Vision, Modeling, and Visualization (2010)

Geometry-aware video registration

Gianpalo Palma[†1,2], Marco Callieri[2], Matteo Dellepiane[2], Massimiliano Corsini[2], Roberto Scopigno[2]

[1]Department of Computer Science, University of Pisa, Italy
[2]Visual Computing Lab, ISTI-CNR, Pisa, Italy

Abstract

We present a new method for the accurate registration of video sequences of a real object over its dense triangular mesh. The goal is to obtain an accurate video-to-geometry registration to allow the bidirectional data transfer between the 3D model and the video using the perspective projection defined by the camera model. Our solution uses two different approaches: feature-based registration by KLT video tracking, and statistic-based registration by maximizing the Mutual Information (MI) between the gradient of the frame and the gradient of the rendering of the 3D model with some illumination related properties, such as surface normals and ambient occlusion. While the first approach allows a fast registration of short sequences with simple camera movements, the MI is used to correct the drift problem that KLT tracker produces over long sequences, due to the incremental tracking and the camera motion. We demonstrate, using synthetic sequences, that the alignment error obtained with our method is smaller than the one introduced by KLT, and we show the results of some interesting and challenging real sequences of objects of different sizes, acquired under different conditions.

Categories and Subject Descriptors (according to ACM CCS): I.4.1 [Image Processing and Computer Vision]: Digitalization and Image Capture I.4.8 [Image Processing and Computer Vision]: Scene Analysis

1. Introduction

The camera tracking problem has been extensively studied in the last few years, and several interesting and robust solutions have been proposed. The purpose is to identify and track the most salient 2D features of the video and to use these features and their trajectories to recover the motion of the camera and some three-dimensional information about the scene.

Due to the main aim of these techniques, which is to provide a way to render additional elements inside a real-world video, the camera motion and scene information recovered by these approaches are correct up to a scale factor that depends on the characteristics of the scene and of the camera motion and that is difficult to evaluate. Additionally, in most cases, this scale is *non-linear* and changes in time and even across the scene. While using this type of data it is possible to render a 3D model as an additional component of the scene, every attempt to project/unproject data between the video and the 3D model is bound to fail. This problem remains even using more advanced methods of structure from motion [PGV*04] or state-of-the-art camera tracking software [TB09].

On the other side, the alignment (registration) of a 2D image with a 3D model is a very well know issue in the computer graphics field. Different solutions, both semi-automatic and completely automatic, have been proposed in the late years, which are able to align images to dense geometries coming, for example, from 3D scanning. However, despite the availability of such methods, the trivial idea of applying the semi-automatic or even the more automatic methods for 2D-to-3D registration to *each and every* frame of the video would result in a high computation time.

Given the amount of works in the 3D computer graphics field which make a profitable use of 3D-registered images to enrich digital models, being able to exploit the advantages of video sequences (frame-to-frame coherence, redundancy of data) could be a great help in different applications. If an accurate registration of the video on the 3D model is obtained, the bi-directional data transfer could be used for a number of interesting applications (color transfer, estimation

† gianpaolo.palma@isti.cnr.it

DOI: 10.2312/PE/VMV/VMV10/107-114

of reflectance properties, recording of appearance-varying scenes). Up to now, no solutions have been proposed to accurately align a video sequence over a mesh using the redundancy and the high frame-to-frame coherence of the video.

This paper presents a method to efficiently but accurately align a video sequence to a dense 3D geometry, combining the speed and flexibility of the feature-based tracking and the high precision and geometrical consistency of the image registration approaches. The proposed method combines the KLT tracking with a state-of-the-art image registration technique based on Mutual Information [CDPS09], a statistical measure of the information shared by the image and a rendering of the model. These two approaches can be considered as orthogonal, since they deal with different information extracted from data (feature vs. statistical analysis). Both the approaches are needed because the MI corrects the drifting effect of KLT tracking, while KLT tracking speeds up the registration and controls the convergence of MI towards good camera parameters.

2. Related Work

The work proposed in this paper is related to two important and different topics: camera tracking by point features and image-to-geometry registration. In this section we summarize the state-of-art of these topics.

2.1. Camera Tracking

The camera tracking based on point features is important and intensively studied in the field of Augment Reality. The more challenging aspect is the detection and the tracking of image features and the creation of correspondences between 2D features and their 3D coordinates.

Some solutions proposed a marker-based tracking, where artificially designed markers, easy to detect with image processing algorithms, are used to simplify the detection and the creation of 2D-3D correspondences [NF05] [KB99]. Even if the detection and tracking of markers are very reliable, in some cases the preparation of the scene with them is not possible. In such cases, a *markerless* tracking based on the natural features of the environment can be used.

The markerless tracking is based on two components: a feature detector and a feature descriptor for matching. A good detector should be repeatable and reliable. Repeatability means that the same feature can be detected in different images. Reliability means that the detected point should be distinctive enough so that the number of its matching candidates is small. Several detectors have been designed: rotation invariant [HS88]; scale invariant [MS01] [Low99]; affine invariant [TG00]. A descriptor should be invariant to rotation, scale, affine transformation and changes of illumination so that the same feature on different images could be characterized by almost the same values. Some common solutions are Sum of Square Differences (SSD) and Normalized Cross Correlation (NCC) of patches around the feature, SIFT descriptor [Low99], with its different versions, and SURF descriptor [BTG06]. A recent framework with SURF local tracking was proposed in [TCGP09].

The KLT tracker, presented in [ST94], is a specific tracker for video sequences which uses the high frame-to-frame data coherence. It extends the local estimation of optical flow proposed in [LK81] to track a template patch under an affine transformation model with the assumption of small brightness changes between consecutive frames. This type of tracking presents a drift problem due to several causes: image noise, geometric distortion, illumination changes, occlusions, fast camera movements, 3D features which leave the camera's field of view and reappear after in the sequence. Different solutions were proposed to compensate illumination changes [ZZCW07] [JFS01] and merge unconnected features track [CVG04] [THWS08]. A further extension of KLT tracker was proposed by Dame [DM09], where the SSD is substituted by MI for the feature matching between images.

2.2. Image-to-Geometry Registration

The image-to-geometry registration allows to align one or more images of an object taken at different times and from different viewpoints to the geometry of the object itself. Robust manual approaches have been proposed [FDG*05] for general cases, where an interactive tool allows to select a set of correspondences both between the 3D model and an image, and between images, in order to minimize the user intervention.

On the other side, the creation of automatic registration procedures is more challenging. This goal can be achieved by analyzing the image features [NK99] or using the reflectance value acquired during scanning [IOT*07]. These semi-automatic approaches need a preliminary calibration of the intrinsics of the camera, and require a constant illumination for all images. Another approach relies on the analysis of the silhouette of the object [LHS00]. Unfortunately, the use of silhouette matching has two important limitations: it must be easy to distinguish the object with respect to the background and this needs controlled setup acquisition or a time-consuming manual or automatic preprocessing; the object must be entirely present inside each image. A recent work for 3D-3D and 2D-3D automatic registration [LSY*06] can be applied in a more general case, but under the assumption that the 3D scene contains clusters of vertical and horizontal lines, like urban scenes. An more robust extension for indoor environment was proposed by Li et al. [LL09], where the lack of features on large uniform surfaces are resolved by projection of special light patterns to artificially introduce new features.

Other methods for automatic registration are based on the maximization of Mutual Information. The first methods proposing this technique were developed by Viola and

Wells [VW97] and by Maes et al. [MCV*97]. The Viola's alignment approach uses the mutual information between the surface normal and the image brightness to correlate shading variations of the image with the surface of the model. Leventon et al. [LWG97] extended this alignment framework to use multiple views of the object when a single image does not provide enough information. Since then, several registration methods based on MI have been proposed, especially for medical images [PMV03]. A more recent approach was proposed in [CDPS09], where Viola's approach is extended using several types of rendering, such as ambient occlusion, normal map, reflection map, and combined versions of them, with a new optimization strategy based on the recent algorithm NEWUOA [Pow08].

3. Video Registration

Our algorithm assumes a perspective camera model defined by two groups of parameters: intrinsic parameters related to the internal characteristics of the camera; extrinsic parameters associated with the position and the orientation of the camera in the space. The intrinsic camera parameters, except for the focal length and the lens radial distortion, are assumed as being pre-determined. More specifically, the skew factor is assumed to be zero, the principal point is set as the center of the image and the horizontal and vertical scale factors are assumed to be known from the image resolution and the CCD dimensions. The focal length is assumed constant for the whole video sequence and it is estimated only for the first frame. The lens radial distortion is estimated only once, using a single frame of a black and white checkerboard to automatically extract the position of the corners to give in input to the camera calibration method defined in [Tsa87] in the case of coplanar points. The extrinsic parameters define the rotation matrix, parameterized by the Euler angles $(\theta_x, \theta_y, \theta_z)$, and the translation vector (t_x, t_y, t_z) that are needed to transform the camera coordinate system into the world coordinate system.

The algorithm takes in input a video sequence of the object acquired with a constant zoom factor and a dense triangular mesh of this object; then, it computes the camera parameters for each frame. The algorithm is composed by two tasks, the feature-based registration and the registration by MI, preceded by a preprocessing step to extract the 2D features tracks from the video.

3.1. Preprocessing

The output of the preprocessing is composed by the camera parameters of the first frame and the 2D features tracks extracted by the video. First of all, the video is deinterlaced (if necessary), and noise is removed by bilateral filtering in order to allow a more robust 2D features tracking. Then the radial distortion introduced by the camera lens is eliminated from all frames.

Starting from the processed frames, we execute the last two subtasks to produce the needed data for the algorithm. The first subtask is the alignment of the first frame over the 3D model by manual selection of a set of 2D-3D correspondences to use in the Tsai's calibration method [FDG*05], followed by a further refinement with the MI [CDPS09]. In this way, the focal length and the extrinsic parameters of the first camera are computed. The second subtask is the extraction and saving of the 2D feature tracks of the video by using the Voodoo Camera Tracker tool [TB09]. This tool uses a KLT tracker to detect and track the features and applies a RANSAC approach to make a robust estimation of the fundamental matrix by eliminating the outliers.

3.2. Registration algorithm

The registration algorithm works in an incremental manner: to align the i-th frame, we start from the registration of the $(i-1)$-th frame. From the camera parameters of the previous frame and the 2D features tracking information, we extract a set S of 2D-3D correspondences to solve a non-linear least square problem to compute the camera pose with the Levenberg-Marquardt algorithm. For the extraction of the set S we compute a validity mask from the depth map of the frame F_{i-1}. This mask allows to discard all the 2D features of the frame F_i with a corresponding 2D point in the previous frames that does not belong to the object or that lie near to depth discontinuities. Then for all valid 2D features, we assign the 3D point computed by projection the corresponding 2D features in the frame F_{i-1} onto the 3D model.

To estimate the quality of the registration, given the set S of 2D-3D correspondences $\langle m, M \rangle$ and the camera projection matrix P, we compute an alignment error E:

$$E = \frac{1}{|S|} \sum_{\langle m,M \rangle} d(M, P^{-1}m) \qquad (1)$$

where the function d computes the geometric distance between the 3D point assigned to the 2D features by the previous frame and the 3D point computed by backward projection of the 2D features with the camera P onto the 3D model. We compute the error E as the average distance of 3D points keeping constant the 2D features positions. The averaging permits to have a comparable error for all frames because the number of correspondences is not constant during the sequence. If the alignment error E is above a threshold, we apply the registration by MI. This threshold is adaptive and is proportional to the objects surface area sampled by a single pixel of the camera. To be more precise, it is equal to the ratio between the width of the camera frustum at the distance of the object from the camera center and the width in pixels of the image. The distance of the object from the camera center is computed as the average between the near and the far plane of the camera to display only the portion of the object in the frustum.

After the alignment by MI, we recompute the correct 2D-

3D correspondences of the current camera needed for the registration of the following frames. Subsequently, we update all cameras between the current frame F_i and the last one F_{i-k} aligned by MI. For each camera in this interval we extract the correspondences with the frames F_i and, for those cameras which have a minimum number of correspondences, we recompute new extrinsic parameters with the Levenberg-Marquardt algorithm based on the 2D features shared with the frame F_i. Finally, for each of these frames we linearly interpolate the new extrinsic camera parameters with those computed before with the forward tracking. With this step, we obtain a continuous and smooth camera path without gaps. The final task of the algorithm is to update the set of 2D-3D correspondences with the new 2D features of the current frame which were not detected in the previous frame.

For each frame this process is iterated until the set of 2D-3D correspondences is updated with the addition of the new 2D features.

3.3. Registration by Mutual Information

Mutual Information measures the information shared by two random variables A and B. Mathematically, this can be expressed using entropy or joint probability. Following this interpretation, the Mutual Information $\mathcal{MI}$ between two images I_A and I_B can be defined as:

$$\mathcal{MI}(I_A, I_B) = \sum_{(a,b)} p(a,b) \log\left(\frac{p(a,b)}{p(a)p(b)}\right) \qquad (2)$$

where $p(a,b)$ is the joint probability of the event (a,b), $p(a)$ is the probability that a pixel of I_A gets value a and $p(b)$ is the probability that a pixel of I_B gets value b. The joint probability distribution can be estimated easily by evaluating the joint histogram ($\mathcal{H}$) of the two images and then dividing the number of occurrences of each entry by the total number of pixels. A joint histogram is a bi-dimensional histogram made up of $n \times n$ bins; the occurrence (a,b) is associated with the bin (i,j) where $i = \lfloor a/m \rfloor$ and $j = \lfloor b/m \rfloor$ and m is the width of the bin. We use a joint histogram of 256×256 bins.

We extend the approach proposed in [CDPS09]. We generate a rendering of the 3D model with some illumination related properties given the current camera parameters, we compute the image gradient of the rendering and the image gradient of the frame and then we evaluate the mutual information of these gradient maps (Figure 1). An iterative optimization algorithm updates the camera parameters and recalculates MI until the registration is achieved. The image gradient is computed by applying the Sobel operator to the images' CIE luminance.

For the rendering of the 3D model we combine the information provided by the ambient occlusion and the normal map, as suggested in [CDPS09]. The ambient occlusion is precalculated and stored in the 3D model as per-vertex color. During the rendering the value of ambient occlusion is interpolated by Gouraud shading among the triangle vertices. The final color C is obtained by weighting the normal map C_N with the value C_A of the ambient occlusion map (that is normalized between 0.0 and 1.0):

$$\begin{aligned} C_x &= (1-C_A)C_A + C_A C_{Nx} \\ C_y &= (1-C_A)C_A + C_A C_{Ny} \\ C_z &= \sqrt{1-(C_x^2+C_y^2)} \end{aligned} \qquad (3)$$

For the iterative optimization we use the algorithm NEWUOA. This algorithm iteratively minimizes a function $F(x)$, $x \in R^n$, by approximating it with a quadric Q. A trust region procedure adjusts the variables looking for the minimum of Q, while new values of the function improve the approximation.

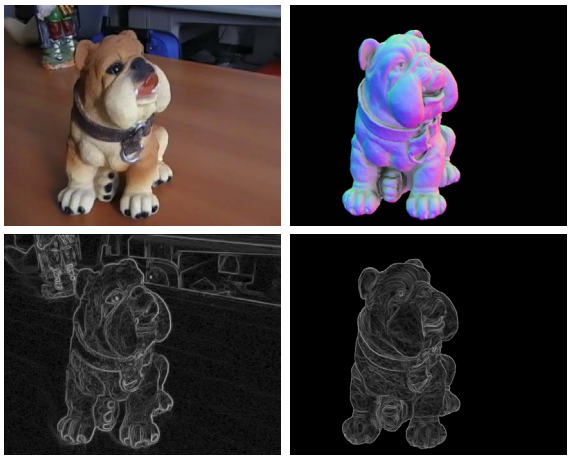


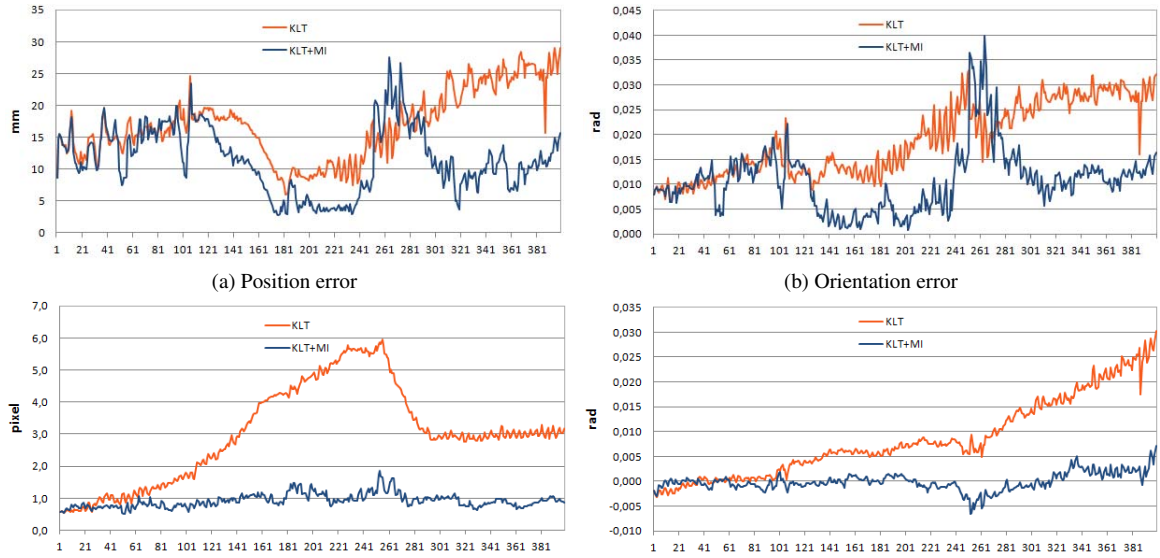
Figure 1: (Top-Left) Video frame. (Top-Right) Rendering of the 3D model with normal map and ambient occlusion. (Bottom-Left) Gradient map of the frame. (Bottom-Right) Gradient map of the rendering.

4. Results

In this section we present the results for two different types of input sequences: a synthetic video to evaluate the registration error and the effectiveness of the method, and a set of real video sequences of objects of different sizes.

4.1. Synthetic sequences

We prepared a synthetic video sequence of 400 frames with known camera parameters to evaluate the quality and the precision of the registration of the proposed method. We compared the camera estimated by our method and the camera estimated only with KLT tracking data. In this sequence we render a colored 3D model (200k faces) of a medium height (50 cm) statue of a shepherd in a complex lighting



(a) Position error (b) Orientation error

(c) Projection error (d) Roll angle error

Figure 2: Charts of the registration errors: KLT + MI registration (blu line); KLT registration (orange line).

environment, composed by an area light and an environment map, simulating a set of possible effects, like motion blur, jittering, noise and unstable lighting conditions, that characterize a real video sequence due to the environment, the camera characteristic and the type of camera motion.

For each frame we show in the Figure 2 the charts of four different types of misalignment measures of the cameras, which are estimated with our method (blue line) and with only the KLT tracking data (orange line), with respect to the real camera. The chart 2a shows the distance in millimeters of the position of the estimated camera from the real one. The chart 2b shows the angle of the quaternion which defines the rotation needed to align the orientation of the estimated camera with the real camera. The chart 2d shows the error in radiant of the roll angle of camera around the optical axis. The chart 2c shows the projection error, which is computed by projecting a set of points uniformly distributed over the surface of the object in image space and calculating the average distance between the image points obtained by the real camera and the image points obtained by the estimated camera. The graphs show that the estimation of the cameras with the proposed method is better and less sensitive to the drift problem with respect to the camera recovered only with the tracking data. This is particularly evident in the chart 2d. Another advantage of our method is the very low and stable projection error (chart 2c). The analysis of the charts 2a and 2b requires more attention, especially between the frames 250 and 280. In this interval our method recovers a camera position and orientation with a bigger error than the camera estimate with only the tracking data, but on the

Figure 4: Registration results obtained in the synthetic sequence with KLT (Left) and KLT+MI (Right): frame 80 (Top); frame 264 (Center); detail of the frame 264 (Bottom).

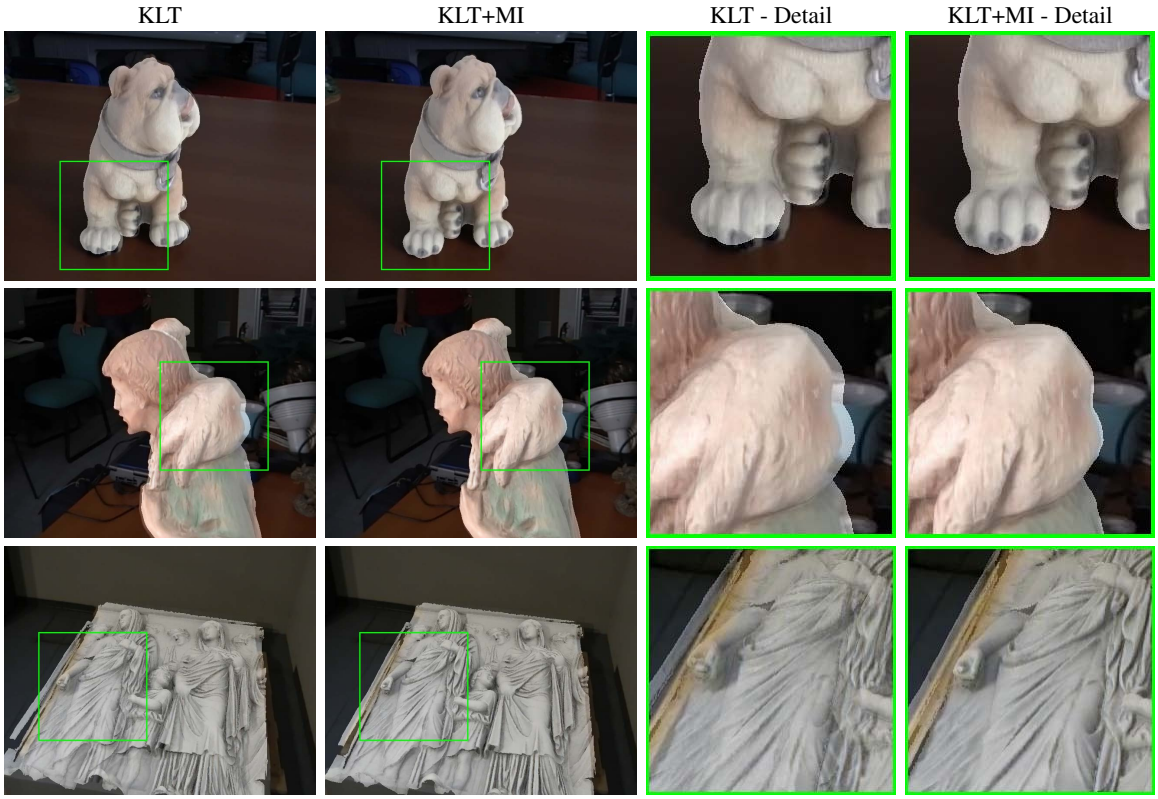


Figure 3: Comparison of the registration obtained in 3 different real sequences: Dog, frame 290 (Top); Shepherd, frame 400 (Center); Ara Pacis, frame 740 (Bottom)

other hand the projection error is lower. This behavior is due to the statistical nature of the registration by MI that in this case converges towards a camera which is quite far away in space from the real camera, but very similar from the point of view of the projection as we can see in the chart 2c and in the Figure 4.

4.2. Real-world sequences

We took four real sequences of different objects of known geometry acquired by 3D scanning: a dog's small statue (about 20 centimeters in height); a shepherd's statue (about 50 centimeters); a marble reproduction of an Ara-Pacis' bas-relief (about 2 meters); the Nettuno statue (about 6 meters) situated in the fountain on Piazza della Signoria in Florence. The sequences were acquired with a consumer video camera with standard PAL resolution of 720×576 pixels and using a constant zoom factor. In Figure 3, we show a visual comparison on a specific frame of the results obtained by the proposed registration algorithm and the results obtained using only the tracking data. A detail of the frame is shown to better visualize the misalignment. These results show the significant improvement introduced by the use of the MI.

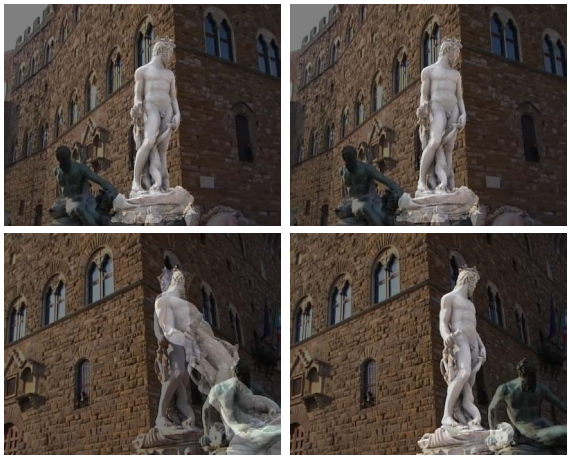


Figure 5: Results on Nettuno sequence obtained by KLT registration (Left) and KLT+MI registration (Right): frame 20, before the occlusion (Top); frame 200, after the occlusion (Bottom).

The results obtained in the sequence of the Nettuno statue are very interesting (Figure 5). In this sequence a major occlusion appears during the video. We don't apply any strategy to discard the features that appear on the occluder object during the occlusion. As we can see in Figure 5, using only the tracking data does not allow to estimate the camera due to the incorrect correspondences and the incremental working of the registration. Conversely, our algorithm permits to preserve a good alignment even if the final registration is not very precise. In the specific, our algorithm estimates an unstable camera during the occlusion, but in the subsequent frames it is able to recover a good registration. Conversely, using only the tracking data, we completely lose the registration. In this case a further improvement in the precision of the registration can be obtained implementing a strategy to automatically discard the features on the occluders, taking into account the camera motion and the error information returned by the algorithm for each 2D-3D correspondence.

For all sequences, we present in Table 1 some data about the length of the sequence, the 3D model used for the registration, the time required for the preprocessing of the video (deinterlace, denoise, removing of the lens distortion and tracking) and for the computation of the registration parameters, and, finally, on how many frames we apply the registration by MI. The tests have been executed on a Intel Core2 Quad Q9400 with 4GB of RAM and a NVIDIA GTX260 896MB. From the table we can note the highest preprocessing time in the sequence of Shepherd's statue and the highest registration time in the sequence of the Ara Pacis' bas-relief. The former is due to the high number of features to track in the video, the last is due to the alignment by MI that requires more iterations of the optimization algorithm NEWOUA to converge for each frames.

5. Conclusion and future work

We presented a new algorithm for the registration of a video sequence of a real object over its dense digital representation, taking advantage of the high frame-to-frame coherence. We put together the strong-points of two different alignment approaches: feature-based by KLT video tracking; statistical-based by maximizing the MI between the gradient map of the frames and the gradient map of the rendering of the 3D model with two illumination related properties, normals and ambient occlusion values. The registration by MI is able to correct the drift problem introduced by the KLT tracker in long and complex sequences, while KLT tracking speeds up the registration and controls the convergence of MI towards good camera parameters. We demonstrated the accuracy of the registration of our algorithm with respect to the KLT tracking on a synthetic sequence. Results are extremely positive, especially for the very low projection error. Then, we showed the results obtained on four different real video sequences of objects of different sizes.

The algorithm can be useful in the applications that use the bi-directional data transfer between the 3D model and the video, like color transfer, estimation of reflectance properties and recording of appearance-varying scenes.

As future work, the algorithm can be improved in three different aspects. The first is the improvement of the registration in the case of major occlusion, like in the sequence of the Nettuno statue. A possible solution could be to automatically delete the 2D features on the occluders taking into account several info, like the camera motion or the error info returned by the algorithm for each 2D-3D correspondence, or implementing a multi-step registration algorithm with several step of forward and backward registration. Another improvement is the GPU implementation of some portions of the algorithm, like the computation of the MI, in order to speed up the methods. The last improvement can be the possibility to make the entire algorithm completely automatic, removing the need of an initial manual alignment of the first frame.

6. Acknowledgment

We acknowledge the financial support of the EC IST IP project "3D-COFORM"(IST-2008-231809).

References

[BTG06] BAY H., TUYTELAARS T., GOOL L. J. V.: SURF: Speeded up robust features. In *ECCV* (2006), vol. 3951, pp. 404–417. 2

[CDPS09] CORSINI M., DELLEPIANE M., PONCHIO F., SCOPIGNO R.: Image-to-geometry registration: a mutual information method exploiting illumination-related geometric properties. *Computer Graphics Forum 28*, 7 (2009), 1755–1764. 2, 3, 4

[CVG04] CORNELIS K., VERBIEST F., GOOL L. J. V.: Drift detection and removal for sequential structure from motion algorithms. *IEEE Transactions on PAMI 26*, 10 (2004), 1249–1259. 2

[DM09] DAME A., MARCHAND É.: Optimal detection and tracking of feature points using mutual information. In *ICIP* (2009), pp. 3601–3604. 2

[FDG*05] FRANKEN T., DELLEPIANE M., GANOVELLI F., CIGNONI P., MONTANI C., SCOPIGNO R.: Minimizing user intervention in registering 2d images to 3d models. *The Visual Computer 21*, 8-10 (sep 2005), 619–628. 2, 3

[HS88] HARRIS C., STEPHENS M.: A combined corner and edge detector. In *Fourth Alvey Vision Conference* (1988), pp. 147–151. 2

[IOT*07] IKEUCHI K., OISHI T., TAKAMATSU J., SAGAWA R., NAKAZAWA A., KURAZUME R., NISHINO K., KAMAKURA M., OKAMOTO Y.: The great buddha

	Frames	Geometry (triangles)	Preprocessing (mm:ss)	Registration (mm:ss)	MI (N. Frames)
Dog	347	195k	5:58	3:43	27
Shepherd	837	200k	16:18	11:43	73
Ara Pacis	749	350k	11:19	16:06	49
Nettuno	360	400k	7:16	5:56	81

Table 1: Test data

project: Digitally archiving, restoring, and analyzing cultural heritage objects. *International Journal of Computer Vision 75*, 1 (Oct. 2007), 189–208. 2

[JFS01] JIN H. L., FAVARO P., SOATTO S.: Real-time feature tracking and outlier rejection with changes in illumination. In *ICCV* (2001), pp. I: 684–689. 2

[KB99] KATO H., BILLINGHURST M.: Marker tracking and hmd calibration for a video-based augmented reality conferencing system. In *IWAR* (1999), p. 85. 2

[LHS00] LENSCH H., HEIDRICH W., SEIDEL H.: Automated texture registration and stitching for real world models. In *PACIFIC GRAPHICS* (Oct. 3–5 2000), pp. 317–327. 2

[LK81] LUCAS B., KANADE T.: An iterative image registration technique with an application to stereo vision. In *DARPA Image Understanding Workshop* (1981), pp. 121–130. 2

[LL09] LI Y., LOW K.-L.: Automatic registration of color images to 3d geometry. In *CGI* (2009), pp. 21–28. 2

[Low99] LOWE D. G.: Object recognition from local scale-invariant features. In *ICCV* (Washington, DC, USA, 1999), IEEE Computer Society, pp. 1150–1157. 2

[LSY*06] LIU L., STAMOS I., YU G., WOLBERG G., ZOKAI S.: Multiview geometry for texture mapping 2d images onto 3d range data. In *CVPR* (2006), pp. 2293–2300. 2

[LWG97] LEVENTON M. E., WELLS W. M., GRIMSON W. E. L.: Multiple view 2D-3D mutual information registration. In *Image Understanding Workshop* (1997), pp. 625–630. 3

[MCV*97] MAES F., COLLIGNON A., VANDERMEULEN D., MARCHAL G., SUETENS P.: Multimodality image registration by maximization of mutual information. *IEEE Transactions of Medical Imaging 16*, 2 (Apr. 1997), 187–198. 3

[MS01] MIKOLAJCZYK K., SCHMID C.: Indexing based on scale invariant interest points. In *ICCV* (2001), pp. 525–531. 2

[NF05] NAIMARK L., FOXLIN E.: Encoded LED system for optical trackers. In *ISMAR* (2005), pp. 150–153. 2

[NK99] NEUGEBAUER P. J., KLEIN K.: Texturing 3D Models of Real World Objects from Multiple Unregistered Photographic Views. *Computer Graphics Forum 18*, 3 (Sept. 1999), 245–256. 2

[PGV*04] POLLEFEYS M., GOOL L. J. V., VERGAUWEN M., VERBIEST F., CORNELIS K., TOPS J., KOCH R.: Visual modeling with a hand-held camera. *International Journal of Computer Vision 59*, 3 (Sept. 2004), 207–232. 1

[PMV03] PLUIM J. P. W., MAINTZ J. B. A., VIERGEVER M. A.: Mutual information based registration of medical images: A survey. *IEEE Transactions of Medical Imaging 22*, 8 (2003), 986–1004. 3

[Pow08] POWELL M. J. D.: Developments of NEWUOA for minimization without derivatives. *IMA Journal of Numerical Analysis 28*, 4 (Oct. 2008), 649–664. 3

[ST94] SHI J., TOMASI C.: Good features to track. In *CVPR* (June 1994). 2

[TB09] THORMAEHLEN T., BROSZIO H.: Voodoo Camera Tracker. `http://www.digilab.uni-hannover.de/docs/manual.html`, 2009. 1, 3

[TCGP09] TA D. N., CHEN W. C., GELFAND N., PULLI K.: SURFTrac: Efficient tracking and continuous object recognition using local feature descriptors. In *CVPR* (2009), pp. 2937–2944. 2

[TG00] TUYTELAARS T., GOOL L. J. V.: Wide baseline stereo matching based on local, affinely invariant regions. In *BMVC* (2000), pp. 412–425. 2

[THWS08] THORMÄHLEN T., HASLER N., WAND M., SEIDEL H.-P.: Merging of feature tracks for camera motion estimation from video. In *CVMP* (2008). 2

[Tsa87] TSAI R. Y.: A versatile camera calibration technique for high-accuracy 3D machine vision metrology using off-the-shelf TV cameras and lenses. *IEEE Journal of Robotics and Automation 3* (1987), 323–344. 3

[VW97] VIOLA P. A., WELLS W. M.: Alignment by maximization of mutual information. *International Journal of Computer Vision 24*, 2 (Sept. 1997), 137–154. 3

[ZZCW07] ZHU G., ZHANG S., CHEN X., WANG C.: Efficient illumination insensitive object tracking by normalized gradient matching. *IEEE Signal Processing Letters 14*, 12 (Dec. 2007), 944–947. 2