THE QUEEN'S AWARDS FOR ENTERPRISE 2001

BBC

R&D White Paper

WHP 072

September 2003

New methods of image capture to support advanced post-production

G. A. Thomas[1], M. Koppetz[2] and O. Grau[1]

[1]BBC R&D, UK and [2]ARRI, Germany

Research & Development
BRITISH BROADCASTING CORPORATION

BBC Research & Development
White Paper WHP 072

New methods of image capture to support advanced post-production

G. A. Thomas[1], M. Koppetz[2] and O. Grau[1]

[1]BBC R&D, UK and [2]ARRI, Germany

Abstract

Advances in post-production technology make it ever easier to manipulate captured images in new ways, to perform functions such as changing of frame rate and insertion of virtual objects. However, such processes could be improved or made significantly easier by capturing additional data at the camera. The MetaVision project, funded by the EU's IST programme, has been investigating how to capture and record such data and how to use it to assist the post-production workflow. Now in its final year, the project has developed a camera system demonstrator capable of capturing additional data in the form of high frame-rate images, and images for depth estimation. Experiments are being conducted using this camera system, which will allow the concepts pioneered by the project to be tested. This paper discusses some of the innovative aspects of the camera system, and presents some results.

This document was originally published in the Conference Publication of the International Broadcasting Convention (IBC 2003) Amsterdam 11th-15th September 2003

White Papers are distributed freely on request.

Authorisation of the Chief Scientist is required for publication.

NEW METHODS OF IMAGE CAPTURE TO SUPPORT ADVANCED POST-PRODUCTION

G. A. Thomas[1], M. Koppetz[2] and O. Grau[1]

[1]BBC R&D, UK and [2]ARRI, Germany

ABSTRACT

Advances in post-production technology make it ever easier to manipulate captured images in new ways, to perform functions such as changing of frame rate and insertion of virtual objects. However, such processes could be improved or made significantly easier by capturing additional data at the camera. The MetaVision project, funded by the EU's IST programme, has been investigating how to capture and record such data and how to use it to assist the post-production workflow. Now in its final year, the project has developed a camera system demonstrator capable of capturing additional data in the form of high frame-rate images, and images for depth estimation. Experiments are being conducted using this camera system, which will allow the concepts pioneered by the project to be tested. This paper discusses some of the innovative aspects of the camera system, and presents some results.

INTRODUCTION

For over a century, moving pictures have been based on the same principle: the presentation of a series of individual images in rapid succession, exploiting the inertia of human vision to produce the illusion of motion. Each image is a discrete sample of the spatial and temporal continuum being photographed, a framed, two-dimensional representation of a moment in time.

In its most basic form, the technique is remarkably simple: images are captured at a certain rate and then presented at the same rate, a one-to-one transformation of the real time-line to the time-line simulated in the presentation. Early on, film pioneers began to experiment with the parameters of this transformation. At a time when film cameras were still being cranked by hand, cinematographers had discovered that by varying the capture rate in relation to the later presentation rate they could produce quite remarkable effects that went beyond the powers of observation given by the naked eye. Slow motion photography enabled the resolution of movement so rapid that it appeared only as a blur to the observer while time–lapse photography showed the movement in processes as slow as the growing of grass. Animation even allowed static images to move. Today, modern cameras can be controlled to vary the frame rate during the take, thus introducing completely new effects.

The spatial transformation, too, offers many aesthetic possibilities. By framing and selecting the focal plane, the cinematographer directs his viewers to a very specific slice of space. This sample can be varied, moving the audience's focus of interest from one element to another. The depth of focus can be large, encompassing a large portion of the photographed scene, or it can be small to single out a certain part.

Technological developments have further added to the possibilities of motion picture production. Today each captured image can be individually processed. Elements can be added to the image, either from other real photographers or in the form of synthetically generated graphics.

At the same time, technology has also added new complications to the process. The development of television broadcast led to a multitude of standards, each a slightly different approach to the task of distributing moving images despite limitations of processing rate and bandwidth. The technical legacy of these early developments continues to influence the broadcast industry today. Different frame rates in different markets, backwards-compatibility to older standards, data compression: these are only a few of the aspects that must be considered.

With all these developments, the original principle of cinematography is reaching its limits. The captured images remain very limited samples of the scene they represent. Between two exposures, there exists an interval that contains data. The two-dimensional image only gives an impression of the spatial build-up of the scene. The problems begin when, for example, the frame-rate of the presentation requires intermediate images, or when an element needs to be added to the image in a seemingly correct location in 3D space.

Much effort has been put into surmounting these problems. Interpolation algorithms seek to fill in the missing temporal information. 3D information can be obtained by tracking the relative motion of elements in the scene and calculating their location based on the laws of perspective. Still, all these approaches are based on the estimation of data that is simply not included in the discrete samples provided by cinematographic images, and so are limited in the quality they can achieve.

The capture of additional data might seem to be the solution and, indeed, metadata is the catch-word of the day. Almost anything can be metadata: the time at which an image was taken, the position of the camera, the location of the lighting, etc. Catalogues have been compiled to describe the many different forms of metadata and yet the problem remains: metadata can describe the context in which the images were taken but the images remain discrete samples of a spatial and temporal continuum.

Of course the most straightforward approach would be to simply increase the sampling rate. A higher frame rate could provide a more accurate temporal representation, and multiple cameras could provide a three-dimensional image of the scene. In the real world, however, the data rates that can be transferred and stored are limited, if not by technology, then by cost. A system based on massive redundancy of data would have little chance of success. Most of the data would never actually be used and would only be carried along as ballast, severely limiting efficiency.

This is where the MetaVision project (Walland et al. (1, 2)) seeks to offer an alternative. The underlying concept is simple: to apply compression to the captured data, but to match the degree of compression to the kind of captured image. Thus the captured images are divided into two classes:

- the “primary” image frames, which are used directly in generating the final image sequence - these are compressed losslessly in almost all circumstances;
- the “meta-images”, whose purpose is to support post-processing of the primary frames, and which can tolerate some lossy compression. Two kinds of meta-images are considered:
    - the “secondary” frames, providing additional images of the scene at intermediate times, which are used indirectly, to support post-production processes requiring temporal interpolation - these are compressed with a small degree of loss.

- the “3D” frames, providing images of the scene from which depth information can be derived - these are sampled at a lower resolution, and could also be mildly compressed.

MetaVision has investigated the actual implementation of this principle and intends to demonstrate its advantages in practical applications all along the motion image production and distribution chain.

The first aspect of the project has been the development of a suitable method of image capture. This has involved an investigation of the additional forms of data-capture that can be used to produce the required meta-images, along with the developments of methods with which to link them to the original images.

To obtain the additional temporal data, a camera has been developed which not only records high-quality images at a standard frame rate, but also provides intermediate images at a higher frame rate. Methods have also been devised to capture additional depth information. The favoured approach involves two or more ancillary video cameras that are offset in relation to the main camera axis, and capture images synchronously with the primary frames of the main camera. This additional information can later be used to calculate a depth map of the photographed scene.

The remainder of this paper is organised as follows. The requirements for the main camera are set out in the next section, followed by a discussion of how these have been realised in the camera system demonstrator. The following section discusses the use of additional cameras to provide depth information via a stereoscopic approach. The approach implemented in the project demonstrator is then described. Finally, some results are presented and conclusions are drawn.

CAMERA REQUIREMENTS

One of the major goals of the project was to ensure that the novel concepts listed above could actually be realised within the context of a professional working environment. A survey was conducted to determine the needs of potential users in the different applications addressed by the project.

The requirements for the image capture components were found to be particularly demanding. First of all, professional image capture is still largely influenced by film-based technology which sets the standard not only in terms of image quality but also with respect to working style. Film cameras represent a mature technology embedded within an evolutionarily grown system of ancillary components and accessories.

Secondly, the creative aspect of capturing moving images plays an important role in the user requirements. Going beyond the mere representation of an existing scene, cinematographers have developed subtle story-telling techniques, employing optical and temporal effects to manipulate the audiences’ perception. It was of particular interest to the project to determine the role that the limitations of existing technology have had in influencing the artistic techniques that viewers are used to. In some cases, such limitations could today be overcome, while the technique that was developed to compensate them remains an important element of the cinematographer’s artistic repertoire.

Developments in post-processing and distribution have also influenced the camera requirements. Due to the increased use of digital image processing such as colour manipulation, image compositing or integration of synthetically generated elements, new standards have been set for the quality and information content of the captured images. It is not only important to capture images with a high level of resolution, dynamic range and colour fidelity: additional data which describes the exact circumstances is necessary to

facilitate the further handling of the images in post-production.

Of course, not every application in which motion pictures are captured will combine all requirements at the highest level. Nonetheless, in defining the goals of the project, an effort was made to validate the concepts against the most stringent demands of the users to show the viability of the technology not only within the limited functionality of a demonstrator, but also to prove their usefulness for future developments.

The following sections summarise the user requirements for the camera:

Lenses

A high-quality camera should allow for the use of lenses comparable to 35mm-format prime and zoom lenses. Since such lenses are readily available in the required optical quality, it is seen as a benefit for the camera to be able to use existing lenses. These lenses also set the standard required in terms of size, weight and ruggedness.

Viewfinder

Most users stated the need for a colour through-the-lens viewfinder that would allow them to judge not only framing but also focus, colour and lighting conditions. Furthermore, a viewing area outside the actual capture area was deemed to be very useful for dynamic framing e.g. stopping a pan movement before undesired elements such as microphone booms etc. enter the image.

Imaging

It is necessary for the imaging sensor to be as large as possible for it to offer the same possibilities of limiting the depth of focus as are given with 35mm film. The feature was underscored as being as essential creative element in image composition.

As far as the required resolution of the sensor is concerned, most users equated the pixel count to the resolution currently employed in scanning film material. This ranges from 4000 pixels in width for 35mm film down to 1920 pixels for Super 16. It was also found that many television productions are captured on film and the scanned at standard television resolution. In these cases, it is not the absolute resolution but other aesthetic elements that have led to the choice of film as the capture medium.

Operation

The ability to create slow-motion, time-lapse, speed-ramp and other temporal effects is seen as being essential for a production camera.

Furthermore, the camera needs to be robust and mobile; control over the basic functions must be simply and reliable. On the whole, users would prefer to capture as much data as possible on set and decide later how to “fine-tune” the images. At the same time, cinematographers see the need to define their artistic intentions in such a way that they can be accurately reproduced in post-production.

Output

It should be possible to record at least 2 – 5 minutes of non- or losslessly compressed images on board the camera. This data could then be downloaded to a storage medium with greater volume. At the same time, a live viewing output, possibly with reduced bandwidth is required for monitoring purposes.

CAMERA DEMONSTRATOR

The camera demonstrator was designed to show how the concepts developed in the MetaVision project could be used to satisfy the requirements determined in the user survey. The goal was to use innovative technology to achieve the results in a more efficient fashion than with current technology, while at the same time preserving proven elements of the existing working style. Of course, in some cases, the technology that would have been required to achieve the very top-level requirements would have gone beyond the scope of the project. However, it will be shown that such results can be realised by extrapolation of the solutions contained in the demonstrator.

Figure 1 shows a schematic layout of the demonstrator. Unlike most conventional video cameras, the demonstrator features a single large format sensor (Fig. 2). This configuration was necessitated be the requirement for high-quality cine lenses. To achieve their high level of optical quality, these lenses have been designed with a relatively short back-focal distance. This effectively ruled out the possibility of introducing beam-splitter prisms between the lens and the image sensor, the additional optical losses of such elements notwithstanding. Also, it was necessary to ensure that the known imaging characteristics of the lenses e.g. field of view and depth of focus were preserved. Under these constraints, a single, large format sensor with an effective imaging area of 18 x 24 mm was selected. To provide a colour representation, the sensor is fitted with a Bayer-type colour mask that assigns different colour filters to adjacent pixels.

Within this configuration, it was possible to incorporate a mirror-reflex viewfinder system as is used in professional film cameras. A spinning half-disc-shaped mirror alternately allows light to pass to the sensor, then reflects it to a ground-glass plate that can be viewed through an optical viewfinder. This optical system provides a very accurate representation of the photographed image, including such characteristics as colour or depth of focus. Also, it provides the required additional viewing area around the captured image area.

To provide a platform that would most closely address the needs specified by the users, it was decided to base the general layout of the camera on an existing 35mm film camera. The cast metal housing of the camera provides a stable and rugged basis for the optical and electronic components of the camera. An additional benefit of using components based on professional film cameras in the demonstrator is that it facilitates compatibility to existing camera accessories, thus making it possible to test the camera in a realistic fashion.

The sensor has an effective pixel count of 2880 x 2160. In the demonstrator, it is clocked to run at a frame rate of 72 fps. In this mode, the image area is windowed down to 1920 x 1080 pixels to reduce the data-rate and allow the use of conventional interface technology. The sensor is based on CMOS technology, which was chosen over CCD technology for a number of reasons. Most importantly, it allows for high frame rates despite the high pixel counts, which is essential for demonstrating the MetaVision concept at professional quality levels. The technology also provides a better basis for future developments. In comparison to CCD sensors CMOS offers a greater design flexibility under the economic constraints of specialised application with limited production volume. Additionally, CMOS has lower power requirements and the off-chip electronics are not as critical for the final image quality.

In the demonstrator, the image signals are output from the sensor via 32 parallel channels, conditioned and converted to digital signals with parallel A-to-D converters. In firmware-based processing units, the signals undergo gain/offset correction in real-time to remove fixed errors and are passed on to a data bus that can accommodate a maximum bandwidth of 10 Gbit/s. Attached to this bus are two parallel processing/interface units based on HD-SDI interface technology (1.5 Gbit/s gross band-width per channel). Each image is split vertically and one half image transferred per channel. Additionally, a low-resolution output provides a rudimentary processed image for monitoring purposes.

At the HD-SDI interface, the data is still in raw form, i.e. the individual pixels each represent only one colour as determined by the Bayer-mask. By abstaining from further data processing in the camera, the demonstrator makes the most effective use of the available bandwidth. Colour grading decisions made by the cameraman can be documented as metadata and applied to the images in later processing.

Despite the effective use of interface bandwidth, further data-rate reduction is necessary for practical data handling. To achieve this, each data channel is passed through a modified MPEG-encoder. Each encoder carries out a near-lossless compression of every third half-image in the 72 fps stream. The remaining half-images undergo a much higher degree of compression, resulting in a much more manageable data rate which can be recorded on a PC-based server.

This data can now be processed off-line to produce the desired output format. The half-images are rejoined and the RGB-colours are reconstructed using an interpolation algorithm. The near-losslessly compressed images provide the basis for a standard rate image stream at 24 fps. The highly compressed intermediate images can be used to calculate motion vectors, which in turn can be used to interpolate new intermediate images between the high-quality images of the 24 fps stream, thus providing the means to generate rate changes or format conversions.

Although the demonstrator is based on MPEG hardware, the MetaVision scheme can also be implemented using other approaches. Ideally, the compression of the primary images should be lossless, and research carried out within the project on the basis of JPEG-2000 compression algorithms has shown this to be feasible.

DEPTH CAPTURE REQUIREMENTS

Potential applications that can make use of “meta-images” representing depth were presented by Grau et al. (3), together with a review of possible depth capturing methods. Applications that were identified included the insertion of virtual objects into real scenes, where effects such as shadows, lighting and occlusions could be modelled by using depth information of the real scene.

Capture requirements

The review concluded that no single depth acquisition method meets all the potential requirements. There are promising new methods under development that may be applicable over short ranges in indoor environments, but stereo-based methods appear to be the only practical method for outdoor environments, and can operate with long ranges by using wide baselines. Given that no single technique suits all applications, the architecture of the MetaVision system was chosen to be independent of the choice of depth measurement method. The system captures “raw” depth information (in whatever image-based form it is produced) and stores it together with an indication of the kind of processing that should be applied in order to generate depth information.

We chose to develop a stereo-based system in the project, and in this case the raw information consists of two additional image streams from two auxiliary cameras placed either side of the main camera.

Processing requirements

The captured information is processed off-line to generate the required 3D data. The depth estimation process needs to be robust to brightness differences between the images, since

it is impossible to obtain perfectly-matched images from two different kinds of cameras under operational conditions. Similarly, it should be robust to small calibration errors between the images. Temporal consistency is important, to avoid time-varying artefacts in the processed sequence. Although real-time processing is not required for the applications considered, the time to process a frame should be no more than 10 seconds or so, otherwise the time taken to process a long sequence becomes prohibitive.

DEPTH CAPTURE DEMONSTRATOR

The project demonstrator has two auxiliary cameras mounted rigidly either side of the main camera. The separation between the main and auxiliary cameras is adjustable, but a value of about 25cm has proved to be a good compromise between depth resolution and camera size. The cameras are standard TV-resolution RGB cameras, triggered synchronously with the 24Hz primary frames from the main camera and capturing progressively-scanned images. They have zoom lenses, which are manually-adjusted to roughly match the field-of-view of the main camera. They are then calibrated automatically using a calibration chart. In a fully-engineered system, the cameras' zoom and focus could be matched automatically with the main camera. The video signals from the cameras are captured directly onto a disk array.

To generate the depth information, a disparity estimation algorithm using a 3D recursive search is being used, a description of which was given by Thomas and Grau (4). Although the auxiliary images are only of standard TV resolution, the disparity map is required at the full resolution of the primary image. The algorithm has therefore been extended to use edge information from the primary image to 'steer' the edges in the disparity map to coincide with these edges.

From the disparity map, 3D information can be generated in various forms, as required by the particular post-production tool being used. Some applications can use the disparity map directly, for example after placing it in the depth channel of an RLA file. For other applications, we generate a 3D mesh representing the scene. The original image can be texture-mapped onto this mesh, and imported into a 3D modelling package for applications such as inserting virtual objects which cast and receive shadows from real objects. The mesh conversion process includes polygon reduction, smoothing, and the detection of discontinuities corresponding to edges of objects (so that no mesh is generated at these boundaries).

RESULTS

This section presents some of the results available at the time of writing, however further results and a live demonstration should be available during IBC2003.

Camera performance

An example of the use of additional temporal data is shown in Figure 3. This shows the benefit that this data provides when interpolating a new frame at a mid-point in time between two primary frames, as would be required when simulating a camera with a continuously-variable frame rate.

Depth sensing

An example of the use of a depth signal for depth-based keying is shown in Figure 4. A depth map was generated for an image of an outdoor scene, using images from the two

auxiliary cameras. The depth map was thresholded at the depth corresponding to the person to generate a matte signal, which was then used to composite a virtual object into the scene.

CONCLUSION

We have described how additional "meta-images" may be captured, in order to better describe the scene in both time and 3D space. The resulting high data rate may be reduced by carefully-targeted data compression, so that the primary images may be stored with no (or negligible) loss. A demonstrator system to test this concept has been described, and some results have been presented that demonstrate the kind of post-processing that the "meta-images" can support.

ACKNOWLEDGEMENTS

The authors would like to thank their colleagues in the MetaVision project for their help.

REFERENCES

1. Walland, P.W. et al, 2002. The application of intimate metadata in post production, Proc. Of International Broadcasting Convention, Amsterdam, September 2002

2. http://www.ist-metavision.com

3. Grau, O., Minelly, S. and Thomas, G.A., 2001. Applications of Depth Metadata, Proc. Of International Broadcasting Convention, Amsterdam, September 2001

4. Thomas, G.A., Grau, O., 2002. 3D Image Sequence Acquisitions for TV & Film Production, Proc of 1st Int. Sym. On 3D Data Processing Visualization and Transmission (3DPVT 2002), Padova, Italy, Jun 19-21, 2002

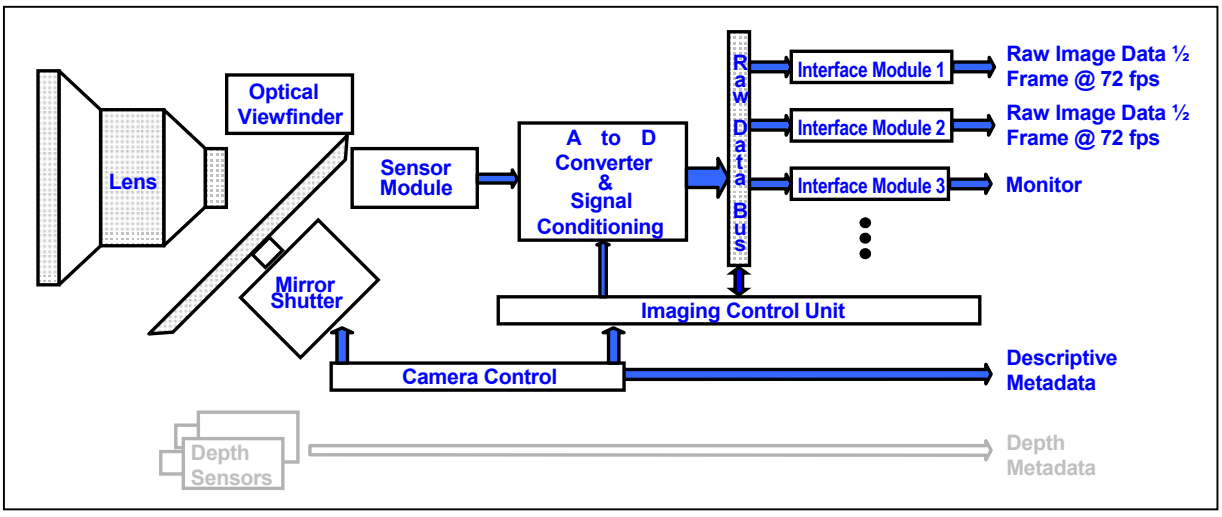

Figure 1 - Schematic layout of the camera demonstrator

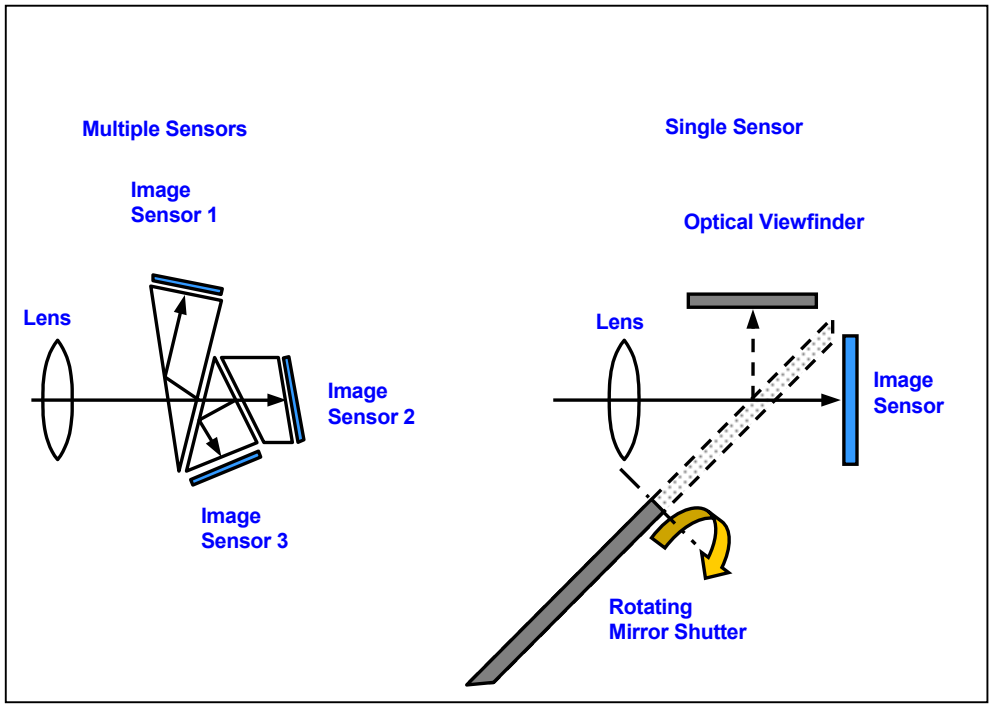

Figure 2 - The chosen single-sensor arrangement compared to the use of multiple-sensors

Linear interpolation between 24fps frames | Motion compensated between 24fps frames | Motion compensated between 72fps frames

Figure 3 - Use of additional temporal data for image interpolation

Original image | Depth map | Object inserted using depth matte

Figure 4 - Use of depth information for 3D compositing