ENHANCED METHODS FOR EFFICIENT VISUALIZATION
OF MULTIMEDIA CONTENTS

Son Minh Tran

Ph.D. Dissertation Summary

Supervised by

Dr. András Gschwindt
Department of Broadband Infocommunications and Electromagnetic Theory
Budapest University of Technology and Economics

1 Introduction
Multimedia is any combination of text, graphic art, sound, animation and video delivered to the user by computer or other electronic means. It is richly-presented sensation. When content-author weaves together the sensual elements of multimedia — dazzling pictures and animation, engaging sounds, compelling video clips, and raw textual information — he / she can electrify the though and action centers of people’s minds. Therefore the resulting multimedia product can excite eyes, ears, fingertips, and, most importantly, the head. It helps end-users (they are readers, listeners, viewers, players and manipulators at the same time) capture the ideas and the services conveyed in that multimedia content in depth.

The superior advantages in presenting information of multimedia attract more and more researchers in the new but very perspective field: multimedia computing. Research and development efforts in multimedia computing have been divided into two areas. As the first area of research, much effort has been centered on the stand-alone multimedia workstation and associated software systems / tools, such as music composition, computer-aided learning and interactive video. However, the combination of multimedia computing with distributed systems offers even greater potential. Promising new applications based on the latter area include multimedia information systems, collaboration and conferencing systems, on-demand multimedia services and distance learning. The fundamental characteristic of both areas is that they incorporate digital data of various media types, such as image, (animated) graphics, voice, music and video. It implies — among the others — a crucial question for the success of multimedia: efficient data compression for handling, storing and transmitting.

The work in this dissertation has been carried out as an effort toward a partial solution for the above question. At present, a large range of innovative facilities for efficiently compressing and conveniently manipulating multimedia components are intensively comprehended. Among them, the dissertation addresses a low cost, three-dimensional (3D) presentation of visual object, compression of object-based video and multimedia content enabling high interactivity. The fact that these three topics are also the hot key-techniques of the recent MPEG-4 international standard, emphasizes their important roles in the current research of multimedia computing.

2 Objectives of the research
The dissertation introduces several new results achieved in the three relatively new fields of current research on multimedia computing: 3D displaying methods, object-based video compression and interactivity-enabling multimedia content. The obtained goals can be summarized as the followings:

The Single Image Random Dots Stereogram (SIRDS) is a well-known technique for representing 3D object on a normal two-dimensional (2D) plane without the lost of the vivid depth information. The general formula for computing
the so-called pixel-constraint in SIRDS is deduced, ensuring the geometrically correct calculation for complex 3D scene. Hidden surface removal plays an important role in the visibility of an autostereogram (resulting image of SIRDS algorithm). A new method for detecting hidden surface is proposed. On one hand, this method reduces the artifacts in identifying hidden surfaces; on the other hand, it offers scalable removal of hidden surfaces. Another feature of the new method is that the existing hidden surface removals can be derived as a specific case of the proposed one. SIRDS is also extended to the domain of moving objects. The reason of the difficulty in viewing a moving autostereogram is found, namely color-constraint. Three methods are proposed for solving the compromise emerging from the coexistence of color- and pixel-constraint, which plays a key-important role in generating visible autostereogram for moving objects.

The recent multimedia compression standard MPEG-4 is the first to address the encoding method of video on the basis of separate objects with arbitrary shapes. The introduction of video objects (VOs) in compression opens a possibility for individual treatment (therefore it provides better compression ratio) as well as convenient manipulation of video components in a multimedia scene. A new type of such objects is proposed, namely integrated mesh video object (IMVO). The goal of introducing IMVO is twofold. Firstly, it demonstrates a full solution to fill up the gap between the compression of conventional frame-based video and object-based video with arbitrary shape, which is not specified in MPEG-4. Secondly, the encoding method for IMVO implies several improvement and integration of various enhanced tools. Constrained Delaunay mesh is used to represent the temporal snapshot of a VO with arbitrary shape — Video Object Plane (VOP). The topology of the mesh is also exploited to reconstruct some later VOPs in motion compensation fashion, the famous key-technique for video compression. The most bit-consuming information of the technique, the residual errors between the estimated and original VOPs are carefully handled. Polynomial estimation and wavelet transform are proposed to encode residual errors. They can be used alternatively to produce a graceful scalable solution for efficient compression of residual errors and the whole codec scheme of IMVO in general.

The recent MPEG-4 standard also deals with composition and interactivity in multimedia content. The driving motor of this new feature, the Binary Format for Scene (BIFS) is studied. Effort was made to ease the usage of BIFS with a high abstraction level, script-free authoring tool. The capability of BIFS is evaluated through the creation of several scenarios of interactive multimedia contents. The completion of this innovative feature provided by BIFS is investigated in a comparative manner with Flash technology, a mature industry platform for authoring interactive contents. New BIFS node is also proposed to enhance the capability of BIFS in creating multimedia games, the most sophisticated scenario supported by MPEG-4 BIFS.
3 Methodology of the research

The algorithms and formulas deduced in the field of SIRDS are mainly based on geometry mathematics. They are simulated with software implementation. The results obtained on the visibility of the generated autostereograms are evaluated with both mathematical model and subjective human perception.

The results related to the MPEG-4 video standard firstly emerge from a comprehensive study of the proposed tools in the standard. A bottom-up design was applied to construct the new visual object IMVO for encoding: all functional units are named and specified to give an overview on a complete scheme of the new codec; then a fine-tuning process is applied to individually optimize the operation of critical units. Mathematical tools, such as decomposition, wavelet transform, interpolation as well as graph theory are deployed in these optimizations.

Dealing with the system information BIFS of the MPEG-4 standard — a revolutionary new feature in video compression technique — an incremental learning method is applied. First, an authoring tool for user-friendly creation of rich multimedia scene is presented. More than just for demonstrative purpose, the authoring tool is intended to assist the scene-designer to explore the new issues supported by BIFS easily and comprehensively. Then a comparison between MPEG-4 BIFS and matured Flash technology is studied for better persuasion on the new perspectives of the MPEG-4 application. Finally, new elements are proposed to the structure of the MPEG-4 BIFS in order to improve its performance in creating complex content, such as multimedia games.

4 Structure of the dissertation

The dissertation is structured as follows:

Chapter 1 briefly describes the circumstances which induced the research of the dissertation. A summary of objectives and the structure of the discussion in the dissertation are given in this chapter.

Chapter 2 discusses the SIRDS algorithm and the new contributions. First, it gives an overview of the algorithm. The existing method is then generalized in the formulas for the calculation of pixel-constraint and in hidden surface removal. Proposals for solving the trade-off between color and pixel-constraint in motion domain are also discussed in this chapter.

Chapter 3 contains the description of the video object concept in multimedia compression. A review of motion compensation for video compression is made. Two new methods for encoding residual errors based on polynomial and wavelet are introduct. Both are then integrated in a full compression scheme of IMVO, whose input is frame-based video sequence, and output is VO of interest.


Publications

Journal papers


Conference papers


[C3] S. M. Tran, M. Preda, F. J. Preteux, K. Fazekas, “Case study: a simple composing tool for editing MPEG-4 system information”, 4th European Workshop on Image Analysis for Multimedia Interactive, Queen Maryland, United Kingdom, April 2003


Chapter 4 describes the structure of BIFS defined in MPEG-4 for enabling the composition and interactivity in multimedia content. The usage of BIFS is demonstrated through an authoring tool of high abstraction level. The interactive capability of MPEG-4 content is discussed in comparison with Flash, a successful technology for composing multimedia content. Proposal of a new BIFS node is also included in this chapter.

Chapter 5 closes the dissertation with the conclusion, applications of the results proposed in the previous chapters as well as some perspectives for the future research of the related fields.

5 Summary of the new results
The new results achieved in the research can be arranged into three main groups: the generalized SIRDS algorithm, the new IMVO codec scheme and the exploration of MPEG-4 BIFS.

5.1 The generalized SIRDS algorithm
SIRDS algorithm is one of the rare techniques that can reconstruct the depth information of a 3D scene at low cost, for both content providers and viewers. The outputs of the algorithm — the autostereograms — are normal 2D images but the true 3D scene can be perceived by viewing properly. An example of an autostereogram is shown in Figure 5-1.
A new scalable hidden surface removal (HSR) is introduced, which not only inherits all the virtues of the current HSRs, but also takes into account the psycho factor, that is the limitations in the human viewing system. I address the problem occurring while applying SIRDS to moving 3D scenes. Three solutions are proposed to alleviate the problem. With statistic tests on volunteers, I point out some crucial issues for generating autostereograms with less risk to artifacts.

According to the conventional implementation of SIRDS [3], [5], [9], [12], the portrayed 3D object is located between two imaginary planes, namely the Near plane and Far plane as in Figure 5-2. Every point $A$ on the surface of that 3D object is projected to the image plane IP (the plane of the final autostereogram image) with two rays, $S_1A$ and $S_2A$ going into the two eyes ($S_1$ and $S_2$) of a virtual viewer. The incident points marked with Left image and Right image in Figure 5-2, as well as the distance ISS between them take an important role in the visibility of the true 3D point $A$ when the autostereogram is viewed by someone.

In order to calculate the position of Left image, Right image and ISS, two conditions are defined in the conventional implementation, as the followings:

1. The Far plane is positioned behind the IP at the distance $D$ as far as the eyes of a virtual viewer are in front of the IP.
2. The eyes $S_1$ and $S_2$ of the virtual viewer are floating over the IP so that for all possible positions of points $A$, the connected triangle $S_1AS_2$ is always isosceles and the plane of this triangle (referred to as Calculating Plane, CP) is constantly horizontal.

Figure 5-2: Geometry on which the conventional SIRDS algorithm is based.

All the pairs Left image and Right image, calculated for every point $A$ of the 3D object being portrayed, expose constraints on assigning texture to pixels of IP to construct the final autostereogram: they must have the same intensity / color because they are the two images of the unique 3D point. For distinguishing, I call this type of constraint pixel constraint (PC).

In some cases, the role of PCs can be omitted due to the occlusion caused by the transition between the surfaces of the 3D objects. Figure 5-3 denotes such situations. Although the pair $x_1$ and $x_2$ participates in a PC of the 3D point $A$, the color / intensity at these positions are not necessarily identical. That is because the point $3D$ is not visible to the virtual viewer: the ray $AS$ is interrupted by a portion of the 3D surface. Hidden Surface Removal (HSR) is therefore specified in the conventional SIRDS algorithm for recognizing such invalid PCs.

Figure 5-3: Geometry for hidden surface removal.

**Thesis 1.1:** By gradually lifting off the conditions in the original algorithm, I generalized the calculation of ISS and HSR.

By retaining the floating eyes supposition [J2], [J4], I deduced the calculation of ISS and HSR for an arbitrary position of the Far plane behind the IP as the following:

$$
ISS = \frac{(1 - \mu \cdot z) \cdot E}{1 + a - \mu \cdot z}
$$

(5.1)
Point A = \[
\begin{cases}
\text{Visible if } \forall i \in \{1, 2\} z_c < z_s + \frac{2 \cdot t \cdot (1 + a - \mu \cdot z_s)}{\mu \cdot E} & \text{otherwise}
\end{cases}
\] (5.2)

The double precision parameter \(a\) is introduced to present the proportion between the distance of the Far plane behind the IP and of the eyes \(S_1, S_2\) in front. It implies the arbitrary location of the Far plane after the IP.

By removing both two conditions [J1], [C11], [C13], I developed the formulas for ISS and HSR in a normal viewing situation of an autostereogram, based on the 3D Cartesian coordinate system, as follows:

\[
\begin{align*}
 x_{\text{left}} &= \frac{X_{\text{ISS}}}{E} \cdot y + \left(\frac{1}{E} - \frac{X_{\text{ISS}}}{E}\right) \cdot x ; \\
 x_{\text{right}} &= \left(\frac{X_{\text{ISS}} + E}{E}\right) \cdot y + \left(\frac{1}{E} - \frac{X_{\text{ISS}} + E}{E}\right) \cdot x
\end{align*}
\] (5.3)

Point A = \[
\begin{cases}
\text{Visible if } \forall i \in \{1, 2\} z_c < z_s + K_i \cdot \frac{(1 + a - \mu \cdot z_s)}{\mu} & \text{otherwise}
\end{cases}
\] (5.4)

where ISS can be reused from formula (5.1) and

\[
K_i = \frac{\left(\frac{x_{i\text{left}} - x_{i\text{right}}}{}\right) + \left(\frac{y_{i\text{left}} - y_{i\text{right}}}{}\right)}{\left(\frac{X_{i\text{left}} - x_{i\text{right}}}{}\right) + \left(\frac{Y_{i\text{left}} - y_{i\text{right}}}{}\right)} \quad i \in \{1, 2\}
\]

The meanings of the symbols used in formulas (5.3) and (5.4) are described in Figure 5-4. The 3D object being portrayed is marked with SO (Solid Object) in the figure.

dimensional” view, where only the relative arrangement between 3D objects (therefore only half dimension more) is required and there is no way to transmit all data for full three dimensional reconstruction.

Video is always the most important component of a multimedia content. It is not only the most attracting, enriched resource but also the most bit-consuming information. Video compression hence plays a crucial role in the feasibility of multimedia applications. It challenges researchers to find more and more efficient solutions. The new IMVO scheme in the dissertation also contributes to this final goal. On one hand, the scheme deals with video data as an independent entity, therefore the output of the codec — the object-based video can be easily reused in a composing process to construct a complex multimedia content. On the other hand, the IMVO scheme is fully specified: it includes object-based video compression and also a specific solution for object extraction from the conventional frame-based video capturing systems. Therefore the scheme can be directly used by the content-provider without any external preprocessor. The interface between the functional units in the scheme is clearly defined for the individual adaptation of each unit to satisfy the predefined requirements. The scalability of the scheme widens its applicability in various situations of the communication channel.

The outcome related to MPEG-4 BIFS in the dissertation is intended to participate in the latest issue in multimedia research: the capability of composition and interactivity. The structure of the proposed authoring tool can be further improved in implementation. As a result, the authoring tool becomes more and more convenient for creating complex multimedia content, which can highlight the interactive feature supported by MPEG-4. The modification proposed to the structure of BIFS in the dissertation adds enhanced services to MPEG-4, so that a complex multimedia scenario like games can be implemented with the light layer of system information in MPEG-4: the BIFS layer. MPEG-4 is an international open standard, MPEG-4 based games (applications) can widely conquer the market with various scenarios of terminals, such as personal computer, set-top box, Personal Digital Assistant (PDA), mobiles, etc. Improvement for creating sophisticated multimedia content with MPEG-4 just opens huge perspectives of its utilizations in the near future.

Acknowledgement

This dissertation could never have been written without the support from a large number of people. I am especially grateful to the following persons:

My parents and my sister, for their endless support and encouragement. By bringing the dissertation work to the end, I would like to express my deepest gratitude for all they have done for me.

My wife, Quynh Phuong Tran, for love and happiness, for constantly reminding me that there are other important things in life.
Table 5-1: Compatibility between MPEG-4 BIFS and high level managing classes which build up Flash application.

<table>
<thead>
<tr>
<th>Classes in high level manager of Swiff file (storage of Flash)</th>
<th>MPEG-4 BIFS structure</th>
</tr>
</thead>
<tbody>
<tr>
<td>HFMovie</td>
<td>SFTopNode: Group, Layer2D,…</td>
</tr>
<tr>
<td>HFFrame</td>
<td>Temporal BIFS commands construct the scene at a time.</td>
</tr>
<tr>
<td>HFObject</td>
<td>SFWorldNode: general collection of nodes.</td>
</tr>
<tr>
<td>HFSound</td>
<td>SFAudioNode: AudioBuffer, Sound2D,…</td>
</tr>
<tr>
<td>HFBButton</td>
<td>SFGeometryNode (Circle, Rectangle,…)+ SensorNode</td>
</tr>
<tr>
<td>HFShape</td>
<td>SFGeometryNode: Circle, Rectangle,…</td>
</tr>
<tr>
<td>HFRectangle</td>
<td>Rectangle</td>
</tr>
<tr>
<td>HBBBitmap</td>
<td>SFTexture: ImageTexture, MovieTexture,…</td>
</tr>
<tr>
<td>HFCircle</td>
<td>Circle</td>
</tr>
<tr>
<td>HFPolygon</td>
<td>IndexedFaceSet, IndexLineSet, …</td>
</tr>
<tr>
<td>HFOval</td>
<td>IndexedFaceSet, PointSet,…</td>
</tr>
<tr>
<td>HFFont</td>
<td>FontStyle</td>
</tr>
<tr>
<td>HFText</td>
<td>Text</td>
</tr>
<tr>
<td>HFEditionText</td>
<td>Text</td>
</tr>
<tr>
<td>HFAction</td>
<td>Script</td>
</tr>
</tbody>
</table>

6 Applications of the results

The recent explosion of multimedia and its related issues has posed numerous new challenges to researchers and developers. The results obtained in the research of this dissertation are connected to the most attracting and considerate issues of the current research on advanced multimedia.

Viewing a multimedia content in the natural 3D fashion — the most realistic and also the most expensive way — is still an endless struggle for researchers. A comprehensive research on SIRDS mentioned in the dissertation gives a better understanding of the key-technique of the SIRDS algorithm, a very efficient method for reconstructing depth information. The generalized algorithm and the evaluation of different combination of parameters used in the algorithm can be exploited to create an autostereogram with better visibility and depth-perception. The results I obtained while applying SIRDS to moving scene can be used as a first step for further improvement in the motion domain of autostereograms. The information stored in PCs can be adopted for reconstructing a “two and half
1. If any two 3D points $A$ and $B$ of the SO have common \textit{Left image or Right image} on IP ($A_{l/r}=B_{l/r}$), and

2. If they are viewed in a smaller angle (angle $\angle AS_B$) than a predefined threshold (controlling the scalability),

then the points $A$ and $B$ can be considered to obscure each other. The point having shorter ISS covers the other.

When applying SIRDS to a moving 3D object, a new type of constraint must be taken into consideration. I call it Color Constraint (CC) to distinguish it from the PC discussed above. With CC, the temporal consistence in the texture of the 3D object is ensured.

**Thesis 1.3:** [J1], [C11], [C12] I showed that when generating moving SIRDS, one can not retain CC without violating PC and vice versa. I proposed three solutions for solving the compromise between CC and PC at different levels as follows:

1. \textit{CC-based generation}: CCs are mainly taken into account when there is a conflict between CC and PC.
2. \textit{Static background-based generation}: CCs are retained periodically at the cost of broken PCs.
3. \textit{Cost-based generation}: CCs’ rules overwrite PCs so that they minimize the cost function, whose value is proportional to the number of current broken CCs and to the inverse distance from the pixel with broken CC to the currently examined one.

The last method offers the best solution in finding compromise between PC and CC. The distribution — controlled by the cost function — can be scattered apart. It reduces the noise in the texture of moving objects throughout frames.

With the assistance of proper authoring tools, MPEG-4 BIFS can take over the functionality of any third-party multimedia content generating tools, such as Flash technology. As part of the dissertation work [C1], [C2], I proved the compatibility between the two technologies in composing a complex multimedia content. The proof is not derived from either the current support of authoring tools or the currently available applications. I pointed out the correspondence between them in their building elements, the root of the associated technology. Table 5-1 summarizes this close relationship.

**Thesis 3.2:** [R1], [C1] I detected the lack of temporal constraints in the structure BIFS while creating a sophisticated multimedia content, such as gaming application. I proposed two alternatives to get over this problem.

I proposed a new node, namely \textit{Cookies} node to BIFS structure. The philosophy of \textit{Cookies} node emerged from the functionality of Cookies in HTML (HyperText Markup Language) technology: maintaining some client-based information for later access to the same portion of a Web server. Currently, the service provided by BIFS can be considered as those provided by dynamic web pages, then the similarity can be easily derived (the node is named after this fact). \textit{Cookies} enhance the functionality of interaction in a scene / web page by means of retaining the temporal states, the conditions for reusability in future access / handling the same scene / web page.

---

Figure 5-6: Delivery of broken CCs (black pixels) produced by cost-based generation.

Figure 5-11: Functionality of authoring tool.
Figure 5-10 shows a so-called Pinball game, in which MPEG-4 BIFS is used as a “programming language” to provide the environment of the game. Some active elements of a video content are available in MPEG-4, which were impossible in the predecessors of the MPEG family. For instance, there is a button-like text, which is “clickable”; there is a movable shape (racquet), whose position can be set by viewer, and so on. Unfortunately, this new capability of MPEG-4 is still not exploited in the current MPEG-4 based applications.

**Thesis 3:** I designed a new model for user-friendly manipulation of the scene composition and interaction between users and multimedia content. I pointed out some shortcomings of the current structure of BIFS. Proposals for overcoming these limitations are also presented.

**Thesis 3.1:** [C2], [C3] I showed that with high level abstraction authoring tool, BIFS structure becomes friendlier for the scene-designer. The structure of such authoring tool is proposed to reduce the programming work of the scene-designer.

High level authoring tools should make the complex structure of BIFS nodes transparent for the scene-designer. Besides multiplexing several media data, the authoring tool modeled in the dissertation contains two more layers: Graphics Interface Layer and Scene Node Layer (Figure 5-11). The former provides familiar components of a multimedia scenario for the scene-designer, such as audio, video, graphics, images and text. However, they are enhanced with properties supported by BIFS. For instance, graphic behaves as a button for interaction; video audio data are constrained in start-time and stop-time, etc. The lower layer — Scene Node Layer — translates these high abstraction components and their settings into binary information in the form of BIFS nodes. Therefore this layer should be covered as much as possible from the scope of the scene-designer.

### 5.2 The new IMVO codec scheme

The recent MPEG-4 standard offers numerous advanced tools for encoding multimedia content. Among many others, object-based video encoding proves to be an efficient way for encoding and manipulating video components in a complex multimedia scenario.

**Thesis 2:** [J3] [C9], [C10] I proposed a new scheme for encoding a video object with arbitrary shape. A joint name of the scheme is Integrated Mesh Video Object (IMVO) codec, which is a full chain of functional units, starting from conventional frame-based video data, through object detection and encoding processes. The scheme contains a linked chain of functional units — black boxes. The interfaces between them are clearly specified, ensuring the upgrade or replacement of individual units. The implementation of the central black boxes are presented in detail with optimized solutions.

**Thesis 2.1:** I specified the structure of a full codec scheme for video object with arbitrary shape. The scheme ensures the template design, which means that the scheme itself is an applicable process, but it guarantees further optimizations to satisfy various requirements of actual applications.

The IMVO codec model [J3], [C9], [C10] has the following properties:

- It is a joint solution of two encoding methods specified in the MPEG-4 standard: animated mesh for standstill texture and video object with arbitrary shape.
- It is a full chain of functional units, which is directly applicable to a normal frame-based video capturing system.
- It is a semi-automatic video object detection, which means human interaction is involved only once, at the beginning of the video sequence.

The components of the scheme can be summarized as in Figure 5-7. The encoding process is actually performed in the block **Mesh-based encoder** of the figure, which deploys several techniques designed in the framework of my dissertation.
In Mesh-based encoder, textures of VOs are encoded in a new motion compensation fashion [3], [9], [10]. Nodal backward motion vectors (BMVs) are encoded in a modified predictive manner. They are interpolated at decoder-side to produce a dense field of BMVs. The results of interpolation are then used to find the prediction values for texture of the VOs. The final texture is fully reconstructed with the assistance of residual errors (REs), which are also encoded and sent to the decoder-side. It is worth noting that all motion vectors in the scheme are given for every vertex of the mesh and not for macro block as in MPEG-4. It expresses the best the combination of 2D mesh animation and motion compensation defined in MPEG-4.

Figure 5-7: Functional units of IMVO scheme.

An important encoded target of the Mesh-based encoder is geometry data, i.e. the structure of the meshes representing VOs. I proposed optimal graph traversal implemented with Prim algorithm [3], [9], [10] for ordering the vertices of these constrained Delaunay meshes. This traversal is processed prior to encoding the geometry information for saving bitrate. That is, while encoding nodal positions of these meshes, instead of traversing along the internal vertices and encoding their difference in distance as in MPEG-4, the IMVO scheme reorders the vertices with respect to the optimal graph traversal, whose cost function is proportional to the distance between the vertices. The gain in bitrate for several intra frames is shown in Figure 5-8.

Figure 5-8: Effect of Prim algorithm (implementation for optimal graph traversal) on geometry compression of the Interview sequence.

Figure 5-9: Overall performance of IMVO codec scheme.

5.3 The exploration of MPEG-4 BIFS

The MPEG-4 is the first standard which addresses the scene composition and enables the interaction-support in multimedia content. These revolutionary features are available thanks to introducing a new type of data: BInary Format for Scene