Real-Time 3D and Color Camera

Frank Forster[1], Manfred Lang[2], Bernd Radig[3]

[1]Siemens AG, CT MS3, Otto-Hahn-Ring 6, 81730 Munich, Germany
[2]Institute for Human-Machine Communication, Munich University of Technology, Arcistr. 21, 80290 Munich, Germany
[3]Informatics IX, Munich University of Technology, Orleansstr. 34, 81667 Munich, Germany
Email: frank.ex.forster@mchp.siemens.de

ABSTRACT

This paper presents a novel real-time 3D-data acquisition system. The system that also provides a normal 2D color image of the scene is based on the structured light approach. It is designed to combine conflicting requirements for 3D-data acquisition systems – a low overall system cost, dense as well as accurate range data and a minimization of the constraints imposed on the scene. Following an extensive review of the state of the art a description of the system is given. First results and the plausibility of the concept are evaluated.

1. INTRODUCTION

Many rapidly growing application fields including augmented and virtual reality require the real-time acquisition of 3D data of arbitrary real-world objects. Yet there are no inexpensive, reasonably accurate 3D acquisition systems available that can provide this data reliably. This paper presents a corresponding system, i.e. a real-time 3D camera designed to combine three conflicting requirements – a low overall system cost, dense as well as reasonably accurate range data and a minimization of the constraints imposed on the scene.

The 3D camera is based on the Structured Light Approach (SLA) whose principle is visualized in figure 1. A projector illuminates the scene with rays or planes of light (called pattern elements) while a conventional camera acquires an image of the scene (called pattern image). If a visible scene point is illuminated by a known pattern element, its 3D co-ordinates are given by the intersection of this pattern element and the line of view determined by the corresponding image pixel and optical center of the camera. The only fundamentally difficult task of the SLA is the identification of the pattern elements in the pattern image if more than one element is projected at a time. Figure 1 shows a simple solution for this task: by assigning each pattern element a different color they can be identified by their color. Figure 1 as well visualizes a problem that occurs if the pattern elements have finite size, i.e. the light planes rather represent cones of light. In that case only the boundary of two pattern elements determines an exactly defined ray or plane of light. The more such boundaries exist, the more dense is the resulting range map.

This work has been partially supported by the European IST program (IST-1999-10087-HISCORE)

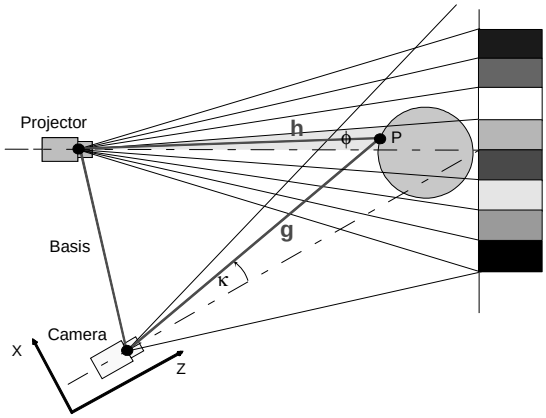

Figure 1: The co-ordinates of P are given by the intersection of the camera's line of view g and the projected pattern element h identified by its color

2. STATE OF THE ART

Many implementations of the SLA (e.g. [1] and [2]) use time-multiplexing to label the pattern elements, i.e. project several distinct patterns sequentially. This approach is usually limited to scenes that are static during the acquisition period. For that reason approaches that use only one camera and projection image (single-shot approaches) have been proposed, e.g. by [3], [4], [5] and [6]. In their case the pattern elements have to be labeled distinctly, i.e. encoded. The primary means of encoding correspond to the light properties measured by the camera on a per-pixel basis. With conventional cameras the options are black-and-white [5], gray levels [7] and colors (with a color camera only).

However, the recognition of these primary means of encoding is very difficult with one-shot approaches since the camera captures only the reflection of the projected light which is strongly influenced by the unknown scene reflectivity. Even with binary (black-and-white) encoding a scene spot that appears black in the pattern image might actually have been illuminated, but have a very low reflectivity. This dependency on the scene reflectivity is particularly strong with color encoding unless each color used consists of monochromatic light. Most single shot approaches using color encoding are for that reason restricted to neutrally colored scenes ([4], [6]), the ones using binary encoding to scenes with a low contrast texture [5]. Even with such scenes or monochromatic colors the identification of colors is very

From: Proc. EUROIMAGE ICAV3D 2001

sensitive to interferences such as mutual illumination of reflecting scene surfaces or an instable camera response if the difference between certain colors is subtle [2] – which is unavoidably the case if many different colors are used. Environment illumination can complicate the identification considerably. It is usually simply ignored, i.e. assumed to be absent or negligible compared to the projection illumination.

In the general case only very few, if any, distinct colors or gray levels can be identified reliably with one-shot approaches and the primary means alone do not suffice for encoding. For that reason spatial encoding has been proposed (e.g. [4], [5] and [6]) where the identification information is locally distributed in the vicinity of a pattern element, i.e. the vicinity forms an identifiable subpattern [5]. This way a large number of pattern elements can be labeled uniquely with only a few distinct colors. Yet a pattern element can only be identified if the subpattern can be established correctly – which is e.g. generally impossible at object boundaries. The impact of this requirement can be kept minimal with a very small size of the subpatterns since then surfaces still smaller than the subpatterns are very difficult to distinguish from erroneous range data in any case.

If only two colors are used as with binary encoding, the succession of colors within a subpattern is pre-determined – elements adjacent to a black pattern element have to be white and conversely. For that reason subpatterns then have to be composed of differently shaped primitives (e.g. [5], [9]). Beumier and Acheroy [9] require the inclination of the scene surface to be approximately constant in the vicinity of a pattern element. Then the relative size of pattern elements is preserved from the pattern to the camera image and they can be labeled by the thickness of their neighbors. However, the use of differently shaped pattern elements results necessarily in sparse range data, since at least some of the pattern elements have to be comparatively large in the camera image – distinguishing between several one or two pixel wide shapes would be too susceptible to noise.

Proesmans et alii [8] describe an implementation of the SLA that requires the pattern elements to retain their spatial arrangement relative to a certain pattern element in the pattern image. This pattern element is considered as origin and remaining elements are identified recursively using their position relative to the origin. Missing pattern elements or artifacts can thus lead to holes in the depth map or in the worst case to systematically erroneous data since identification errors can be propagated. Since the origin is arbitrarily chosen the resulting depth map represents dimension-less shape data rather than accurate range data.

This description of the state of art illustrates a major shortcoming of existing systems based on the SLA: their narrow field of application since each implementation requires either

(1) a completely static scene

or several of the following scene constraints:

(2) The scene texture has a low contrast.
(3) The scene reflectivity is neutral.
(4) The ambient light is controlled or insignificant.
(5) The vicinity of pattern elements is preserved.
(6) The spatial arrangement of the pattern elements is preserved relative to an origin.
(7) The surface inclination is approximately constant in the vicinity of pattern element.

3. SYSTEM OVERVIEW

In the following sections a new 3D camera approach is developed that follows the design objectives of a low system cost, dense as well as accurate range data and minimal scene restrictions.

3.1 General Approach

Since it seems impossible to obtain dense depth data with binary encoding colors are chosen as primary means of encoding. Only a very limited number of colors will be used to ensure a reliable color identification. Monochromatic colors of sufficient intensity are too expensive to produce for a low cost system.

The restriction to static scenes limits the field of application drastically and is therefore not acceptable for our 3D camera. However, dynamic scenes seem to require single shot approaches which in turn leads to the equally unacceptable reflectivity constraint if poly-chromatic color encoding is used. Thus at least one of the two strong constraints seem to be inevitable under the given circumstances. Our approach solves this dilemma as follows: First a reference image of the scene with white illumination is acquired. This image provides the camera's color image output and is at the same time used to determine the scene reflectivity. With this information the pattern colors of the subsequently acquired pattern image can be identified reliably. This technique works even with scenes with deeply saturated hues as long as identical pixels in both images refer to scene spots with similar reflectivity. Of course this imposes an upper limit on the scene movement that is admissible within the time span between the acquisition of the two images. This limit depends on the spatial variation of the scene reflectivity and the duration of the time span. While the former is scene dependent, the latter can be optimized by the system (see section 3.2).

The presented 3D camera has to encode its pattern elements since it projects all pattern elements at once. Given that our approach uses only a small number of different colors, spatial encoding has to be used. As an additional benefit the latter allows the use of redundant codes, i.e. codes with error-detection and error-correction capabilities as proposed by e.g. [5]. Such codes are rarely used with the SLA, but very useful since noise is generally a significant problem with the SLA and especially with low-cost 2D cameras as used for our 3D camera (see section 3.2). The current prototype uses a corresponding spatial encoding scheme based on vertical light planes as the basic pattern elements. Each light plane has a different color where the colors used represent the eight corners of the RGB-cube. These colors have a maximal distance per channel with standard color cameras based on RGB or complementary filters. The subpatterns used are triples of adjacent light planes. Equating each color with a code letter this corresponds to a code word length of 3. The resulting

code space has a size of 392 (7·8·7). To allow the detection of single errors, each pair of code words differs by at least two letters, i.e. the code has a Hamming distance of at least 2. As a consequence only 56 code words out of the code space can be used. Since by far more than 56 light planes are needed for dense 3D data, the code words are currently repeated. This temporarily limits the maximal admissible depth gap of the scene. Currently a two-dimensional encoding is explored which does not require this limit, but is more difficult to decode.

3.2 Hardware Components

The hardware components of the proposed 3D acquisition system consists of a 2D color camera and a projecting device. Both should be inexpensive and at the same time meet the requirements of the proposed approach. The main requirement for the 2D color camera is a minimal time span between the acquisition of the pattern and the reference image. To accomplish that end the following approaches are currently explored:

The minimization can be achieved most directly by using non-interlacing cameras with a high frame rate which are usually rather high-priced. However, the current progress in the area of inexpensive CMOS cameras will allow the use of such cameras even in low-cost systems. CMOS cameras would as well provide the benefit of a high dynamic range and thus allow the system to work with scenes of very high contrast, e.g. in the presence of highlights. Their main disadvantage, the high noise level, is counterbalanced by the robustness of our approach.

An even better, but more costly solution would be the use of two separate cameras that share a common field of view via a beam splitter. The time span between the two acquisitions could thus be reduced to 1 ms or less.

Alternatively sequentially exposed fields can be used with standard interlacing cameras at the cost of a reduced spatial and color resolution.

The required projecting device has to switch between projecting the pattern and the white reference light within milliseconds. Possible devices for this task are slide projectors and computer-controlled projectors. The latter are better suited for switching between different projection images quickly, but substantially more expensive than slide projectors. Therefore a slide projector will be used for our 3D camera that is modified to allow switching between two distinct slides within milliseconds by using a rotating mirror. A corresponding slide projector can be implemented easily at low cost.

3.3 Data Processing

The resulting data processing is very straightforward: The pattern and the reference image are acquired in quick succession. The reference image is used to compute the scene reflectivity, which is in turn utilized to determine the color of the projected light for every pixel of the pattern image. The resulting image (called classified image) consists solely of identified colors, i.e. code letters. Roughly vertical sequences of pixels with identical colors, i.e. the stripes representing the projections of the light planes, are traced in the classified image. Their spatial context, the code letters of the adjacent left and right pattern elements, is established. Using that information every light plane is assigned a code word. Each code word is error checked. Valid code words are then used to identify a light plane. After interpolating the position of the light planes the 3D co-ordinates for every pixel illuminated by an identified light plane are computed, resulting in a depth map of the scene. This color image and the 3D data are provided as output.

4. FIRST RESULTS

The outlined approach has been implemented in a first prototype set-up. Examples of the 3D data obtained with the 3D camera are displayed in figure 2. The major conclusions that could be drawn from first experiments were as follows:

The presented camera principally still imposes two scene constraints: the preservation of the vicinity of pattern elements and the limited scene movement. Yet these restrictions proved to be only minor in praxis for a wide field of application; the width of the pattern elements (light planes) in the pattern image is usually about two or three pixels. Thus only surfaces of this size – which could be hardly distinguished from erroneous range data in any way – can not be measured by the proposed 3D camera.

The restriction of the scene movement has a larger impact on the practical use of the camera, yet occludes only fast moving object unless these are uniformly (including neutrally) colored such as human skin. The proposed camera allows the acquisition of dense 3D data of fairly slow moving objects of arbitrary color – which has previously not been possible.

Yet the most notable result of the experiments so far is the robustness of the outlined approach. Even the prototype works very well with arbitrary scenes with considerable environment illumination. In this context the following measures proved to be very successful:

- The use of only eight distinct colors and of a reference image make the identification of the colors highly reliable.
- The decoding based on stripes rather than on single pixels is barely susceptible to noise.
- The error detecting code allows to detect and resolve systematic decoding errors as caused by object borders.

5. CONCLUSIONS

The first results indicate that a broad range of scenes can be acquired with the HISCORE camera - its only major limitation is currently the restriction of the permissible scene movements with certain scenes. Yet the solutions outlined in section 3.2 are promising to minimize the impact of this constraint for practical applications. The only way to completely drop this constraint might be the use of two separate cameras. This would represent a partial departure from the strict low-cost approach – at least as long as the price of suitable 2D color cameras does not drop substantially.

Figure 2: Reference Image – Pattern Image – Point Clouds (rotated to visualize the depth differences)

6. REFERENCES

[1] G. J. Agin and T.O. Binford, "Computer Description of Curved Objects", Proc. IJCAL, pp. 629-640, 1973

[2] D. Caspi, N. Kiryati and J. Shamir, "Range Imaging with Adaptive Color Structured Light", IEEE PAMI 20(5):470-480, 1998

[3] J. Tajima and M. Iwakawa, "3D Data Acquisition by Rainbow Range Finder", Proc. ICPR, pp. 309-313, 1990

[4] K.L. Boyer, A.C. Kak, "Color-Encoded Structured Light for Rapid Active Ranging". IEEE PAMI 9(1):14-28, 1987

[5] P. Vuylsteke, A. Oosterlinck, "Range Image Acquisition with a Single Binary-Encoded Light Pattern", IEEE PAMI 12(2):148-164, 1990

[6] J. Salvi, J. Batlle and E. Mouaddib, "A Robust-Coded Pattern Projection for Dynamic 3D Scene Measurement", Pattern Recognition Letters 19(11):1055-1065, 1998

[7] B. Carrihill and R. Hummel, "Experiments with the Intensity Ratio Depth Sensor", Computer Vision, Graphics and Image Processing 32(1): 337-358, 1985

[8] M. Proesmans, L.J. Van Gool, A. Oosterlinck, "Active Acquisition of 3D Shape for Moving Objects", Proc. ICIP, pp. 647-650, 1996

[9] C. Beumier and M. Acheroy, "3D Facial Surface Acquisition by Structured Light", Proc. International Workshop on Synthetic-Natural Hybrid Coding and Three Dimensional Imaging, pp. 103-106, 1999