure 5. Due to the outliers on the moving parent, significant misalignments occur especially at the pole, the number 6, and the blue circle, as compared to the images on page 1. When shooting panoramic footage without a tripod, typically the photographer rotates in place and the camera traces a circle around her, so that in effect the images are captured on a cylinder. When we project a cylinder of images onto a flat printed page, the registered images will stretch and become skewed as we move away from the reference image. However, the extreme skewing and stretching on the right side of Figure 5 is due to accumulated errors in registration, caused by the walking parent.

More generally, if the scene is very dynamic, such as a playground with children, airport terminal, sports field, or ADSC’s building lobby at lunchtime, then the frequent occlusions, rapid scale change, and large number of moving objects with similar appearance will yield significant errors in FMR registration. This happens because moving individuals or even visual intersections between moving people in successive frames contribute to the homography, rather than only the hardscape that is in fact the static part of the scene. Such significant registration errors will definitely impact the performance of higher level analysis, such as human tracking and action recognition. Not surprisingly, biological evidence suggests that humans use a very sparse representation for interesting things that they see, such as faces, and do not use an FMR-like approach for registration.
Whether these errors matter depends on the application. When registering MRI images to prepare for brain surgery to remove a suspected blood clot, then a minimal error rate is crucial. When aligning images taken at 1 minute intervals as a contrast medium flows into a patient’s eye, for an automated check on the progress of macular degeneration, then an error of 3-4 pixels in successive frames will cause the analysis program to significantly misinterpret the shape of the patient’s blood vessels and their leaks, reducing the chance of clinical acceptance of automated assistance. When registering video of a sports playing field, so that we can track the movements of players on the field and analyze each team’s strategy, FMR registration makes major errors, as shown in Figure 6. For example, the 10-yard-line is misplaced by about 7 yards.

If we forego registration entirely and perform player tracking directly using state-of-the-art particle-based filtering methods, then the tracking errors are extensive. Observe the jaggedness of the tracked trajectories depicted in the bird’s-eye view in Figure 7 for a 237-frame kickoff play in American football, which even includes a detour off the field. We cannot understand and analyze players’ movements on the field without accurate registration.

The efficiency and accuracy of FMR registration techniques have made them the dominant state-of-the-art algorithms for registering many kinds of images and videos. Extending the benefits of registration to scenarios that are too dynamic for FMR requires a fundamental change in how the problem of registration is conceptualized and addressed.
3 Sparse Representation and Sparse Registration

Mathematically, we can think of an image as a point \( I \) in a very high-dimensional space, where each dimension represents the color at a particular location in the image. But there is too much detail in that high-dimensional space, and it is very hard for a computer algorithm to know which details matter and which should be ignored. The solution is to map \( I \) to a point \( m(I) \) in a much lower-dimensional space that omits all the irrelevant detail in the image, and analyze the image in that lower-dimensional space. The analysis will be much more tractable there. But in the computer vision world, it was never clear which lower-dimensional space(s) are the best choice for quickly getting high quality analysis results, for almost any kind of high-level analysis task.

This began to change in 2006, when mathematicians proved that under certain conditions, with very high probability, we can efficiently find a lower-dimensional space and mapping \( m \) such that the Manhattan distance of \( m(I) \) from the origin is as small as possible, yet from \( m(I) \) we can still reconstruct \( I \). These conditions imply that the representation of \( I \) is as simple as possible in this lower-dimensional space, and we call \( m(I) \) the sparse representation of \( I \). Many natural phenomena satisfy those conditions, including most images, so this result was very exciting. For example, Illinois professor Yi Ma’s research group used sparse representation to build a system that is almost perfect at recognizing people’s faces, even when they are wearing dark glasses [WYGSM09]. More precisely, they showed that sparse representation is highly suitable for recognizing faces under extreme conditions of lighting, scale, and occlusion, and even some spatial transformations. ADSC subsequently extended this approach to speed up the analysis [GA11].

ADSC researchers were the first to apply sparse representation to the problem of video registration. We hypothesized that the outliers in a video scene – the moving parts that should not register – should be sparse. For example, the dynamic aspect of the playground scene is the parent’s trajectory, which should have a sparse representation. More precisely, there should be a sparse representation for the difference between each image and its successor, even if the camera pans, tilts, or zooms, and/or people move around in the scene. Thus by maximizing the sparsity of the changes in successive frames, which we call the sparse registration of the video [GZA11], we can register the video while simultaneously picking out exactly the objects of interest for many applications, which are the outliers. Technically speaking, given frame \( I \) and its successor \( I' \), we search for the homography \( h \) to apply to \( I \) such that \( h(I) - I' \) has the sparsest representation, i.e., the point representing the error image \( h(I) - I' \) is as close as possible to the origin, under the appropriate distance measure. The homography \( h \) is estimated by an iterative process, which our experiments have shown to converge to the best possible homography, even if the camera pans, tilts, or zooms. In this way and unlike FMR methods, sparse registration avoids the instability of detecting and matching points of interest, lines, or other
primitive structures, and instead matches entire images or patches of those images. Sparse registration assumes only that outlier pixels are sufficiently sparse in each successive frame; no other prior knowledge is needed or used. In particular, sparse registration has no tuning parameters, so users never need to adjust the system before registering a new video.

Figure 8 shows the result of applying sparse registration to the playground video excerpted on page 1. Note the crisp, accurate edges and minimal stretching compared to the corresponding FMR registration in Figure 6. The remaining visible stretch and skew comes from the projection of a cylindrical sequence of photos onto a flat page; a tripod would have reduced the stretch to insignificance, as shown in the next example.

Figure 9 shows the result of applying sparse registration to footage of a kickoff play in American football, during which the camera pans and zooms. Note that the field and its markings look crisp and stretching is negligible, since the camera was not hand-held and so the cylindrical effect is minimal. The players are in their configuration from the first frame of the video, with no confusion from later frames.

By coupling sparse video registration with adaptations of the particle-filter-based object tracking methods mentioned earlier, ADSC has increased the accuracy of tracking in dynamic scenes by 35-40% [ZGA11], compared to popular and traditional methods such as the boosted particle filter and online Adaboost methods, which address the tracking and registration problems independently. Some of these methods fail dramatically when tracking real-world videos. For example, in sports broadcast video, human-to-human interactions lead to severe occlusions; the camera motion from panning, tilting, and zooming leads to apparent motion and scale change in the scene; and the team uniforms lead the tracking algorithm to drift between similar-looking players who are close to each other. ADSC’s joint registration-tracking algorithm overcomes these challenges in many real-world videos, including those from the sports domain.
For example, Figure 10 shows the result of sparse registration plus tracking of the kickoff starting configuration in Figure 9, up through and including image $I$, producing a proposed registration of the field in the next frame $I'$ that consists of $h(I)$ plus the outliers, which are exactly the players. The resulting trajectories for the kickoff play are very smooth and rational, as shown in the bird’s-eye view of the field in Error! Reference source not found..

The run-time of ADSC’s current implementation of sparse registration is similar to that of the standard FMR method. We are developing an optimized implementation that capitalizes on the parallelized nature of the problem, which will make sparse registration particularly computationally attractive.

Overall, ADSC’s tuning-parameter-free multi-scale implementation of sparse registration is always more accurate than the standard FMR method on consecutive video frames. The degree of improvement over FMR depends on how dynamic the scene is. Footage of nearly static scenes typically shows a 5% improvement in accuracy, while highly dynamic sports footage can have 40% improvement. The typical average accuracy improvement in significantly dynamic footage is 20%. As applications that are impractical with 70%-accurate registration become feasible when the accuracy exceeds 90%, sparse registration’s boost in accuracy will open many doors for new applications in computer vision [GKFZA11]. Since video registration is very important for dynamic scene understanding and is traditionally used as early stage processing for higher level operations, we expect sparse registration to replace the conventional FMR registration approach that has been the de facto standard for quite some time.
About Us

The Advanced Digital Sciences Center (ADSC), established in 2009 and located in Singapore, is a wholly-owned subsidiary of the University of Illinois at Urbana-Champaign. ADSC’s research program in interactive digital media is funded by the Agency for Science, Technology, and Research (A*STAR).

Narendra Ahuja is the Donald Biggar Willet Professor in the Department of Electrical and Computer Engineering at the University of Illinois at Urbana-Champaign. He is a Fellow of the IEEE, American Association for Artificial Intelligence, International Association for Pattern Recognition, Association for Computing Machinery, American Association for the Advancement of Science, and International Society for Optical Engineering. Narendra received his Ph.D. from the University of Maryland in 1979.

Bernard Ghanem is a senior research scientist at ADSC, where he is exploiting techniques in computer vision and machine learning for real-world applications including semantic sports video analysis, biomedical image analysis for diagnostics and discovery, and efficient object detection in video. Bernard received his Ph.D. from the University of Illinois at Urbana-Champaign in 2010.

Tianzhu Zhang is a postdoctoral fellow at ADSC. His research interests are in computer vision and multimedia, including action recognition, video surveillance, and object tracking. Tianzhu received his Ph.D. from the Institute of Automation, Chinese Academy of Sciences, Beijing, in 2011.

References