Illumination Analysis for Synthetic/Natural Hybrid Image Sequence Generation

Jürgen Stauder[1]
University of Hannover, Institut für Theoretische Nachrichtentechnik und Informationsverarbeitung
Appelstr. 9A, D-30167 Hannover, Germany
Email: stauder@irisa.fr

Abstract

For virtual studio TV production, multi-site video conference applications using a virtual meeting room and for synthetic/natural hybrid coding according to the forthcoming ISO/MPEG-4 standard, a synthetic and a natural scene are mixed to generate a synthetic/natural hybrid image sequence. For realism, the illumination in both scenes should be identical. In this paper, the illumination of the natural scene is estimated automatically and applied to the synthetic scene. For illumination estimation, moving real objects are segmented in the natural image sequence and 3D modeled. The 3D shape, 3D motion and the displaced frame difference between two succeeding images are evaluated to estimate three illumination parameters. The parameters describe a distant point light source and ambient light. Using the estimated illumination parameters, the synthetic scene is rendered and mixed to the natural image sequence. Experimental results with a moving virtual ball mixed automatically into a real video telephone sequence show that virtual objects appear naturally having the same shading and shadows as the real person.

Keywords: MPEG-4, synthetic-natural hybrid coding (SNHC), virtual studio, shading, cast shadow, illumination model

1. Introduction

For virtual studio TV production as in RACE MONA LISA project [1], for multi-site video conference applications using a virtual meeting room [8], and for synthetic/natural hybrid coding according to the forthcoming ISO/MPEG-4 standard [13], synthetic and natural scenes are mixed to generate a synthetic/natural hybrid image sequence. This may be a real object like a speaker in front of computer-generated synthetic background objects.

If the illumination of the natural scene is non-diffuse, e.g. by a spot light or sun light, shading effects [4] will be visible in the natural image sequence, first described by J.H. Lambert in 1760 [6]. Shading is the spatial change of appearing brightness of a real object. For example, a speakers face will appear brighter at the side facing a spot light. Shading may change temporally. For example, the speakers face will get darker, if he rotates its head away from the spot light.

A synthetic/natural hybrid image sequence appears realistically only, if the viewing geometry as well as the illumination of the synthetic and the natural scenes are identical. To align the viewing geometry, the camera parameters of the real camera have to be estimated [11]. This paper is concerned with the illumination aspect. The case of one synthetic and one natural scene is discussed. To let the real and synthetic objects appear as being illuminated by the same illumination, the illumination of the natural scene has to be estimated from the natural image sequence. Then, the estimated illumination can be applied to the synthetic scene. To the know/-ledge of the author, this has not been done before automatically.

Most references for illumination estimation evaluate one image of a natural image sequence assuming the shaded objects having unicoloured texture, see e.g. [12][21] for one or [20] for several point light sources.

In [7], two images are evaluated to estimate a single point light source. The shaded objects need not to be unicoloured as in [12][21][20]. On the contrary, they should be highly textured, because image signal gradients are evaluated. The shape and motion of the real objects have to be known in advance. In [16], the illumination model of a single point light source from [7] is extended by ambient diffuse light that is often caused by indirect illumination from walls or from cloudy sky. A maximum likelihood estimator derives the illumination parameters evaluating two images assuming Gaussian noise as observation error. On the contrary to former approaches, the shaded objects can be textured arbitrarily because the image signal is evaluated directly.

In this paper, the illumination parameters are estimated

[1] Jürgen Stauder is now with IRISA/INRIA, Campus Universitaire de Beaulieu, F-35042 Rennes Cedex, France

automatically from the natural image sequence as in [16]. The illumination estimator requires the segmentation and 3D modeling of the real moving objects. The objects will be segmented assuming static camera and background as in [3]. Because illumination is expected to be non-diffuse, shadows are considered for object segmentation as in [18]. The derived object masks will then be used to estimate the 3D shape, 3D motion and texture of the real objects as in [9]. Using the estimated illumination parameters, synthetic objects will be rendered as illuminated from the natural illumination. Synthetic and real objects will be integrated into one hybrid image sequence using the derived object mask of the real objects as key signal.

The paper is organized as follows. In Section 2, the automatic segmentation and 3D modeling of real objects from a natural image sequence is described. Based on this knowledge, the illumination estimator is derived in Section 3. In Section 4, the synthetic/natural hybrid image synthesis is described. An example of a generated image sequence is shown in Section 5 followed by the conclusion in Section 6.

2. Automatic segmentation and 3D modeling of moving real objects

The method for illumination estimation requires the 3D shape, 3D motion and texture of the real objects from the natural scene in advance. Therefore, the real objects are segmented (Section 2.1), 3D modeled (Section 2.2), textured (Section 2.3) and motion compensated (Section 2.4) automatically. The algorithms will evaluate a current and the previous image of the natural image sequence.

2.1. Segmentation of moving real objects

For segmentation of moving objects in the natural image sequence, the objects are assumed to be highly textured, moving, opaque and not covering each other. The camera as well as the background are assumed to be static. The segmentation algorithm is taken from [3] and is extended to use 3D modeling results from former images as in [9] and to consider cast shadows as in [17][18].

For segmentation, first a change detection mask is estimated from the current and the previous natural image using a noise adapted threshold [3]. The detected changed regions are or-connected with object regions of the previous image if already available. Then, the uncovered background is deleted from the change detection mask using motion information. The motion information is derived using former estimates of the 3D shape and 3D motion [9] (see Sections 2.2 and 2.4). The resulting object mask is finally adapted to luminance edges in the current image.

The segmentation algorithm has been extended to consider cast shadows. Moving cast shadows are assumed on the background of the natural scene, only. They are caused by objects moving between a light source and the background. Moving cast shadows cause a marking frame difference between two succeeding images of the natural image sequence. These signal changes should not be misunderstood as caused by moving objects. Therefore, these regions of temporal signal changes due to moving cast shadows are detected and eliminated from the change detection mask that is derived by the described object segmentation method.

Image regions of temporal signal changes due to moving cast shadows are detected by applying four assumptions to the unknown natural 3D scene. The assumptions lead to four criteria. First, the camera and background are assumed to be static. Image regions changed by moving cast shadows are detected by static edges of static background texture [18] [10]. Secondly, the background of the scene is assumed to be plane. Image regions changed by moving cast shadows are detected by a spatially homogeneous luminance ratio between the current and the previous image. In regions of static background, the frame ratio depends only on changes of illumination. The changes of illumination are low frequent and are rendered to moving cast shadows [18] [10]. Thirdly, the light source causing the shadow is assumed to be non-zero in size compared to the distance between light source and moving object. In this case, shadows have a penumbra causing a slow transition of image luminance from shadowed to non-shadowed regions. Image regions changed by moving cast shadows are detected by soft shadow contours [18] [17]. The fourth assumption is a strong light source. Image regions changed by moving cast shadows are detected by an image difference larger than camera noise. These four criteria are evaluated by heuristic rules resulting in a binary mask of image regions changed by moving cast shadows [18].

2.2. Estimation of 3D object shape

From an estimated object mask, a 3D wireframe is automatically generated. First, the 3D shape of the real object is assumed to follow a distance transform. This transform defines the depth of an object for each point in the image plane inside the object mask evaluating the horizontal distance to the border of the object mask [9]. The result is a object-mask-adapted ellipsoid-like 3D shape. Then, the defined shape is approximated by a closed 3D wireframe.

2.3. Estimation of 3D object texture

The 3D wireframe is textured by mapping the current natural images onto the wireframe using texture mapping. Therefore, a 2D position on the image plane is attached to a 3D position on the surface of the wireframe by perspective projection [9]. The focal length of the camera is assumed to be known. The textured wireframe is called model object. The model object surfaces are matt causing diffuse reflection. Specular reflections [5] are not described. No assumptions are made on the texture of the objects.

2.4. Estimation of 3D object motion

The rigid motion of a real object between two images of the natural image sequence is described by the motion of the model object. The motion is defined by six parameters. Three rotation angles define a 3D rotation matrix $\underline{\mathbf{R}}$ and three parameters define a 3D translation vector $\mathbf{T}$ [9]. The model object is motion compensated by moving each of the corner points of its wireframe from the 3D position $\mathbf{P}_{k-1}$ to $\mathbf{P}_k$ according to

$$\mathbf{P}_k \ = \ \underline{\mathbf{R}}\mathbf{P}_{k-1} + \mathbf{T} \ . \tag{1}$$

3. Illumination estimation

The illumination of the natural scene is described by a time invariant distant point light source and ambient, diffuse light, see Fig. 1. Therefore, the irradiance

$$E(\mathbf{N}) \ = \ 1 + c\, max\{0, \mathbf{L}^T\mathbf{N}\}, \tag{2}$$

is defined as the incident light power per receiving object surface with $\mathbf{N}$ the surface normal vector, $\mathbf{L}$ the unit point light source direction defined by two spheric angles θ, ϕ and c the ratio of point light source and ambient irradiance. The point light source and the ambient light are assumed to have the same colour. Eq. 2 is a subset of Phongs illumination model [14].

For illumination estimation, temporal changes of shading between two succeeding images of the natural image sequence are evaluated. Temporal changes of shading on the surface of a real object will occur, if the real object undergoes a motion including a rotation around an axis different from the illumination direction [16]. Cast shadows, mutual illumination and specular reflections are neglected. The temporal changes of shading are described by a signal model. The signal model describes the image luminance at an image position $\mathbf{p}_k$ in a natural image s_k according to

$$s'_k(\mathbf{p}_k) \ = \ \frac{E(\underline{\mathbf{R}}\mathbf{N})}{E(\mathbf{N})}\, s_{k-1}(\mathbf{p}_{k-1}) \ . \tag{3}$$

In Eq. 3, $s_{k-1}(\mathbf{p}_{k-1})$ is the image luminance at the corresponding image position $\mathbf{p}_{k-1}$ at time instant $k-1$ in the previous image s_{k-1}. The image positions $\mathbf{p}_{k-1}$ and $\mathbf{p}_k$ are linked by motion compensation using 3D object shape and motion (see Sections 2.2 and 2.4). The image position $\mathbf{p}_{k-1}$ corresponds to a point $\mathbf{P}_{k-1}$ on the surface of the model object with the surface normal $\mathbf{N}$.

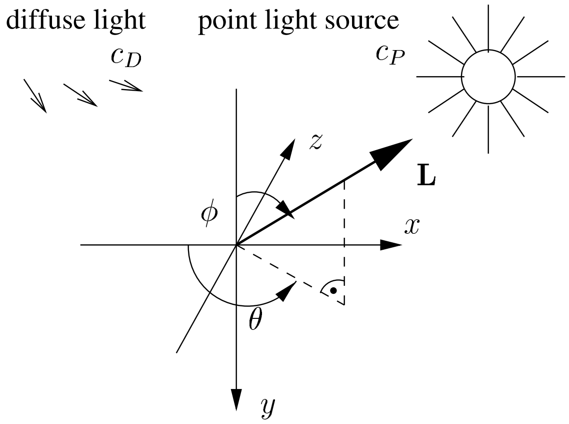


Figure 1. Illumination model: The illumination parameters θ, ϕ define the direction $\mathbf{L}$ of a distant point light source, and parameter $c = c_P/c_D$ defines the ratio of irradiances c_P of point light source and c_D of the ambient light.

Assuming Gaussian camera noise, the maximum likelihood estimate for the illumination parameters c, θ, ϕ is derived by minimizing the sum of squared shading compensated displaced frame difference

$$sdfd(\mathbf{p}_k) \ = \ s'_k(\mathbf{p}_k) \ - \ s_k(\mathbf{p}_k) \tag{4}$$

with $s'_k(\mathbf{p}_k)$ being the predicted luminance according to Eq. 3 and $s_k(\mathbf{p}_k)$ the measured luminance from the current natural image [16]. To build the sum of squares, image positions $\mathbf{p}_k$ with low image luminance gradient are preferred to lower the influence of possibly erroneous motion compensation. Typically, 3000 points are chosen, equally distributed in the object mask. The sum of squares is minimized employing the Gauss-Newton optimization technique [15]. Points with a shading compensated displaced frame difference larger than the standard deviation of differences of all considered points are handled as outliers and excluded from estimation at the beginning of the optimization. Points with $\mathbf{L}^T\mathbf{N} < 0$ are excluded too.

Before the very first illumination estimation, an initial guess for the illumination parameters is derived from a rough systematic search in the parameter space evaluating the cost function. Else, the estimate of the previous image is used as initial guess for the current image. During the image sequence, the estimation results are temporally integrated using a Kalman filter as explained in [16].

4. Synthetic/natural hybrid image synthesis

For image synthesis, the distances between the camera and each of the real objects of the natural scene are assumed to be known up to a scale factor. In this early stage, the static background is modeled as a plane parallel to the image plane and textured as described in Section 2.3. The model objects and the virtual objects are then positioned in a virtual scene. The virtual objects are animated artificially, the model objects are motion compensated according to the real objects. The composed scene is rendered applying the estimated illumination parameters only to the virtual objects. Using a z-buffer algorithm [19], the renderer considers cast shadows generated by virtual objects as well as cast shadows that darken virtual objects [2][19].

5. Experimental results

In this Section, each step and the final result of synthetic/natural hybrid image sequence is shown. In Fig. 2, the second image from the image sequence “Matthias” can be seen. The person translates slowly, rotates his head and talks. The scene is illuminated by a spot light source. Walls and ceiling are bright and reflect some light back into the scene. A cast shadow moving together with the person can be seen on the static background. The camera is static.

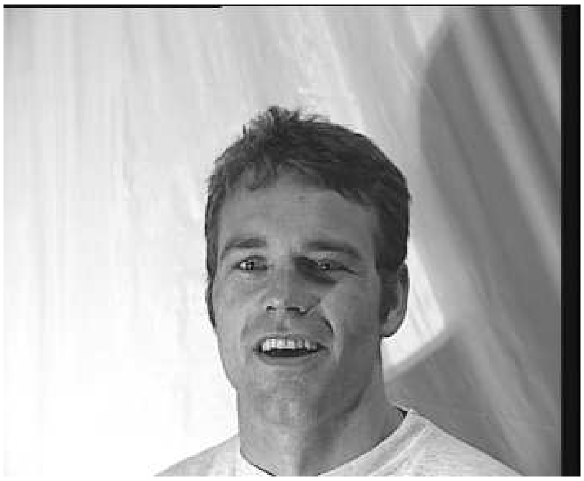


Figure 2. Sample image 2 of natural image sequence “Matthias”, Common Intermediate Format (CIF) 352x288 pel, 10Hz frame rate; The person is illuminated by a spot light shading the persons face and causing a cast shadow at the wall.

Fig. 3 shows the object mask for the second image of “Matthias” (see Section 2.1). Up to a small artifact at the neck, the person has been found accurately. Image regions changed by the moving cast shadow are not included in the object mask. More segmentation results can be found in [17] [18]. In Fig. 4, the 3D wireframe is shown that is generated automatically from the object mask as described in Section 2.2.

Figure 3. Person automatically detected by the described segmentation method for the second image of the natural image sequence “Matthias”

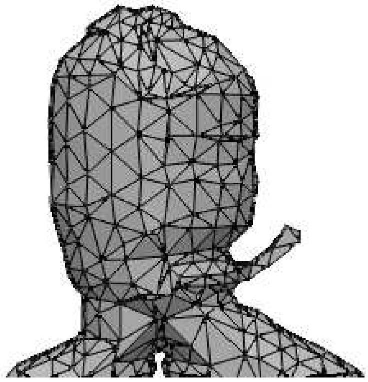
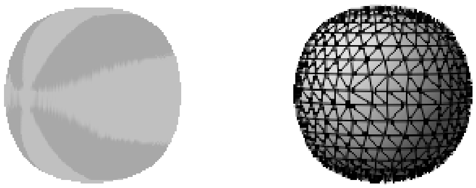


Figure 4. 3D wireframe of the real person, automatically generated from the object mask in Fig. 3, artificially shaded for presentation purposes

As synthetic scene, the synthetic ball in Fig. 5 (a) is used. Its 3D wireframe is shown in Fig. 5 (b). The ball follows an animated motion: Coming from the position of the viewer, it bumps against the background and returns to the front. The background is modeled as a plane, textured by the background part of the natural image sequence.



(a) (b)

Figure 5. (a) Virtual textured ball and (b) its 3D wireframe, shaded artificially for presentation purposes

In Fig. 6, the synthetic/natural hybrid image sequence without consideration of illumination is shown. The image sequence lacks in realism because the virtual object is rendered as illuminated diffusely. In Fig. 7, the same image of the synthetic/natural hybrid image sequence with applied illumination parameters is shown. The synthetic ball and the real person are shaded and generate a cast shadow on the background as being illuminated by the same illumination.

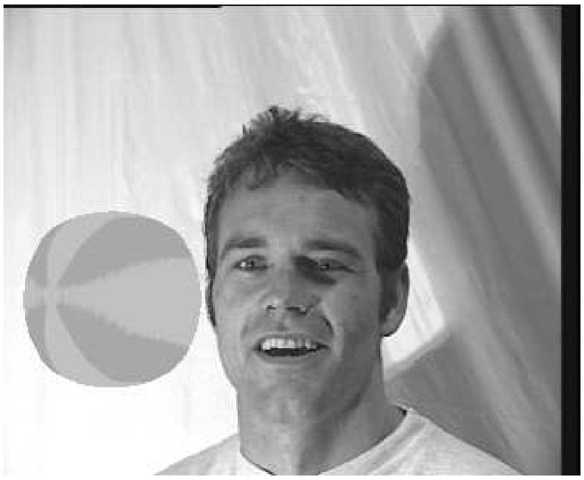


Figure 6. Sample image 2 of synthetic/natural hybrid image sequence without considering illumination

Figure 7. Sample image 2 of synthetic/natural hybrid image sequence with application of the estimated illumination parameters. The ball is shaded and generates a cast shadow like the real person.

6. Conclusion

This paper proposes a method for generation of synthetic/natural hybrid image sequences under consideration of illumination. Real objects from a natural scene and synthetic objects from a synthetic, computer generated scene are mixed into a hybrid image sequence such that the synthetic objects are illuminated by the same illumination as the real objects. Therefore, the illumination of the natural scene has to be estimated automatically from the natural image sequence.

The illumination is described by three illumination parameters. Two spheric angles define the direction of a distant point light source. A third parameter is the ratio of irradiances of the point light source and ambient, diffuse light. The point light source and the ambient light are assumed to be of the same colour. The illumination parameters are estimated automatically from two succeeding images of the natural image sequence evaluating the displaced frame difference, 3D object shape, and 3D object motion. The real objects in the natural image sequence are assumed to undergo a motion including a rotation around an axis different from the illumination direction. No assumptions are made on the texture of the real objects.

The displaced frame difference, 3D object shape and 3D object motion are derived in advance. First, the moving objects are segmented automatically. The segmentation method considers moving cast shadows to distinguish them from moving objects. The objects are assumed not to cover each other. From the object masks, the 3D shape of the objects is calculated assuming an ellipsoid-like shape.

The 3D shape is representated by a wireframe. The wireframe is then textured using the first image of the image couple. After that, a gradient method estimates the 3D rotation and translation of the real object from the first to the second image. The 3D textured wireframe is then motion compensated. Also, its temporal changes of shading are compensated. The illumination parameters are estimated such that the remaining displaced and shading compensated frame difference is minimized.

The estimated illumination parameters are then applied to the synthetic objects. Experiments with a synthetic ball mixed into natural video telephone sequences show that shading and cast shadow of the synthetic ball matches that of the real person resulting in a realistic impression of the synthetic/natural hybrid image sequence.

In the future, the problem of mixing different real objects from different natural image sequences has to be solved. Here, the estimated illumination has to be applied to real instead of synthetic objects.

References

[1] L. Blondé, M. Buck, R. Galli, W. Niem, Y. Paker, W. Schmidt, G. Thomas: "A virtual studio for live broadcasting: The MONA LISA project" , IEEE Multi Media, Vol. 3, No. 2, Summer 1996.

[2] M. Fouad: "Object based image synthesis based on the model of global illumination", technical report, University of Hannover, 1993.

[3] M. Hötter, R. Thoma: "Image segmentation based on object oriented mapping parameter estimation", Signal Processing, Vol. 15, No. 3, pp. 315-334, October 1988.

[4] B.K.P. Horn: "Robot vision" , MIT Press, Cambridge, 1987.

[5] G.J. Klinker, S. A. Shafer, T. Kanade: "The measurement of highlights in color images", International Journal of Computer Vision, Vol. 2, Nr. 7, Juli 1988, S. 7-32.

[6] J. H. Lambert: "Lamberts Photometrie (Photometria sive de mensura de gratibus luminis, colorum et umbrae)", Wilhelm Engelmann, Leipzig, 1892.

[7] N. Mukawa: "Estimation of light source information from image sequence", Systems and Computers in Japan, Vol. 23, No. 10, october 1992, pp. 92-99.

[8] H. Nicolas, J. Motsch: "Very low bitrate coding using hybrid synthetic/real images for multi-sites videoconference applications", Proc. of Visual Communication and Image Processing Conference, 12.-14.2.97, San Jose, USA, SPIE Vol. 3024, pp. 1330-1341.

[9] J. Ostermann: "Object-based analysis-synthesis coding (OBASC) based on the source model of moving rigid 3D objects", Signal Processing: Image Communication, Vol. 6, No. 2, May 1994, pp. 143-161.

[10] J. Ostermann: "Segmentation of image areas changed due to object motion considering shadows", in Y. Wang u.a. (Ed.): "Multimedia Communications and video coding", Plenum Press, New York, 1996.

[11] J.-I. Park, C. W. Lee: "Robust estimation of camera parameters from image sequence for video composition", Signal Processing: Image Communicartion, Vol. 9, No. 1, November 1996, pp. 43-52.

[12] A.P. Pentland: "Finding the illumination direction", Journal of the Optical Society of America, Vol.. 72, No. 4, April 1982, pp. 448-455.

[13] F. Pereira: "MPEG4: a new challenge for the representation of audio-visual information", Picture Coding Symposium, 13.-15.3.1996, Melbourne, Australia, pp. 7-16.

[14] B. Phong: "Illumination for computer generated pictures", Communications of the ACM, Vol. 18, No. 6, june 1975, pp. 311-317.

[15] L.E. Scales: "Introduction to non-linear optimization", Macmillan, London, 1985.

[16] J. Stauder: "Estimation of point light source parameters for object-based coding", Signal Processing: Image Communication, Vol. 7, No. november 4-6, 1995, pp. 355-379.

[17] J. Stauder: "Segmenation of moving objects in presence of moving shadows", Int'l Workshop on Coding Techniques for Very Low Bit-rate Video, Linkping, Sweden, 28.-30. July 1997, pp. 41-44.

[18] J. Stauder, R. Mech, J. Ostermann: "Detection of Moving Cast Shadows for Object Segmentation", submitted to IEEE Transactions on Cirtuit and Systems for Video Technology.

[19] A. Watt: "3D Computer graphics", Addison-Wesley, 2nd edition, 1993.

[20] Y. Yang, A. Yuille: "Sources from shading", Int'l Conf. on Computer Vision, 1990, pp. 534-539.

[21] Q. Zheng, R. Chelappa: "Estimation of illuminant direction, albedo, and shape from shading", IEEE Trans. on Pattern Analysis and Machine Intelligence, Vol. 13, No. 7, july 1991, pp. 680-702.