Authoring of Visual Information in Multimedia

Dr. D. B. Lowe, A/Prof A Ginige
School of Electrical Engineering,
University of Technology, Sydney

Abstract:

Multimedia involves the use of multiple forms of communication media in an interactive and integrated manner. In current multimedia systems the interactivity is predominantly derived from the textual information, rather than other media such as images, video or audio. Multimedia is unlikely to achieve its full potential until all media can actively support the full level of functionality expected - including visual information. Part of the process of achieving this goal is to improve current techniques for authoring these additional media in more sophisticated ways. This paper looks at current research into the authoring of visual information and ways in which this information can be more tightly integrated into multimedia applications.

Introduction

Multimedia is a rather eclectic technology, drawing on a number of enabling technologies - information theory, man-machine interfaces, information technologies, and database handling, to name but a few. These technologies are combined to create an application which purportedly integrates a number of different media into an interactive whole. Note the key words in this description: integrates, media, and interactive. To date most multimedia systems have not successfully reached this goal. Although they tend to be interactive and integrated, and include multiple forms of media, these media are not combined into a cohesive whole - i.e. all forms of media are not used to provide the interactivity. In fact most existing multimedia applications would be more appropriately called multiple media hypertext systems. They tend to be hypertext systems (i.e. the textual information provides the interactivity) with the additional media acting as annotations, but not truly integrated into the application.

The few minor exceptions to the lack of visual information interactivity predominantly revolve around hand-crafted small scale applications. These applications are however sufficient to illustrate the significantly enhanced functionality which can be achieved with active visual information. For a good example of this, look at the map images in the Virtual Tourist on the WWW (http://wings.buffalo.edu/world/). These small scale applications have tended to be hand-crafted using low-level tools (such as MapEdit - a shareware Unix tool) and contain hard-coded static syntactic descriptions of the image data. Indeed, we are not truly interacting with the visual information, but rather interacting with an underlying syntactic description of the image (which provides us with the illusion of interaction with the visual information).

We can understand the poor use which is made of visual information better by considering the use of textual data within multimedia, an area that has received (and continues to receive) considerable research attention. The text can be stored, analysed, manipulated, and generated synthetically. Essentially the text can be treated as consisting of lexical components or discrete entities - hypercomponents - (words, sentences, paragraphs, etc.) which obey a series of syntactic and semantic rules describing the inter-relationships (Ginige, 1995). Within the multimedia application these hypercomponents can be used to create nodes, anchors, links, etc. These elements provide the navigation functionality which supplies the interactivity - the core of any multimedia application. A growing number of authoring tools exist which assist in the conversion of textual information into an appropriate structure (Robertson, 1994).

The evolution of visual information is still at a much lower level, and is predominantly treated as a passive media. Visual information in its raw form is highly unstructured and yet very commonplace (documentary video tapes, image and photographic collections and databases, etc). Many visual information applications
such as medical imaging, robotics, and interactive multimedia) make use of the visual information in a highly structured format. Conceptually visual information can be treated in the same way as text; consisting of discrete entities which obey certain syntactic rules. The primary difficulty however lies in identifying these entities and the associated rules, and then interpreting these. When this is achieved visual information can become an active media as powerful (and in many cases, more so) than textual information. It is this premise which has driven this work.

Although there has been considerable work on structuring textual information, there has been minimal work performed on structuring the visual information to suit multimedia applications. For example, the development of practical multimedia systems requires the use of both suitable information and the appropriate structuring of this information. This information structuring is critical to the development of high-quality visual information applications. One of the major obstacles hindering the advancement and commercial acceptance of these applications is the cost of structuring the vast amount of visual information. At present most existing visual information databases have been handcrafted - using the human operator to perform the lexical decomposition of the visual information; a process which is excessively expensive and time consuming.

This paper looks at current research into the authoring of visual information and ways in which this information can be more tightly integrated into multimedia applications. Current techniques range from completely manual mark-up of the visual information (a process which is excessively expensive and time-consuming for large databases) to the use of object recognition schemes (which is only practical for highly constrained problems). We describe a compromise between these extremes - providing intelligent assistance to the author in the identification of prospective syntactic components of the visual information.

Multimedia Systems and Interactivity

Interactivity

Multimedia is a technology which has been enjoying significant attention within the last few years. Multimedia involves the use of multiple forms of communication media (such as text, audio, video, images, etc.) being used interactively. At present the most common forms of multimedia applications are surrogate travel, telemarketing, telemedicine, entertainment and interactive learning. These applications can all benefit significantly from the extensive use of a number of different media, including text, audio, and particularly visual information.

One of the key elements of any multimedia application is interactivity - the ability to interact with the media in such a way that improves the accessibility, useability and presentation of the information. The level of interactivity (along with many related factors, such as the appropriateness of the links) will greatly affect the overall success of the application. The degree to which we achieve interactivity will be strongly related to the way in which we model and analyse each of the media.

We can identify two main factors which influence our ability to achieve appropriate interactivity. The first is the ability to structure the information in such a way that appropriate sections of the information can be specified and utilised - e.g. the identification of key words or phrases in textual information. The second factor is our own ability to interpret and utilise this information - e.g. how easily we can locate a specific textual phrase from a page of text. We will discuss both of these factors.

Media Structuring

If we are going to interact with various media then we need to be able to identify structure within the media to support the interactivity. This is typically part of the purpose of the multimedia authoring process - to identify the relevant inherent structure and make this explicit in a way which can be utilised by the multimedia applications. For example, with text we may identify relevant words or phrases which can be linked to related sections of information.
Forms of electronic information such as text, numerical data, etc. are highly structured. This structure (or more correctly, the ability to identify this structure) is typically critical to the effective use of this information in supporting interactivity. For example, in multimedia systems text is typically stored, analysed, manipulated, etc. In order to achieve this, applications (or authoring packages) typically structure the textual information into ‘nodes’ (where each node is typically information on a specific topic). Each node consists of discrete entities (words, sentences, paragraphs etc.). Appropriate entities can be used as specialised components. For example, a word can act as an anchor point for links to other nodes (e.g. if the user clicks the mouse on the word, the link is traversed, and the destination node is displayed). The effective use of this textual information relies strongly on the ability to make explicit this inherent structure of the textual information. This is also true for most alternative uses of similar information.

Recent authoring frameworks have begun to make this structure more explicit. For example, our own work on MATILDA (Lowe, 1996) separates the information domain from the application domain, and then within the information domain makes explicit the various levels of information structure. These levels include lexical (i.e. identification of information components), syntactic (structural relationships between these components), and semantic (meaning-based relationships between the components).

Despite these advances in the concepts of media structuring, the evolution of visual information (such as images and video) is still at a much lower level. In general it is either poorly structured, or the structure is poorly understood. Current techniques are relatively primitive in terms of their ability to make the structure of visual information explicit. As a result, in traditional multimedia systems visual information is predominantly treated as a passive media - essentially acting as an annotation to the textual information. This is illustrated in figure 1.
A number of (increasingly sophisticated) tools exist which assist in the process of structuring the raw textual information into a form suitable for multimedia. This typically requires the partitioning of the information into appropriate nodes, the identification of the hypercomponents, and specifying appropriate links. Recent authoring tools, such as HART (Robertson, 1994) have begun to focus on providing a greater degree of support for the author during this process. Support is typically provided through both procedural guidance (assisting in the process of converting the information) and intelligent assistance (providing context-dependant choices for the author - such as identifying and suggesting appropriate keywords from the text). To date, these tools still focus almost entirely on textual information. At present there is minimal support for assisting the authoring process of visual information. By improving this support the development of visual information applications becomes much simpler, and available to a much larger audience.

Identifying Information

Once the information has been structured it still needs to be utilised. The information needs to be accessed, interpreted, searched, manipulated etc. The Human Visual System (HVS) provides a much higher information transfer bandwidth for visual information than is typical of other media (audio, interpretation of text, etc.) - i.e. we have the ability to assimilate and interpret visual information much more rapidly than most other forms of information. This implies that for many applications significant benefits can be gained from the use of active visual information. In particular it can be expected that the resultant application will have improved functionality, and allow significantly enhanced information access. A number of existing hand-crafted applications illustrate this quite well. Examples on the WWW include:

- The Virtual Tourist on the World-Wide-Web (http://wings.buffalo.edu/world/)
- Map of the Berkeley Campus (http://superior.berkeley.edu/Berkeley/campus.html)
- Experimental Internet Resources Metamap (http://www.ncsa.uiuc.edu/SDG/Software/Mosaic/Demo/metamap.html)
- National Center for Supercomputing Applications (NCSA) (http://www.ncsa.uiuc.edu/General/NCSAHome.html)

Achieving Active Visual Information

Current Practice

Once we recognise that active visual information can significantly improve multimedia applications we need to consider appropriate mechanisms for achieving this. The volume of research that has been performed on the role and representation of image data within multimedia systems is relatively small. The primary research focuses for visual information in multimedia have been on information management, data modelling, database issues and standardisation issues (Wong, 1992). In general image and video data, when used, has been stored in the form that was most convenient given the hardware and software tools provided: tools that were typically not designed with multimedia in mind. In many applications, hardware that was developed for alternative imaging applications has been adapted for use. An example of this is the use of the JPEG coding scheme to code and decode the image data. Almost all these schemes suffer from the same limitation; the image representation is not particularly appropriate for a multimedia application. Edgar, Steffen and Newman (Edgar 1992) look at current technologies in this area and the effect that these have had. This includes advances in image handling capabilities, the development of image compression standards, new storage techniques and network applications. The more fundamental issue of the image representation and structuring itself is however not raised.
If we consider the critical aspect of image structuring, then we can identify several approaches. We will briefly discuss a few of these before considering what we see as an acceptable compromise given the current technological limits.

**Manual Image Markup**

In order to adapt image data to a multimedia environment and make the visual information truly active, the usual techniques to date have involved adding additional information to the representation without modifying the underlying structure. An example of this approach is manually specifying the location and orientation of objects within the image (for example, drawing around them using the mouse - cf MapEdit, a public domain Unix utility - see Figure 3) and storing this data with the image to create a 'marked-up' image (Constantopoulos, 1991). This approach has been used in a number of cases with WWW pages which rely on map files which contain descriptions of specific shapes (polygons, squares, circles, etc) within the image and the associated URL to act as the destination of a link from the shape.

This approach has several significant problems. Firstly it means that we are actually interacting with a map file, rather than directly with the image (though this does provide the illusion of interaction with the image). This means that, since we are not interacting with the image, we have the additional problem of ensuring that we maintain consistency between the image and the map. Any changes to the image must be reflected in changes to the map file. More significantly, the effort required to manually mark-up images will be excessive in all but the simplest cases. This situation will be significantly exacerbated in the case of video information where a typical 10 minute video might contain 18,000 frames. At present this has meant that only the simplest cases of visual information has been marked up, and typically only in small isolated cases. Nevertheless, this work does indicate the validity and improved functionality which can be obtained from the use of active visual information.

**Textual Annotation**

An alternate approach to the use of manual markup techniques is to manually add a textual description of the image. This description can then be processed along the same lines as normal textual information. This approach has been used successfully for retrieval in image databases (Al-Hawamdeh, 1991), though this approach would be less useful in situations requiring knowledge of the positions of objects within the scene.
Both manual markup of images and textual annotations are time consuming, cumbersome, and prone to error - especially for large image sets, as each image needs to be independently manually annotated. The use of textual annotations obviously presents the problem that the user cannot interact directly with the visual information, and must instead interact via text.

**Object Recognition**

The major problem with the approaches described above lies in the fact that they are exceedingly manual - requiring considerable time and effort. When structuring text-based systems, assistance is provided by automating the process of extracting words, sentences, paragraphs, etc. If we are to achieve a similar level of support for visual information then we need to automate the process of extracting objects from the visual information, and then identifying suitable objects to be used in the structuring process. The most obvious (and currently the most powerful) method of providing this assistance is to use object recognition schemes. In order to investigate the applicability of this, we previously developed a demonstration application which integrated a simple object recognition scheme into a multimedia application.

**Application Description**

The multimedia application which was developed, although very simplistic in comparison to existing multimedia applications, demonstrates one possible technique for enabling the use of active image data rather than passive image data, as well as providing a platform for investigating the general performance and change in functionality. No attempt was made to create a commercial quality multimedia application, as the primary aim was to investigate the implications of the imaging principles involved, rather than the various multimedia principles. As a result the overall operation and performance of the system from a multimedia point of view is quite simplistic. Only sufficient functionality to investigate the image algorithms was incorporated. The development and usage of this application are detailed elsewhere (Lowe, 1993).

The recognition process which was used (Lowe 1992) formed the central core of the multimedia application. In previous work (Lowe 1990) a representation was developed which decomposed an image into a number of hierarchical information layers. These layers contained image features such as edge information (and other

![Figure 4: Snapshot of screen containing the HyperImages application: A link which has an object from within an image as an anchor has just been triggered and is about to be traversed.](image-url)
primary image discontinuities), texture, shading, and colour. These layers of information were used to perform an object recognition scheme based on hypotheses generated from feature matches. The multimedia application was developed on a Sun SPARCstation running OpenWindows version 3 with an X-Windows interface, and was coded in such a way as to make the user interface as simple and logical as possible. The images are processed offline to extract the information hierarchy. They are then loaded in when required for a particular database element. The display routines decode this data and regenerate the images progressively so that the user has an indication of the image content as early as possible. If the user then selects an object within the image (by clicking on it with the mouse) then the recognition scheme will attempt to recognise the object using the information representation rather than the having to regenerate the canonical image. If the recognition is successful the application will jump to an appropriate database element (i.e. image, text and audio that is relevant to the selected object).

Figure 4 is a snapshot of the screen containing the multimedia application. The user has just moved the mouse over an object in the image and pressed the left mouse button. The object was identified and the appropriate link will be subsequently traced. This application did not attempt to address any of the broader issues of information handling, database management, etc. The sole purpose was to investigate one method for automating the process of generating active image information in a multimedia application, and to then consider the implications of this in terms of the effective change in the application functionality.

**Results of Using Object Recognition**

After the application was developed and a typical database installed, the system was evaluated. This evaluation included both a consideration of authoring issues, and useability issues. The useability encompassed both a subjective users point of view, as well as quantitative measurements of the performance. The process of authoring the database required the author to manually generate the models used for the object recognition. Once this step was completed the author did not need to consider the visual information again. The application which was developed handled the creation of automatic links. The most difficult part of the authoring process was therefore the generation of the object models. This required the specification of a wireframe model of the object and detailing the surface shading, colour and texture. It was found that for this approach to be of practical use, this process would need to be automated. The model specification required a comparable effort to the manual identification of the objects (i.e manual authoring). The primary difference between a manual authoring approach and this automated approach is related to ongoing effort. The effort required is related to the number of objects to be identified, rather than the number of images in the database. Once the models have been created, additional images can be added to the database without requiring any additional authoring effort (apart from any data capture and conversion which is required).

Once the application was completed and the demonstration database installed a number of people were asked to use the system and comment on its useability. The most significant conclusion drawn from this was the effectiveness of the improved functionality of the resultant system. In general the users seemed to be quite impressed with the general idea of active image data. In general the users found it much easier to locate and select an object within an image, than a word within a document.

**Using Object Recognition for Achieving Active Image Data**

Although the application which was developed had a high level of success, this was rather artificial. It certainly illustrated the appropriateness of using active visual information. However it should be recognised that the visual information within the application was limited to relatively simple objects. Every object was rectilinear, relatively simple in shape, and had a simple shading, texture, and colour. It was only because the objects were so simple, that the object recognition scheme had such success in correctly identifying objects.

For an object recognition scheme to be effective in implementing active visual information in multimedia it needs to satisfy at least two criteria. Firstly it must be robust, reliable, and consistent for a very wide range of applications and objects. Secondly, it must be very straightforward to expand the object database which it uses to identify objects. The object recognition scheme used for HyperImage satisfies neither of these criteria. Much research is occurring in the field of computer vision, and great success has been achieved in restricted application domains and for restricted image sets. Nevertheless, a general object recognition scheme which could handle the broad range of visual data present in multimedia applications is likely to be a considerable distance off. In the long term, object recognition will become increasingly important in multimedia authoring
and multimedia applications (as has been foreshadowed by experiments such as HyperImage). In the shorter term however, object recognition would appear to be insufficiently mature, except in perhaps very isolated cases.

A Viable Compromise for Achieving Active Visual Information

If we wish visual information to be widely used as an active media in multimedia applications then we need to improve our ability to handle visual information during the authoring process. As mentioned above, a number of multimedia authoring tools provide assistance for identifying the structure of textual information. This includes identifying the individual hypercomponents within the media. For text, this involves extracting words, sentences, paragraphs, and other suitable blocks of text. For images this should involve identifying regions (foreground, background, surfaces, etc.), objects, and other relevant information. For video this involves identifying episodes, shots in addition to foreground, background and objects. In order to handle visual information effectively during the authoring process we need to investigate methods for providing this assistance.

Having accepted that object recognition is likely to be too impractical for use for general multimedia applications, we need to consider possible alternatives. As was discussed above, information structuring can typically include lexical, syntactic, and semantic structuring. The lowest level of this - lexical structuring - involves identifying the lexical elements within the media. This does not require knowledge of the meaning of the information, or even true object recognition. Lexical decomposition is predominantly a segmentation problem - where we wish to identify things such as regions of images and boundaries between scenes in videos. Considering this view, and recognising that we cannot completely automate the authoring process, we can look at the possibility of semi-automating the process.

In order to provide an effective analysis of the visual information in multimedia, we can either restrict the visual information - which is what we would need to do if we were using object recognition, but is impractical for general authoring applications - or restrict the analysis which we are performing. Authoring assistance can be provided by combining appropriate analysis tools with the interaction of the author to guide the analysis.
For example, assistance can be provided by using analysis tools (such as segmentation) to assist the multimedia author in identifying possible objects. The author will interact with the analysis tools, guiding them where necessary, but providing the necessary control. The analysis tools are used essentially to provide assistance to the author, rather than performing the entire authoring process. The authors current research is following this path - investigating methods of using image analysis tools to semi-automate the authoring process which has previously been performed by hand.

The most substantial example of similar research is the QBIC project (detailed on the QBIC WWW home page - http://wwwqbic.almaden.ibm.com/~qbic/qbic.html) where image analysis tools have been used to support image retrieval based on image content. The QBIC system includes various tools which can be used to identify regions of images. For example, a drawing tool allows the user to select a region of an image and then have the drawn line snap to the nearest boundary (using snakes). Although these tools are not being used for multimedia authoring of visual information (rather they focus on image databases), they have illustrated the power of image analysis tools in visual information handling.

We are currently investigating the inclusion of similar tools into a multimedia authoring system. Figure 5 shows the MATILDA information structuring package with an image loaded into the lexical structuring tool (The MATILDA information structuring system, and how these tools are used in the authoring process, is described in a companion paper in the same proceedings as this paper). At present the tool is still being developed. Over time we intend to incorporate progressively more sophisticated tools, providing an enhanced level of functionality to the authoring of visual information.

Summary

This paper has discussed the role which visual information can play in multimedia applications. For multimedia to achieve its immediate potential, the visual information needs to become truly integrated. This in turn requires that it become an active media. In order for a given media to be used as active data, the individual components of that media need to be identified during the authoring process. Traditionally, image data is manually marked-up during authoring. For large image databases - which will become increasingly common - this is excessively expensive and time-consuming. Object recognition schemes have shown the gains which can be obtained through automation of the handling of visual information. Unfortunately computer vision is still too immature to provide a sufficiently high level of robustness and accuracy for multimedia authoring, except for very specialised applications. We can nevertheless develop tools which will semi-automate the authoring process. A typical example would be an integrated tool, based on image segmentation, which makes suggestions to the user regarding possible objects within images. The user would have control over these analysis tools, which act to assist rather than replace the multimedia author.

The implementation of the active image media is the first phase of what the authors see as the development of the ability to treat all media generically. Thus the application becomes independent of the particular media type. Each media type allows static and dynamic links from its various semantic components, contextual analysis etc. - the range of options now available with textual media. This research has tried to begin the work on treating image data in the same fashion as textual data.

Acknowledgments

The authors would like to thank both the Australian Research Council, and the University of Technology, Sydney for supporting research which has contributed to many of the ideas outlined in this paper.

References


Contact Details

Dr. David Lowe
Senior Lecturer, Computer Systems Engineering
University of Technology, Sydney
P.O. Box 123, Broadway, 2007, NSW, Australia
Tel: +61 2 330 2526
Fax: +61 2 330 2435
Email: dbl@ee.uts.edu.au

Associate Prof. Athula Ginige
Associate Professor, Computer Systems Engineering
University of Technology, Sydney
P.O. Box 123, Broadway, 2007, NSW, Australia
Tel: +61 2 330 2393
Fax: +61 2 330 2435
Email: athula@ee.uts.edu.au