RES: computing the interactions between real and virtual objects in video sequences

Pierre Jancène, Fabrice Neyret, Xavier Provot,
Jean-Philippe Tarel, Jean-Marc Vézien,
Christophe Meilhac, Anne Verroust
INRIA
Domaine de Voluceau, ROCQUENCOURT
78153 Le Chesnay cedex FRANCE
Contact E-mail: Anne.Verroust@inria.fr
http://www-rocq.inria.fr/syntim/analyse/video-eng.html

Abstract

Possibilities for dynamic interactions of people with machines created by combination of virtual reality and communication networking provide new interesting problems at the intersection of two domains (among others): Computer Vision and Computer Graphics.

In this paper, a technical solution to one of these problems is presented to automate the mixing of real and synthetic objects in a same animated video sequence. Current approaches usually involve mainly 2D-based effects and rely heavily on human expertise and interaction. We aim at achieving a close binding between 3D-based analysis and synthesis techniques to compute the interaction between a real scene captured in a sequence of calibrated images, and a computer-generated environment.

1 State of the art

Augmented Reality defines the process by which natural images can be enhanced by superimposition of additional, usually computer-generated, visual inputs. The actual mixing process can be performed by a computer screen or other kind of optical systems. Augmented reality is of a great interest in various areas such as industry [FMS93, CM92], interior design [AKB+95, WCB+95] and medical applications [BFO92]. In most cases, the main problem is to ensure the accuracy of the super-imposition between the real environment and the virtual one.

When mixing 3D synthetic and real objects in a same animated sequence for video applications, one must insure the light and shadow consistency as well as the detection of the parts of the virtual objects hidden by the real ones in the resulting images. Several approaches propose solutions for this problem:

- Usual applications for special effects in video production simply add a synthetic image upon a real image (a technique known as *alpha keying*). If the synthetic image has to be partly occulted, a 2D mask has to be provided by the user, which forbids animated sequences or requires to interpolate the mask evolution along the time.
- Synthetic TV [Fel94], ensures the geometric and kinematic coherence between real images and synthetic ones. This is done by computing the correct position of virtual objects with respect to a real scene, and the correct modeling of a virtual camera with respect to a real one as well.
- Computer Augmented Reality [FGR93, Fou94] studies how to compute the common illumination information between a real and a computer generated scene. An application of this technique is the study of the visual effect of adding a virtual light source to the environment.

Currently all these methods rely heavily on the user's expertise and his ability to correct the mixing process on-line. Our goal is thus to propose a framework in which the mixing process can be automated as much as possible, and to demonstrate our approach on simple test cases.

2 Our approach

The purpose of *Reality Enriched by Synthesis* (RES) is to insert synthetic 3D objects in a 2D sequence of real images, while respecting as much as possible the 3D coherence for occultations, shadows, and collisions. The key idea is to recognize and locate real 3D objects in the 2D images in order to build a kind of 3D mask, which then is used for composition with virtual environments to produce the final enriched image sequence.

More precisely, our approach associates image analysis and synthesis techniques through the construction of compatible *synthesizable* 3D models, allowing a 3D-coherent mixing of the synthetic objects in the 2D images.

A synthesizable model describes a scene composed of objects, light sources and cameras, in such a way that an image of the scene can be generated from any given viewpoint [1] (one assumes that the resulting image has to make sense to the viewer, i.e. is to some extent realistic). Objects have a shape, a position, an orientation, surface properties (e.g. ambient, diffuse and specular colors), and possibly a temporal evolution (rigid motion or deformations). Practically speaking, an interactive tool called *modeler* is used to describe and manipulate synthesizable 3D models. Our lab has developed such a modeler, ACTION-3D, in conjunction with image analysis tools to compute actual relationships between images and models.

Possible applications of the RES technique can be found in a wide range of applications such as visualization (e.g. medical imagery, architecture, design), and video production (movies special effects, advertising, impact studies ...).
The RES process is achieved in batch mode, using the following sequence of actions (see block diagram of figure 1):

[1] For technical reasons, the set of valid observer locations is of course a restricted subset of all the possible ones.

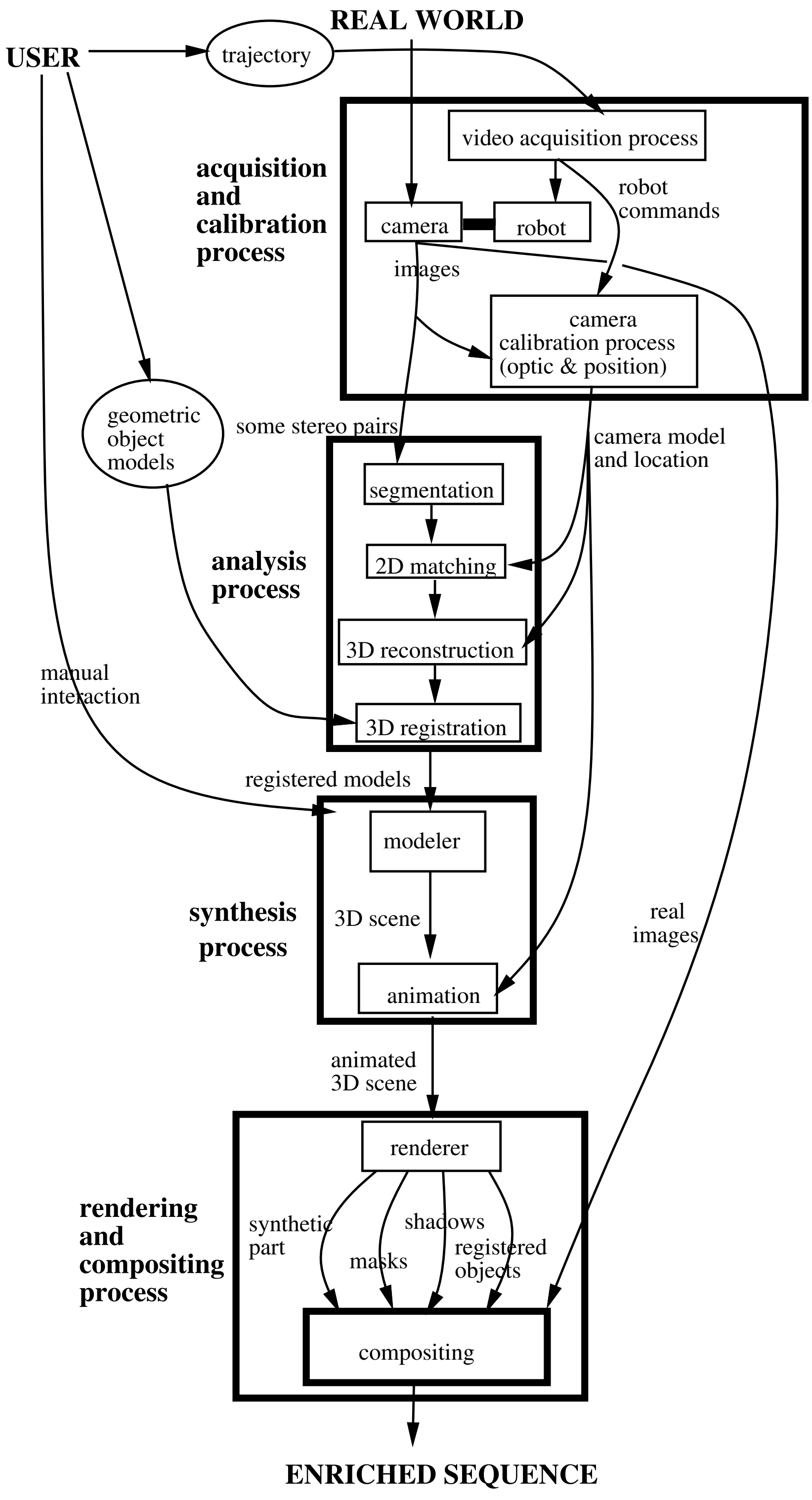


Figure 1: The RES process.

- **Image acquisition and camera calibration process**

  The real sequence is first acquired and digitized using our camera installation according to a user-defined trajectory (cf. section 2.1). Motions and optical characteristics of the camera during the whole sequence are computed from visual inputs for feedback purposes. Other camera characteristics, such as the field of view, are also stored.

- **Analysis process**

  Pairs of key images are then selected in the sequence and Computer Vision stereo algorithms process them to generate a 3D reconstruction of the observed scene. A Computer Graphics database of 3D models is then used to register complete object shapes in the real scene (cf. section 2.2). One so obtains a "synthetic" scene, consisting of a set of existing objects which all have the right location, size and orientation with respect to the original observer.

- **Synthesis process**

  The user generates his own animated scene using ACTION-3D, or retrieves it from existing databases. The "real" scene (composed of the models registered by the image analysis) serves as a starting point, around which all virtual objects are added incrementally (cf. section 2.3).

- **Compositing process**

  For each image of the sequence, the image production manager reads the geometric database containing both recognized and synthetic objects, calls dynamic modules which complete the scene (e.g. if dynamic deformable shapes are present), and finally calls the public domain renderer RAYSHADE to obtain four partial images, that will be mixed with the original one using a compositing module (cf. section 2.4).

The process described above is not totally automatic: as it is shown in figure 1, the user interacts:

- during the acquisition process, to determine the trajectory of the camera and the illumination characteristics of the real scene, as well as the position of the objects. This choice will be crucial for the rest of the analysis process as the occultations and illumination conditions must be satisfactory for at least two key images for each real object to be registered.
- during the analysis process, the key images corresponding to the real objects to be registered are manually selected. In fact, several real objects may be associated to the same two key images.
- during the synthesis process, as the synthetic scene is interactively modeled by the user (the perspective transformations corresponding to the real camera are given to the user so that he can see the images of the synthetic objects from the original viewpoint of the real sequence).

Image sequences presented in section 3 were generated using the above scheme, with the insurance that 3D coherence is satisfied (concerning occultations, shadows and collisions).

2.1 Acquisition and camera calibration process

The video sequence is obtained using a color camera Sony XC007 with a Canon J15x9.5B zoom lens. The images are digitized from the CCD camera using a S2200 color frame grabber which provides high quality images in video format.

The camera is mounted on a motorized three degrees of freedom robotic system which controls:

- the zoom factor and the aperture of the diaphragm
- the movements of the camera (a translation along the main frame axis and two axes of rotation: see figure 2).

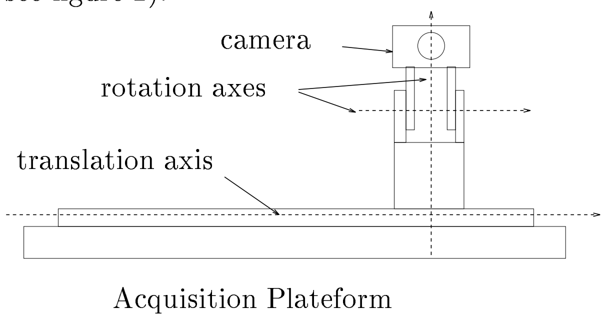


Figure 2: The camera and the robotic frame used in our experiments.

As the acquisition system can be controlled with an excellent precision, the values of the rotation and of the translation movements of the robot are known. A kinematic calibration of the frame [HD95] is necessary in order to obtain these informations. Up to now, only the translation motions are calibrated.

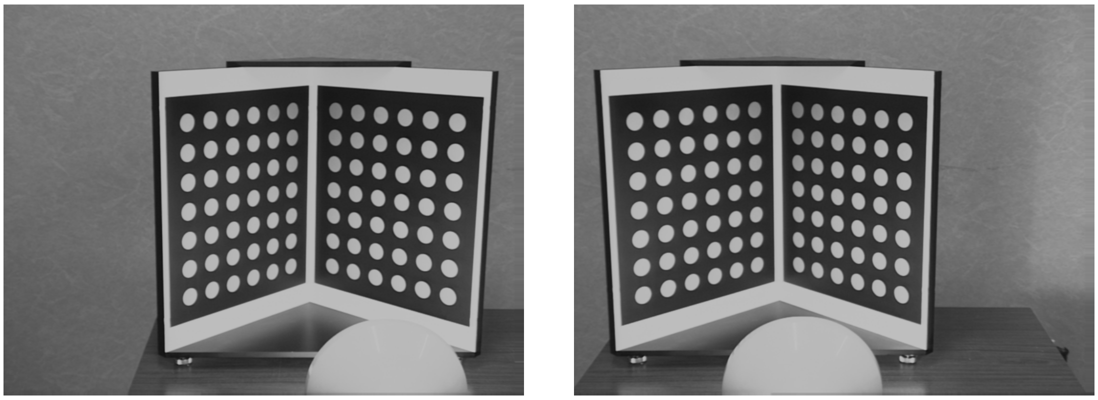
The visual calibration process [TG95] uses a selected set of images of a reference object and the corresponding set of robot commands controlling the movements of the camera to compute, for each real image of the whole sequence, the positions and the optical characteristics of the camera [Wil94].

2.2 Analysis stages

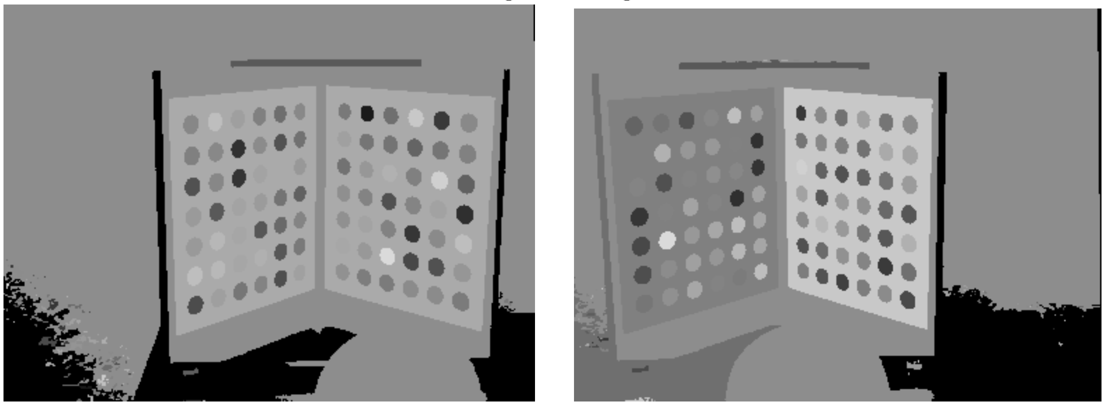
The analysis process must produce as complete and accurate geometric and photometric representations of the observed scene as possible, to allow the composition with synthetic objects in a realistic way. Currently, objects position, camera model, camera moves, lights position and modeling are estimated. No real-time constraint is imposed for the time being, as accurate image analysis techniques are still usually computationally expensive.

The analysis process stages used to generate a compatible 3D representation of the world are the following:

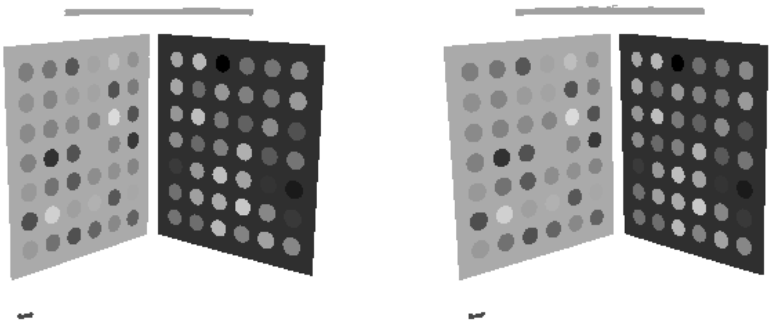
1. Pairs of key images are interactively chosen to construct stereoscopic pairs. Each image of the stereo pair is segmented into homogeneous regions using various algorithms such as region growing [SVCG89], region splitting [RG91] or energy minimization [AMG93] (see figure 8b and figure 3).



Original images



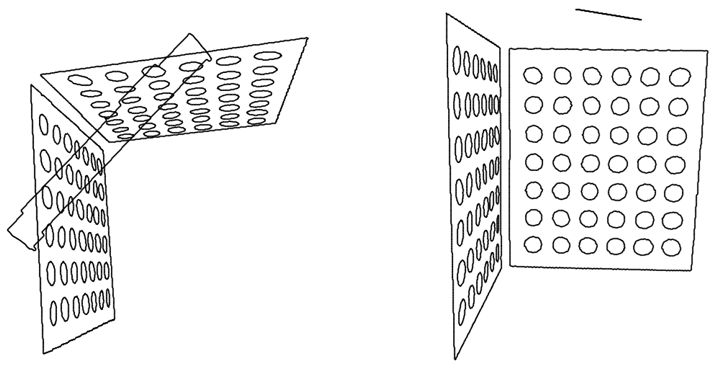
Segmented pair of images



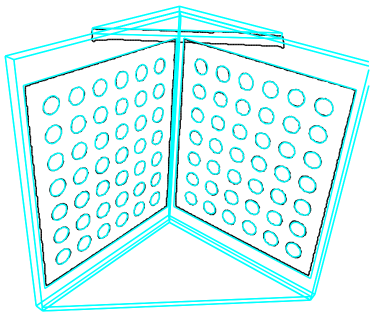
Matched pair of images

Figure 3: The 2D-steps of analysis process.

2. Segmented regions are then matched (one-to-one mapping between images), using epipolar constraints and image intensity-based characteristics [RG91]. At this point, one obtains couples of regions each of which (hopefully) correspond to the projections of a given 3D shape of the original scene (see figure 3).
3. 3D planar patches are reconstructed from their image projections using both geometric and photometric cues [TV95] (see figure 3).
4. Finally 3D reconstruction outputs are interpreted geometrically using prior models generated with the ACTION-3D interactive graphic modeler and stored as databases. "Interpreted" here means that the location of the models of objects known to be in the scene is accurately recovered, i.e. the parameters of the rigid displacement between each object model and its partial reconstruction obtained through image analysis are computed. Currently the algorithms used are adapted from the Iterative Closest Point approach [BM92, Véz95] (see Figures 8c and 4).



3D- reconstruction



Object registred

Figure 4: The 3D-steps of analysis process.

Simplifying assumptions are made during the analysis process which constrain the possible choices in the observed scene as well as in the image sequences:

- **Camera**: camera moves between key images are linear, to allow direct interpolation of the motion within the whole sequence. Variations in focal length (zoom factor) are excluded for now.

- **Scene content**: we chose to extract region features, thus objects actually involved in interactions with the virtual scene must have a minimum of 3 planar surfaces of approximately uniform color to allow a correct registration. Region features were preferred because they convey richer and denser information about the image content, thus allowing a more accurate scene interpretation.
- **Light**: lights are assumed to be far away from the observed scene, to produce constant or linear intensity gradient on the imaged surfaces.

Theses constraints are severe on the scene as well as on the sequence choice, but future work will focus on the improvement of our analysis process, by using additional computer vision techniques. The purpose of this preliminary work is mainly to demonstrate the possibility of a quite automatic mixing of real sequences with synthetic ones by using existing computer vision techniques developed by our team.

2.3 Synthesis

Rigid or geometrically deformed synthetic objects are modeled using the ACTION 3D modeler.
The modeler is able to :

- interactively build polyhedral objects and free-form surfaces,
- deform and animate these objects using our 3D deformation and animation modules, i.e. animated free form deformations (AFFD), axial deformations and 3D morphing,
- interactively design textures and their mapping upon object surfaces.

Dynamic objects (e.g. deformed at run-time using physics-based models) are introduced at rendering time by external modules. The dynamic module used for synthetic cloth animation [Pro95] is a mass-spring model which takes into account interactions such as contacts and collisions with other objects.

At this stage the camera model and and trajectory are integrated in the modeler so that the user in charge of animation can see the virtual objects from the point of view of the real camera during the whole sequence.

At the end of this process, a sequence of shapes and 3D positions of the synthetic objects to be inserted in the real video sequence and the corresponding perspective transformations of the camera are obtained. It is what we called the "animated 3D scene" in figure 1.

2.4 Compositing

The inlaying of a synthetic image upon a real one (see figures 5 and 6) is achieved using a classical mask-based approach: each pixel of each frame of the video sequence is modified by combining it with the corresponding pixel of a synthetic image. The process is of course achieved with sub-pixel resolution to avoid aliasing effects.

Coherent occultation of the real objects by the synthetic ones is thus automatically achieved. But what about the symmetric case ? Real objects occulting virtual ones have to mask them in the final image, but should *not* be present in the synthetic image, for they are already in the original one. This effect is taken into account by

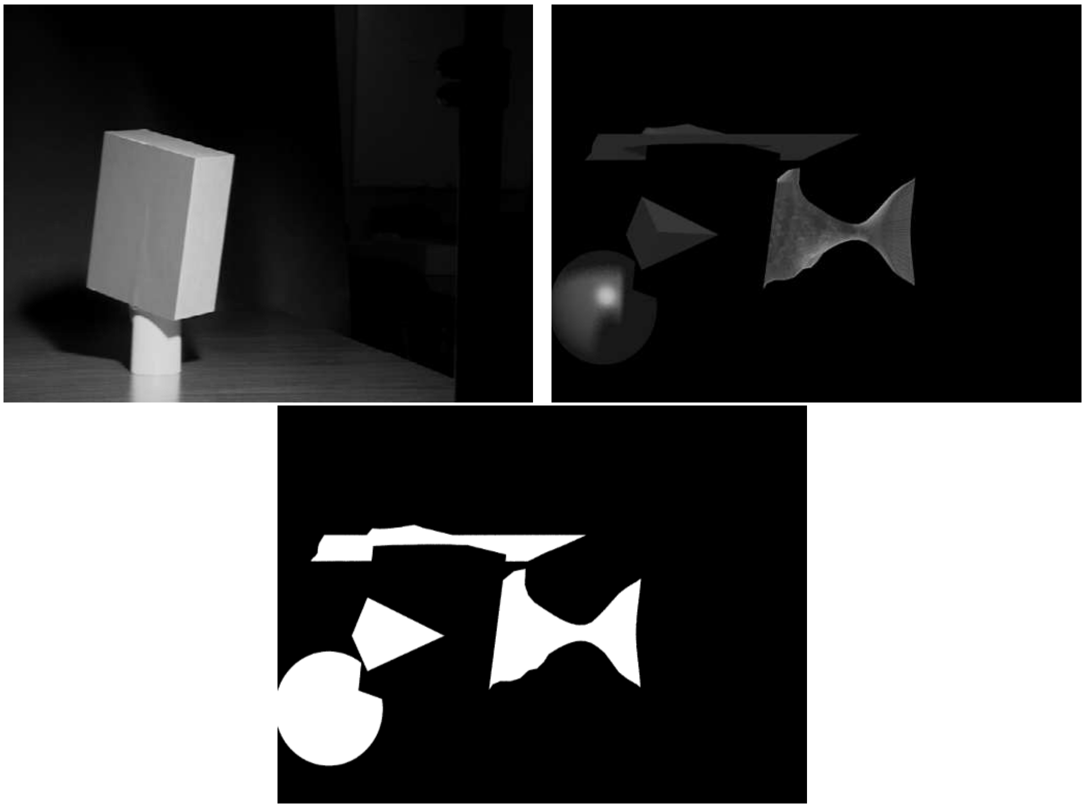


Figure 5: The initial image, the synthetic one and the mask. Note the mark consisting of the registered objects (such as the box), invisible but producing shadows and occultations.

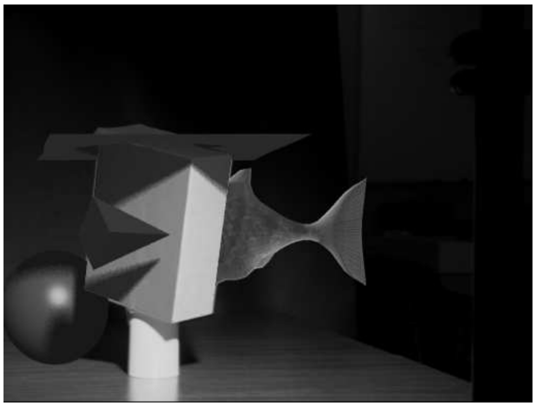


Figure 6: The enriched image.

incorporating the 3D models of the real scene (as computed by the analysis process described above) into the synthetic one, but painting them in black, so that their contribution consists only in occulting parts of the added image. Actually, it is worth noting that shadows of real objects are also cast on the synthetic ones by this process, so this effect is taken care of as well. Also, one can change the reflective properties of the "black" real objects by adding a reflective component, hence creating virtual reflections (see figure 8).

Finally, one has to compute the shadows (and more generally the shading effects) produced by the virtual objects on the real ones. We approximate them by computing an attenuating factor (number between 0 and 1) to multiply to each pixel of the original image. This factor is evaluated by rendering the recognized objects in purely lambertian white, first with and then without the virtual objects. The attenuation factor is thus defined as the ratio of the second image (with shadows) and the first one (the reference image, containing only the real scene).

Finally, the mixing formula defining the compositing phase is:

$$\text{enriched image} = \text{real image} \times \text{attenuation} \times \text{(1-mask)} + \text{synthetic.}$$

One should note that the synthetic image contains:

$$\text{synthetic objects} \times \text{mask} + \text{reflecting black recognized objects .}$$

The masks and the images of the synthetic objects are computed using the public domain ray-tracer RAYSHADE.

3 Examples

Figures 7 and 8 show images extracted from two RES sequences [2]. The different images illustrate the complete RES process (analysis, synthesis, and mask generation) along with the final result.

[2]These animations can be seen at the WWW site given in the header.

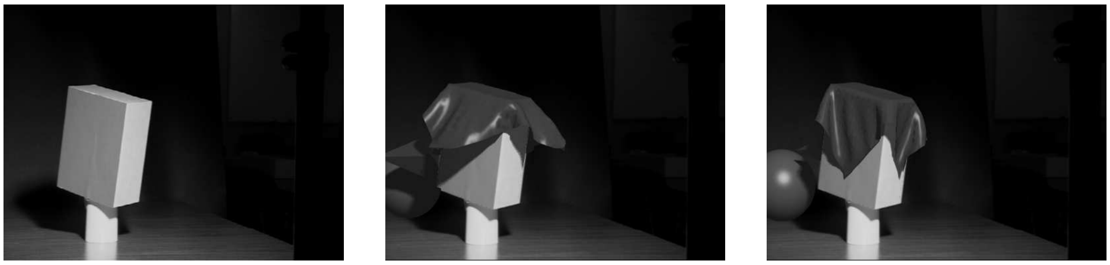


Figure 7: Original image and 2 frames of the sequence enriched by a moving synthetic red cloth, a blue tetrahedron and a green sphere.

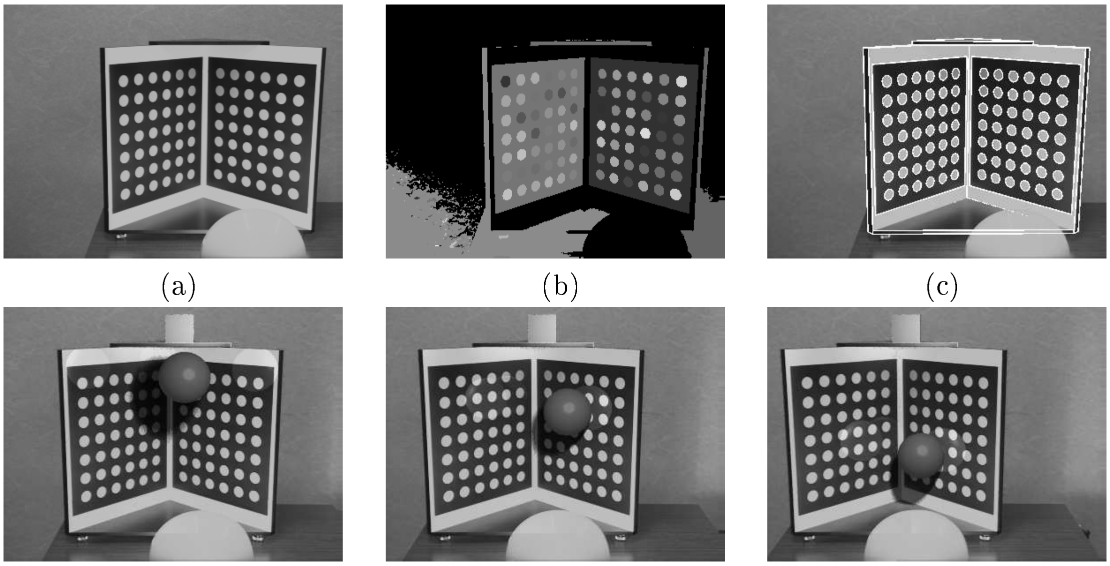


Figure 8: Original (a), segmented (b) and registered (c) images and 3 frames of the sequence enriched by a moving synthetic green sphere and a static pink cube.

Figure 7 illustrate the coherence obtained in the interactions between real and virtual objects: one can note occultations, shadows and contacts with both rigid and non-rigid shapes.

Figure 8 also illustrates how trajectory coherence can be maintained throughout a sequence (the pink virtual cube is attached to the top of the test target), as well as the creation of "virtual" reflections, created at compositing time.

4 Conclusion

We have described a method to compute the interactions between real and virtual objects in video sequences.

We have illustrated this method on a few simple examples to demonstrate the possibility of automatic insertion of 3D synthetic objects in a real image sequence, which should considerably ease and significantly improves the quality of classical video compositing applications.

Moreover, the modularity of the RES process allows us:

1. to use different strategies during the different processes. For example, when it is possible, the analysis process can be solved in parallel, the registration of parts of the real objects present in different video subsequences being computed by different sub-processes.
2. to reuse and "enrich" the same real scene with different synthetic scenes, the analysis process being performed only once and stored as a synthesizable model. Several virtual sequences can serve different purposes, depending on the final application (an expert may have the use for an artificial colormap to identify the critical components of a scene, whereas a standard observer may only be interested in a visually realistic or aesthetically pleasing result).
3. to share a given RES sequence with several users in a distributed networking environment, particularly during the virtual sequence design process. Only graphics-based descriptions of the objects involved have to be specified in a common format, the images being generated on each site according to each user's desire.

Concerning 3D movies production, RES will allow to use image synthesis only where it is necessary, using real scenes or elements of scenes (stored in a Computer Graphics environment) when it is possible (as A. Fellous team has begun to illustrate [Fel94]).

One of the main problem of RES (and all other related techniques) is that it is crucial for the analysis phase to be robust enough to avoid visual artifacts in the enriched images (such as jittering or bad occultation boundaries). Typically this involves obtaining sub-pixel accuracy in the visual projection of aligned 3D models onto the original image sequence.

Up to now, collisions between real and virtual objects are detected and modeled only by using a mass-spring model [Pro95]. To generalize this, work in progress includes attempts to add an interactive tool in the modeling part of the synthesis process, which would automatically prevent interpenetrations during insertion virtual objects in the real scene.

More generally the framework proposed above is not limited to the use of the tools presented here. For example the development of other image analysis algorithms for RES applications is a promising field for future research.

Acknowledgements: We wish to thank all the members of the SYNTIM project and Arghyro Paouri for valuable discussions about various aspects of this research.

References

[AKB+95] K. H. Ahlers, A. Kramer, D. E. Breen, P. Y. Chevalier, C. Crampton, E. Rose, M. Tuceryan, R. T. Whitaker, and D. Greer. Distributed augmented reality for collaborative design applications. In *Eurographics'95*, pages 3–14, September 1995.

[AMG93] A. Ackah-Miezan and A. Gagalowicz. Discrete models for energy minimizing segmentation. In *Proceedings of the 3rd International Conference* *on Computer Vision*, pages 200–207, 11-13 May, Berlin, 1993. http://www-rocq.inria.fr/syntim/textes/iccv93-eng.html.

[BFO92] M. Bajura, H. Fuchs, and R. Ohbuchi. Merging virtual objects with the real world: seeing ultrasound imagery within the patient. *Computer Graphics (SIGGRAPH '92 Proceedings)*, pages 203–210, July 1992.

[BM92] Paul J. Besl and Neil D. McKay. A method for registration of 3D shapes. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 14(2):239–256, February 1992.

[CM92] T. Caudel and D.W. Mizell. Augmented reality: An application of heads-up display technology to manual manufacturing processes. In *Hawaii International Conference on System Sciences*, January 1992.

[Fel94] A. Fellous. The STV-synthetic TV system. In *Journees INRIA Analyse-Synthèse d'images*, pages 28–35, ENST, Paris, January 1994.

[FGR93] A. Fournier, A. S. Gunawan, and C. Romanzin. Common illumination between real and computer generated scenes. In *Proceedings of Graphics Interface '93*, pages 254–262, Toronto, Ontario, Canada, May 1993. Canadian Information Processing Society.

[FMS93] S. Feiner, B. MacIntyre, and D. Seligmann. Knowledge based augmented reality. *Communications of the ACM*, 36, July 1993.

[Fou94] A. Fournier. Illumination problems in computer augmented reality. In *Journees INRIA Analyse-Synthèse d'images*, pages 1–21, ENST, Paris, January 1994.

[HD95] R. Horaud and F. Dornaika. Hand-eye calibration. *Int. Journal of Robotics Research*, 14, Spring 1995.

[Pro95] X. Provot. Deformation constraints in a mass-spring model to describe rigid cloth behavior. In *Proceedings of Graphic Interface'95*, pages 147–154, Québec, Canada, May 1995.

[RG91] S. Randriamasy and A. Gagalowicz. Region based stereo matching oriented image processing. In *Proceedings of Computer Vision and Pattern Recognition*, Maui, Hawaii, June 1991.

[SVCG89] Peter T. Sander, Laurent Vinet, Laurent Cohen, and André Gagalowicz. Hierarchical regions based stereo matching. In *Proceedings of the Sixth Scandinavian Conference on Image Analysis*, pages 71–78, Oulu, Finland, June 1989.

[TG95] J.P. Tarel and A. Gagalowicz. Calibration de caméra à base d'ellipses. *Traitement du Signal*, 12(2):177–187, 1995.

[TV95] Jean-Philippe Tarel and Jean-Marc Vézien. A generic approach for planar patches stereo reconstruction. In *Proceedings of the Scandinavian Conference on Image Analysis*, pages 1061–1070, Uppsala, Sweden, 1995. http://www-rocq.inria.fr/syntim/textes/scia95-eng.html.

[Véz95] J.M. Vézien. *Techniques de reconstruction globale par analyse de paires d'images stéréoscopiques*. PhD thesis, Université Paris-VII, 1995.

[WCB+95] R. T. Whitaker, C. Crampton, D. E. Breen, M. Tuceryan, and E. Rose. Object calibration for augmented reality. In *Eurographics'95*, pages 15–27, September 1995.

[Wil94] R. G. Wilson. *Modeling and Calibration of Automated Zoom Lenses.* PhD thesis, The Robotics Institute Carnegie Mellon University Pittsburg, Pensylvania 15213, January 1994.