Tracking with Reference Images: A Real-Time and Markerless Tracking Solution for Out-Door Augmented Reality Applications

Didier Stricker

Fraunhofer Institute for Computer Graphics
Rundeturmstraße 6, 64283 Darmstadt, Germany
stricker@igd.fhg.de

Abstract

This paper presents the optical tracking solution developed in the ArcheoGuide project (The Augmented Reality based Cultural Heritage On-site GUIDE, IST-1999-11306). The system enables to recover precisely the user head position and orientation at predetermined viewing location without the support of markers.

The tracking approach is novel and bases on the real-time registration of the live video-image with so called "reference images" of the site. Once the matching between a live- and a reference-images could have been established, the virtual information can be presented correctly to the user, either using a 2D image warping transformation or after deduction of the current 3D position/orientation of the camera.

The image matching algorithm represents the core of the tracking system. It must have real-time performance and be particulary robust to intensity and local changes. We opted for an analysis in frequency space and exploit the invariance properties of the Mellin-Fourier transform.

The tracking system has been tested outdoor in the context of the Archeoguide project. It runs on a laptop at around 10 to 15 Hz and provides views of virtual monuments in a Head Mounted Display superposed to the ruins of the archaeological site.

CR Categories: I.4.8 [Image Processing and Computer Vision]: Scene Analysis—Tracking I.4.3 [Image Processing and Computer Vision]: Enhancement—Registration I.4.7 [Image Processing and Computer Vision]: Feature Measurement—Invariant I.3.7 [Computer Graphics]: Three-Dimensional Graphics and Realism—Virtual Reality

Keywords: Markerless Optical Tracking, Mobile Augmented Reality, Cultural Heritage

1 Markerless Tracking for Augmented Reality

1.1 Necessity of a support

Markerless optical tracking is a very complex task. One can apply an image processing operator, like the Shi-Tomasi [12] or Harris [7] corner detector, to the video images and recover the camera trajectory out of the motion of the 2D image features. This approach works well off-line, because all features are accessible at a time, what enables filtering or minimization of the errors by e.g. bundle-adjustment techniques. On contrary for real-time applications, the motions are often abrupt and unpredicable, what makes sequential approaches very uncertain and fragile.

The need of a reference (or a support) for optical trackers appear to be absolutely necessary. Usually, these systems rely on markers [6, 8, 5]. They can be easily detected in the images and provides 3D metric information, which are used to deduct the current camera position/orientation. In our application - Augmented Reality on an archaeological site - markers can not be used: the considered scenes are too wide and the environment must keep unchanged. A solution consists in replacing the markers by another support provided by the scene it-self. We propose to use, not local scene features or a 3D model of the scene, but a set of standard images of the environment.

1.2 Tracking with reference images

The principle of the tracker is illustrated figure 1. At a given view point, a set of reference images is selected. The user view, *ie* the current live video image, is compared to all reference images and a correlation score is computed. The best score is retained and the 2D transformation between the current video-image and the reference image is evaluated.

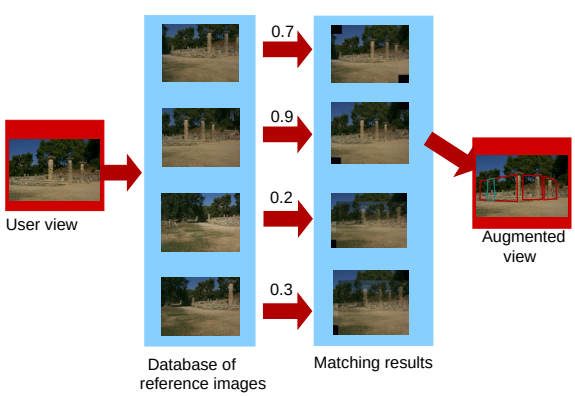


Figure 1: Tracking with reference images

At this step, two alternatives are possible:

1. The virtual information are directly warped in 2D onto the live video image with help of the transformation computed before.

2. The current camera position and orientation are deducted from the 2D image transformation and pass to the rendering system.

The success ot this approach resides in the image-registration technique, which must be particulary robust and fast. These techniques are shortly reviewed and discuss in the next paragraph.

1.3 Image registration techniques

It exists a lot of different approaches to image registration [2], which differ basically in:

1. the kind of the considered transformation - e.g. local transformation, global linear and non-linear transformation.

2. the data, which are used - e.g. corners [15], contours [9], pixel intensity [14].

3. the search space and the search strategie - e.g. graph search, linear bipartite assignment.

As mentioned before, the reliability of the algorithm represents the most important aspect of the choice of the registration technique. A lot of changes may appeared between the live and reference images. New objects or visitors may be present in the scene and new lightning conditions, due for example to another sun direction or clouds, will create new shadows or highlights in the images. In addition, the camera used is an USB-camera, which is a low-cost camera providing images of relatively poor quality.

We prefer therefore to orient our choice to algorithms that exploit rather global than local properties of the image. Basically, this corresponds to algorithms working directly on the pixel-intensity of the whole image or in frequency space.

In this work, we opted for a Fourier based approach. Those algorithms, like the phase-correlation, are well-known for their robustness and are often used to initialize iterative and more exact registration computations [14]. The problem is, that the Fourier transform enables to recover only a few transformation parameters, *ie* rotation, scale, and translation. The image registration bases then on a 2D planar similitude and will be exact only in very special cases. It follows, that the tracking will be valid only at pre-defined viewing-points and for restricted camera motions.

Nevertheless, acceptable results can be expected by taking advantage of the large size of the scene and the 3D/2D motion ambiguity. For a given reference image we consider that, (1) the 2D image shifts are due to a pure 3D rotation (2) the 3D translation and the perspective distorsion are neglectable, (3) a (moderate) scale change corresponds to a motion along the optical axis of the camera, and finally (4) the 2D rotation of the image is due to an 3D rotation of the camera around its optical axis.

2 Image Registration with the Fourier-Mellin Transform (FMT)

This section describes how to recover the 2D translation, the rotation and the scale factor between two images using the Fourier-Mellin transformation [3, 11, 4].

2.1 Fourier Transform

2.1.1 Translation

Let f_1 and f_2 be two images differing only in a 2D translation $t(t_x, t_y)$. The images are related as follows:

$$f_2(x, y) = f_1(x - t_x, y - t_y) \tag{1}$$

The Fourier functions F_1 and F_2 of the two images are given by the Fourier shift-theorem (see for example [1]):

$$F_2(\xi, \eta) = e^{-j2\pi(\xi t_x + \eta t_y)} F_1(\xi, \eta) \tag{2}$$

where F_2 and F_1 are two arrays of complex numbers.

The power spectrum of the Fourier transforms F_1 does not change if the function f_1 is shifted by an amount of (t_x, t_y). The power spectrum is translation invariant.

The translation vector (t_x, t_y) can be easily isolated by computing the cross-power-spectrum of F_1 and F_2:

$$\frac{F_1(\xi, \eta) F_2^*(\xi, \eta)}{|F_1(\xi, \eta) F_2^*(\xi, \eta)|} = e^{j2\pi(\xi t_x + \eta t_y)} \tag{3}$$

where: $F_2^*(\xi, \eta)$ is the conjugate-complex value of $F_2(\xi, \eta)$.

The inverse Fourier transformation (IFT) of an exponential function is a Dirac impulse. Therefore, the estimation of the maximum of the IFT of the equation (3) provides the image shift (t_x, t_y).

2.1.2 Rotation

If the image $f_1(x, y)$ is transformed into the image $f_2(x, y)$ by a translation $t(t_x, t_y)$ and a rotation with angle ϕ, then the relation between f_1 and f_2 is defined as:

$$f_2(x, y) = f_1(\quad x \cos\phi_0 + y \sin\phi_0 - t_x, \\ -x \sin\phi_0 * y \cos\phi_0 - t_y) \tag{4}$$

According to the shift theorem of the Fourier transformation, we obtain:

$$F_2(\xi, \eta) = e^{-j2\pi(\xi t_x + \eta t_y)} F_1(\quad \xi \cos\phi_0 + \eta \sin\phi_0, \\ -\xi \sin\phi_0 + \eta \cos\phi_0) \tag{5}$$

The power spectra of F_1 and F_2 are related as follows:

$$\|F_2(\xi, \eta)\| = \quad \|F_1(\quad \xi \cos\phi_0 + \eta \sin\phi_0, \\ -\xi \sin\phi_0 + \eta \cos\phi_0)\| \tag{6}$$

$\|F_1\|$ and $\|F_2\|$ undergo the same rotation as the images.

2.2 Scale

Let $f_2(x, y)$ be the scaled replica of the image $f_1(x, y)$ with the scale factors (a, b), so that:

$$f_2(x, y) = f_1(ax, by) \tag{7}$$

The Fourier scale property is defined as follows:

$$F_2(\xi, \eta) = \frac{1}{|ab|} F_1\left(\frac{\xi}{a}, \frac{\eta}{b}\right) \tag{8}$$

The equation (8) shows that the scaling of the images f_1 and f_2 by the factors (a, b) scales the spectra magnitudes by $(\frac{1}{a}, \frac{1}{b})$.

In conclusion, the power spectra are invariant for translation but variant for scaling and rotation.

2.3 Fourier-Mellin Transform

For a rotated and scaled image (with $a = b$), the Fourier transform can be written as follows:

$$F_2(\xi, \eta) = \quad \tfrac{1}{a^2} F_1(\tfrac{1}{a}(\xi \cos\phi_0 + \eta \sin\phi_0), \\ \tfrac{1}{a}(-\xi \sin\phi_0 + \eta \cos\phi_0)) \tag{9}$$

$$\tag{10}$$

With an adequate change of coordinate system, from Cartesian (x, y) to polar (r, ϕ), the rotation can be decoupled from the scaling.

The following varaibles are introduced:

$$r = \sqrt{\xi^2 + \eta^2}$$

and:

$$\phi = atan(\frac{\eta}{\xi})$$

Equation (9) becomes:

$$F_{2p}(\phi, r) = \frac{1}{a^2} F_{1p}(\phi - \phi_0, r/a)$$

An image rotation corresponds to a shift along the angular axis (ϕ) of the function $F_{2p}(\phi, r)$. The scaling influences the values of the function by a factor (a^{-2}) and scales down the radial coordinates by the factor a.

Using a logarithmic scale for the radial coordinates, the scaling can be reduced to a translation.

We pose:

$$F_{2lp}(\phi, \rho) = \frac{1}{a^2} F_{1lp}(\phi - \phi_0, \rho - A)$$

with: $\rho = log(r)$ and $A = log(a)$

In logarithmic-polar coordinates, the scaling is expressed as a translation.

By transforming the log-polar of the magnitude of F_{1lp} and F_{2lp} as proposed in paragraph 2.1.1, the shift (ϕ_0, A) can be determined. This transformation is known as Mellin-Fourier transform and is translation, rotation and scale invariant.

In summary, the following computation steps are required:

1. Compute the Fourier-transformations of the images
2. Do a log-polar transformation of the spectrum.
3. Apply the phase-correlation method and recover the rotation angle ϕ_0 and the scale factor a.
4. Rectify the image and compute the translation by phase-correlation (see paragraph 2.1.1).

3 Implementation and Evaluations

3.1 Implementation: the Fast Fourier Transform (FFT)

The computations of the image registration are done on basis of the Fast Fourier Transform (FFT)[10]. Thereby, the image must be square and with dimension 2^n. In our implementation, the left and right borders are cut out and the image is scaled down to the next 2^n dimension.

Because the Fourier transform assumes a periodic function and the image is truncated, it is crucial to apply a window-function, like the Hanning window, to the input images. Another implementation difficulty consists of the numerical instability for coordinates near to the origin, since we have: $\lim_{r\to 0} \rho = \lim_{r\to 0} ln(r) = -\infty$. Therefore, a high-pass filter is apply on the logarithmic spectra. As in [11], we used a filter with the following transfer function:

$$H(x, y) = (1.0 - cos(\pi x)cos(\pi y))(2.0 - cos(\pi x)cos(\pi y)) \quad (11)$$

With: $-0.5 \leq x, y \leq 0.5$

An alternative to the high-pass filtering is to directly set to vlue zero the points near to the origin and inside a circle of ray ϵ [3].

3.2 Evaluation

In order to evaluate the precision, the robustness to noise, and the minimal image overlap of the implemented algorithm, we generate image pairs with given transformations (ground-truth). The initial and a transformed image are cropped in the center of a same large image. The pixel intensity and the noise level are the same and the computation should provide exact results. The size of the original images is 256×256 pixels but as for the real-time implementation, they are first internally scaled down to 128×128 pixels using bilinear interpolation.

3.2.1 Precision and resistance to noise

Two experiments are presented here. First, the image is rotated around its center with rotation angles varying between 0 and 90 degrees. A Gaussian noise with different variance ($\sigma = 0, 100, 200, 300, 400, 500$) is added to the original images. Then, the rotation is recovered with the help of the Fourier-Mellin transform and the absolute error between the true and the estimated angle is deducted. The results are shown figure 2. The dash line represents the angle error with no additive noise, whereby the dot-dash line represents the errors with maximal noise level.

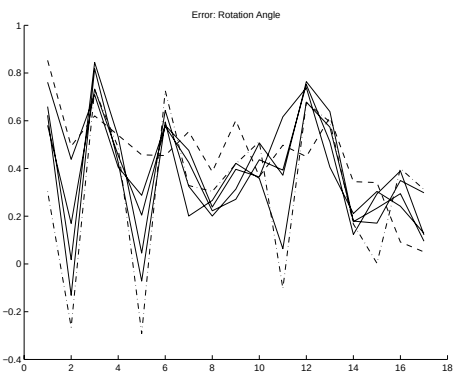


Figure 2: Absolute error of the rotation angle

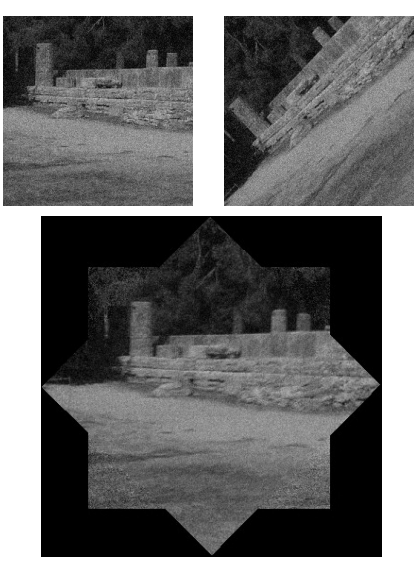
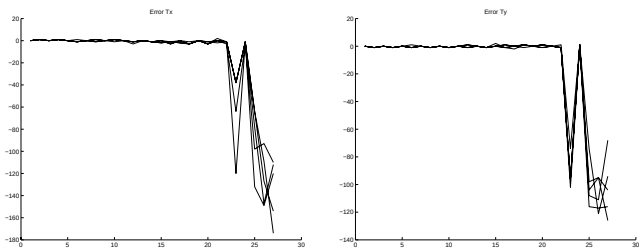


Figure 3: Image registration with noisy image and rotation angle of 45 degrees

It can be noticed that the rotation angle is computed with a satisfying precision, the maximal error beeing 0.9 degree. The other parameters, *ie* the translation and the scale stay also near to their true values. A maximal deviation of 2 pixels have been detected for the translation and a scale change of maximal 0.98 appeared. The experiment shows that the noise has a minimal influence on the registration results and that the rotation angle can be recovered with a

good precision even for high values, like 70 or 80 degrees. An example of an image registration with additive noise is given figure (3).

The second experiment is similar to the first one, but has for goal to estimate the robustness of the computation of the image translation and the minimal necessary image overlap. We let vary the translation from 0 to 130 pixels and add at each step Gaussian noise, as previously.



(a) (b)

Figure 4: Error for the translation component tx (a) and ty (b)

Here again, a good precision is achieved, independently from the noise level and even for small image overlaps ($< 50\%$). The absolute error is inferior to 2 pixels as long as the registration is successful. The minimal necessary overlap represents 30.5% of the image size. The registration result is shown figure (5).

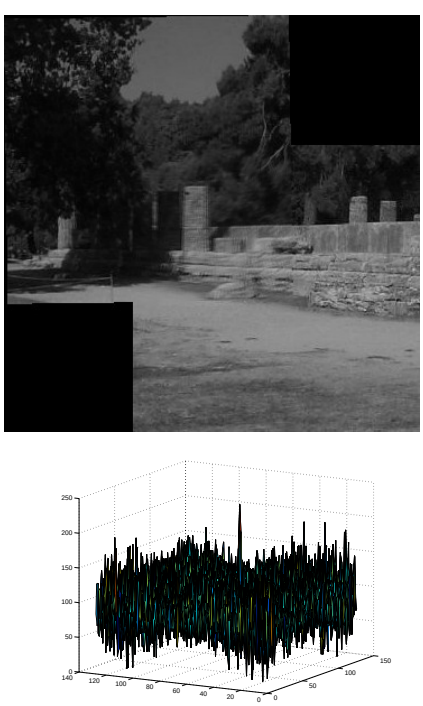


Figure 5: Registration with a minimal overlap (30.5%) and corresponding IFFT of equation (3).

The registration result is correct but the peak of the Dirac impulse can not be reliably distinguished from the other values. In pratice, such results would be preferably rejected in order to avoid instabilities.

3.2.2 Estimation of the registration validity

The validity of the registration is estimated at the end of the registration process, *ie* by the computation of the image shift (see paragraph 2.1.1). We consider that the Dirac impulse is well-defined if its magnitude is significantly superior to the mean over all values. The following condition is defined:

$$\|\delta(x,y)\| > \overline{m(x,y)} + a.\sigma(x,y)$$

Where: $\delta(x,y)$, $\overline{m(x,y)}$, and $\sigma(x,y)$ represent respectively the Dirac impulse, the mean value, and the variance of the IFFT of the exponential signal of equation (2). a is a scale factor determined on basis of the experiments presented in the previous paragraph.

3.2.3 Real-time performance

The last step of the development consists of the real-time implementation of the algorithm.

The first optimization consists of pre-computing the Fourier transformations of the reference images, so that the processing is limited to the live video image. Nevertheless, three FFT and a log-polar transformation are necessary. Additionally, the video image must be rectify after the computation of the rotation angle and scale factor.

Currently, if only one image is considered, the tracking reaches a frame-rate of 10 Hz on a 800 MHz PC. The frame-rate is constant, since the number of operations is fix. This would not be the case for intensity-based registration algorithms, which used iterative non-linear optimization routine, or for feature-based approaches, for which the frame-rate depends on the number of detected features.

4 Application: ArcheoGuide

The tracking system has been originally developed in the project *ArcheoGuide* (The Augmented Reality based Cultural Heritage On-site GUIDE). The ArcheoGuide-system consists of a mobile Augmented Reality unit that allows the visitors to see a computer-generated reconstruction of monuments without cutting them off from the real surroundings of the site. The project has for goal to explore new ways to learn about a cultural site.

Arriving at a "view point"(marked on the ground), the visitors wear on a Head Mounted Display and contemplate views of the virtual monuments on theirs ruins [13]. As a first trial site, the ancient Olympia in Greece, the birthplace of the Olympic games has been selected.

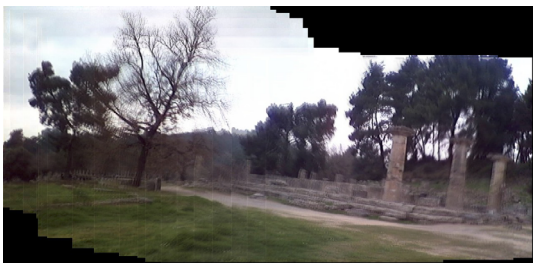
The complete system runs on a laptop PC (PIII 800 MHz, GeForce graphics card) at 10 Hz with an image resolution of 320x240 pixels (see figure 6(b)).

On-site tests demonstrated the validity and practicability of the tracking approach. Because the registration algorithm is robust to noise, light changes (since the analysis occurs in frequency space) and requires only small image overlaps, another day-time or visitors entering the scene didn't disturb the tracking.

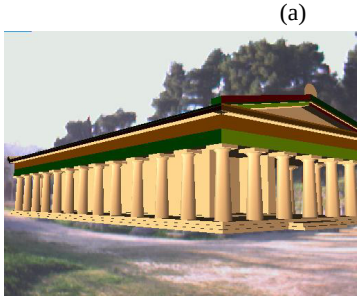
5 Conclusion

In this paper, we point out the necessity to provide a support to optical tracking-systems. Usually, this support is given by the markers and has been replaced successfully with images of the environment.

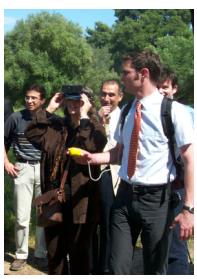
The presented tracking system relies on a fast and robust registration algorithm, which bases on the Fourier-Mellin transformation. A simulation with real images and known image transformations provided a good evaluation of the precision and the robustness of



(a)



(b)



(c)

Figure 6: (a) real-time motion estimation (hand-held camera) (b) Augmented-View of the Hera Temple (c) Trials on-site

the current implementation. The tracking system has been demonstrated outdoor in the frame of a mobile Augmented Reality application for archaeological sites. At given view-points, the head motion could have been recovered correctly, what enabled the insertion of virtual objects in 2D into a head mounted display.

Future works will concentrate on the extension of the working area of the tracker. Currently, for performance reasons, only one reference image is considered at a time. With help of a fast image retrieval module, the most appropriate image could be selected, and so allows different viewing directions and larger viewing areas.

Acknowledgments

This work is funded by the EU (Project Archeoguide, IST-1999-11306). The authors would like to thank all the consortium partners Intracom S.A. (Greece), the Computer Graphics Center (ZGDV) (Germany), the Centro de Computao Grafica (CCG) (Portugal), A&C 2000 Srl (Italy), Post Reality S.A. (Greece) and the Hellenic Ministry of Culture (Greece).

References

[1] Ronald N. Bracewell. *The Fourier Transform and its Applications*. McGraw-Hill, New York, 1986.

[2] Lisa Gottesfeld Brown. A survey of image registration techniques. *ACM Computing Surveys*, 24(4):325–376, December 1992.

[3] David Casasent and Demetri Psaltis. Position-, rotation-, and scale-invariant optical correlation. *Applied Optics*, 15(7):1795–1799, July 1976.

[4] Q.S. Chen, M. Defrise, and F. Deconinck. Symmetrical phase-only matched filtering of fourier-mellin transforms for image registration and recognition. *PAMI*, 16(12):1156–1168, December 1994.

[5] Youngkwan Cho, Jun Park, and Ulrich Neumann. Fast color fiducial detection and dynamic workspace extension in video see-through self-tracking augmented reality. *Proceedings of the Fifth Pacific Conference on Computer Graphics and Applications*, 1997.

[6] Stricker D., Klinker G., and Reiners D. A fast and robust line-based optical tracker for augmented reality applications. In *First International Workshop on Augmented Reality*. Springer Verlag, 1998.

[7] C. Harris and M.J. Stephens. A combined corner and edge detector. In *Alvey88*, pages 147–152, 1988.

[8] H. Kato and M. Billinghurst. Marker tracking and hmd calibration for a video-based augmented reality conferenencing system, 1999.

[9] H. Li, B.S. Manjunath, and S.K. Mitra. A contour-based approach to multisensor image registration. *IP*, 4(3):320–334, March 1995.

[10] W.H. Press, S.A. Teukolsky, W.T. Vetterling, and B.P. Flannery. *Numerical Recipes in C: The Art of Scientific Computing*. Cambridge University Press, Cambridge, UK, 2 edition, 1992.

[11] B.S. Reddy and B.N. Chatterji. An fft-based technique for translation, rotation, and scale-invariant image registration. *IP*, 5(8):1266–1271, August 1996.

[12] J. Shi and C. Tomasi. Good features to track. In *Proceedings of the Conference on Computer Vision and Pattern Recognition*, pages 593–600, Los Alamitos, CA, USA, June 1994. IEEE Computer Society Press.

[13] D. Stricker, P. Daehne, F. Seibert, I.T. Christou, L. Almeida, R. Carlucci, and N.I. Ioannidis. Design and development issues for archeoguide: An augmented reality based cultural heritage on-site guide (icav3d'01). In *International Conference on Augmented, Virtual Environments and Three-Dimensional Imaging*, Mykonos, Greece, May 28-30 2001. IEEE.

[14] Richard Szeliski and Heung-Yeung Shum. Creating full view panoramic mosaics and environment maps. In Turner Whitted, editor, *SIGGRAPH 97 Conference Proceedings*, Annual Conference Series, pages 251–258. ACM SIGGRAPH, Addison Wesley, August 1997. ISBN 0-89791-896-7.

[15] I. Zoghiami, O.P. Faugeras, and R. Deriche. Using geometric corners to build a 2d mosaic from a set of images. In *CVPR97*, pages 420–425, 1997.

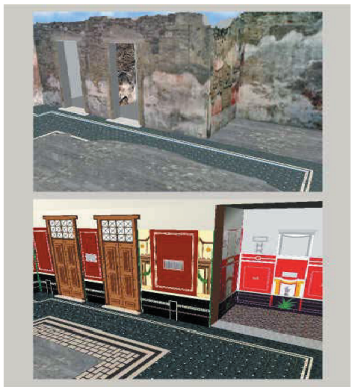
real-time motion estimation (hand-held camera)

Augmented-View of the Hera Temple

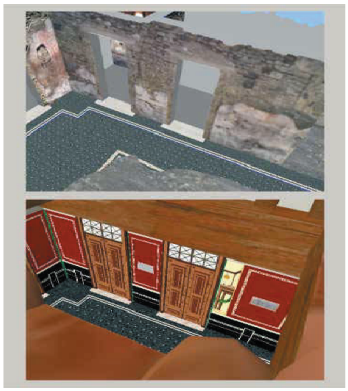
Trials on-site

Stricker: **Tracking with Reference Images: A Real-Time and Markerless Tracking Solution for Out-Door Augmented Reality Applications,** pp. 77-82.

Atrium, view from tablinum

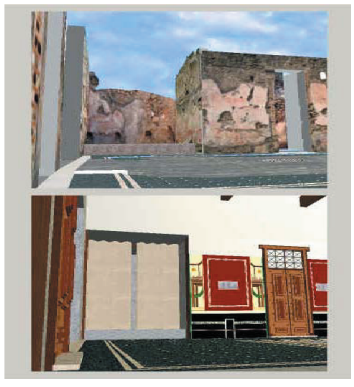


Atrium, Eastern side, view from the roof (compluvium)

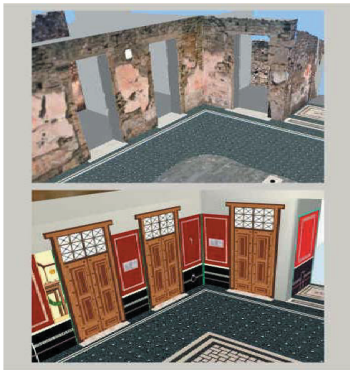


Atrium, Eastern side, view from South-Western corner

Atrium, South-Western corner, view from impluvium, Northern side



Atrium, South-Western corner, view from Eastern ala



Atrium, -North-Western corner, view from Eastern ala

Scagliarini, Coralini, Vecchietti, Cinotti, Roffia, Galasso, Malavasi, Pigozzi, Romagnoli, Sforza: **Exciting understanding in Pompeii through on-site parallel interaction with dual time virtual models,** pp. 83-90.