Web-enabled Advanced Multimedia Systems

Alfons Salden*, Frank Aldershoff*, Sorin Iacob *, Raymond Otte* and Menzo Windhouwer†

*Telematica Instituut, Drienerlolaan 5, 7500 AN Enschede, The Netherlands
Telephone: ++32-53-4850485, Fax: ++32-53-4850400, Email: salden@telin.nl

†CWI, P.O. Box 94079, 1090 GB Amsterdam, The Netherlands
Telephone: ++31-20-5924308, Fax: ++31-20-5924199, Email: Menzo.Windhouwer@cwi.nl

***Abstract*—**

Most of the multimedia objects distributed over the World-Wide Web are unstructured or poorly meta-indexed to be of any use in retrieval tasks formulated by users in natural language queries. In general these dynamic multimedia objects are manually annotated in terms of textual documents. The high costs involved in manually indexing multimedia objects, which grow in volume and are becoming ever more diverse in type, call for automatic sustainable categorization schemata that are accesible and operational on the Web. These categorization schemata comprise indexing, querying and retrieval schemata. We propose a web-enabled advanced multimedia system as a solution to this categorization problem. We lay bare the physical, mathematical and logical framework underlying our system. We demonstrate that this system pays off especially in semantically user-defined summarisation tasks concerning multimedia presentations.

I. Introduction

Web-enabling advanced multimedia systems is one of the hardest problems in hypermedia engineering, because it concerns on the one hand an integral semantic representation, analysis, processing and understanding of various physical fields underlying dynamic multimedia objects. On the other hand one has to offer various indexing, querying and retrieval tools to make dynamic multimedia objects acccesible and available. Dynamic multimedia objects are becoming in addition ever more large in volume, type, and complexity. Furthermore, they are stored, distributed and active on the Web. Therefore, we solve in this paper the categorization problem of dynamic multimedia objects and demonstrate the effectiveness of our solution in an educational setting.

The Acoi system [1] provides a nice way to model, manage, query and retrieve multimedia objects on the Web through meta-indices stored in databases. Unfortunately, the feature detectors and grammars used in this system do not solve a generic multimedia categorization problem. The feature detectors and grammars are rather restricted to manually annotated and still very coarse multimedia meta-indices. Combining, however, the Acoi system with advanced multimedia systems [2], [3], [4] can automate and make explicit the feature detectors and grammars. Feature detectors and grammars are in the Acoi system context free implying that they are still not precisely defined. The advanced multimedia systems on the contrary are context dependent and do initiate those detectors and grammars. They invariantly and robustly categorize multimedia objects irrespective a so-called gauge group possibly covering degradations of the objects due to noise or morphological transformations. Instead of being based on heuristics, as in content-based multimedia systems, like QBIC [5], VisualSEEK [6], Virage [7], and VideoQ [8], our advanced multimedia systems are based on a generic physical, mathematical and logical framework. The physical framework is related to potentials and strengths of the physical fields. The mathematical framework is consequently related to connections and curvatures of the mulimedia objects corresponding to those physical fields. Finally, the logical framework that relies on computer algebra systems, like Mathematica together with Jlink and webMathematica [9], stipulates the (fuzzy) logical semantic web services needed to categorize multimedia objects.

Our paper is organised as follows. In section II we present the mathematical, physical and logical framework underlying advanced multimedia systems. In section III we elaborate on our web-enabled advanced multimedia system demonstrating its added-values in summarisation of multimedia presentations within E-learning contexts.

II. Advanced multimedia systems

Advanced multimedia systems, as proposed by Salden, Aldershoff and Iacob [2], [3], [4], overcome the problems of generic multimedia representation, analysis, processing and understanding. Instead of giving a full theoretical account of their solutions to these problems, we illustrate the main underlying concepts of our logical, mathematical and physical framework.

A. Multimedia representation

The representation of multimedia objects depends on the chosen (en)decoding schemata applied to the corresponding physical fields underlying e.g. audio, video and text. In this context it is important to know which abstraction schemata are used to resolve the physical fields, i.e. which (en)decoding schemata do exist to represent physical fields at a particular resolution. Furthermore, it is essential to know whether and how those representations change whenever other physical fields are also taken into account. What will be the (en)decoding

schemata if e.g. audio, video and text are coupled? Last but not least, what will be the (en)decoding schemata whenever the physical fields are subjected to gauge groups or even severe morphological transformations? Those gauge groups may coincide with deformational fields, whereas the morphological transformations may cover cutting, pasting, insertion and deletion of multimedia objects. If the gauge group is known, then a gauge invariant multimedia representation can be mathematically derived. However, such passive transformations are not common for multimedia objects that are subjected to small and even large scale perturbations. Active transformations such as morphological transformations may have far reaching implications on a semantic as well as contextual level, but may be undone by means of similarity operations inducing inference mechanisms on the multimedia objects. Thus our multimedia categorization problem involves besides the problem of invariance under gauge groups also the problem of robustness under similarity operations.

In Fig. 1 we apply various transformations to an input image, namely spatial deformations, shadowing and adding Gaussian noise. Apparently, humans are perfectly capable in detecting the contours of the visual object. We are also aware of the ambiguity of the visual stimulus. Multiple interpretations of the input image observed by our visual system is no surprise realizing that our brain is an inference machine inducing all possible connections if possible on the visual input to explore its meaning in various contexts.

Fig. 1. From left to right and from top to bottom: Input image of a vase/mirrored faces together with its spatially deformed, shaded and noisy version.

B. Multimedia analysis

The next phase in a multimedia categorization schema is multimedia analysis. This analysis boils down to identifying schemata that yield equivalences of the multimedia objects invariant under a gauge group (see also section II-A). Among such concise sets of equivalences of multimedia objects are geometric, topological and functional invariants that are either of a local, global or joint character. Furthermore, those equivalences may be differential and integral features, or even inductive and deductive inference structures on top of multimedia objects. The natural statistics of those inference structures subsequently underly the syntax, semantics and contexts residing in the multimedia objects. However, the capability of detecting perceptually salient multimedia structures using only (semi-)local information is limited. If the multimedia object is subjected to a well-defined gauge group we can derive theoretically the multimedia content. However, if the multimedia objects are possibly degraded by noise or even larger scale morphological deformations, then the effectiveness of such (semi-)local mulimedia analysis schemata breaks down.

In Fig.2 we demonstrate that topological edgels [2], [3] of the input image and its deformed and shaded version in Fig.1 are readily detected, whereas such a semi-local analysis at inner scale for the noisy version yields no useful contour information at all. The topological edge of the input image, its deformed and shaded version are equivalent under a gauge group that includes certainly deformations of the input image in a spatial as well as a dynamical sense. The question rises why this is not the case in the last image in Fig.1 that to us humans appears to be certainly as equivalent as the input image.

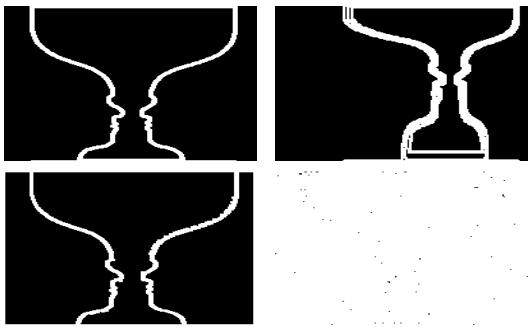

Fig. 2. Topological edges of images in Fig.1.

C. Multimedia processing

In order to ensure robustness and discriminative power of multimedia representation and analysis schemata some form of multimedia relaxation process becomes indispensable. Changing multimedia conditions the observed multimedia physical fields will be subjected to modern geometric, topological and dynamical perturbations consisting of non-integrable and integrable deformations of the physical fields. These multimedia deformations and morphological transformations lead to a change of the equivalences derived by means of multimedia representation and analysis schemata. Therewith our ”golden rules” used to categorize multimedia are under attack. We would like to forget about gory/minor details and even about large scale background fields that obstruct our comprehension of the perceptually meaningful semantics and contexts of multimedia objects on the basis of those golden rules. Lyapunov stability under noise and structural stability under severe morphological transformations can be guaranteed by multimedia

processing that is consistent with the underlying salient physical field dynamics. Note that, the latter salient field dynamics may comprise fields that are related to user behaviour and other dynamic network topology aspects.

In Fig.2 we demonstrate that the topological edge of the noisy version of the input image of Fig.1 can be detected and is up to some tolerance equivalent to those edges, found in Fig.2, for the input image, its spatially deformed and shaded version. Thus we conclude that in order to assure equivalence, discriminative power and robustness of multimedia representations and analyses under various types of transformations processing those objects at various scales is indispensable.

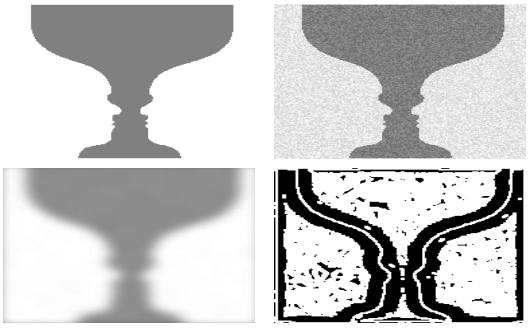

Fig. 3. Input image of a vase/mirrored faces, its noisy version, the nonlinearly filtered version of the noisy image and the topological edge of that image.

D. Multimedia understanding

Multimedia understanding in categorization schemata assumes knowledge of the multimedia consistent similarity operators and related recursion operators used in multimedia processing schemata and consequently in multimedia representation and analysis. Both these types of operators generate hierarchical nestings of gauge invariant and robust equivalences that might themselves be perceptually consistently grouped multimedia objects. These self-similar dynamic multimedia objects come about by segmentations and arrangements of dynamic scale-spaces of equivalences of the primal multimedia objects. Next an ensemble of inductive (multimedia statistics driven) or deductive (multimedia concepts driven) inference structures can be derived on top of those equivalences through combinatorics and enumeration (see for an extensive exposition [2], [3]). Combinatorics and enumeration of semantics by inducing different connections on the multimedia objects perfectly explains the possibility of ambihuity and also the capability of humans to disambihuate such objects given a particular context in which input stimuli occur. Thus in order to come up with a unique interpretation of scenes obviously user context information is indispensable. Finally, these (fuzzy) inference structures give rise to multiple dynamic multimedia object interpretations. The output of these inference structures in terms of gauge invariant and robust multimedia meta-indices, including issues like spatio-temporal and dynamic inclusion and ordering relations for the self-similar multimedia objects, can then also be used to come up with multiple summarisation, synthesis and association schemata.

As in section III we address the problem of web-enabled multiple summarisations of multimedia presentations based on different query criteria we briefly dive into a-symmetric clustering of self-similar multimedia objects as proposed by Iacob [4]. A dynamic multimedia object summary in this clustering paradigm comes about by a recursive grouping and possibly hierarchical reordering of perceptually self-similar and predominant multimedia objects. Now we may base our similarity measure on histograms of gauge invariant and robust multimedia meta-indices such as proposed by Salden and Aldershoff [2], [3]. The hierarchical grouping of the most representative and predominant key multimedia objects using a-symmetric similarity subsequently comes about after the construction of a directed weighted graph . This graph consists of a set of vertices that is in a one-one relation with the original set of multimedia objects and a set of edges. Each edge is then also weighted by a similarity measure for a key-object relative to another. An optimal collection of two-level trees is obtained from the directed weighted graph of self-similar multimedia objects by searching for the minimal number of two-level trees with a maximal number of edges having the highest weights.

III. WEB-ENABLED INTEGRATED SYSTEM

We present a web-enabled advanced multimedia system in Fig.4 that combines our theory of advanced multimedia systems with that underlying the Acoi system [1]. The Acoi system is

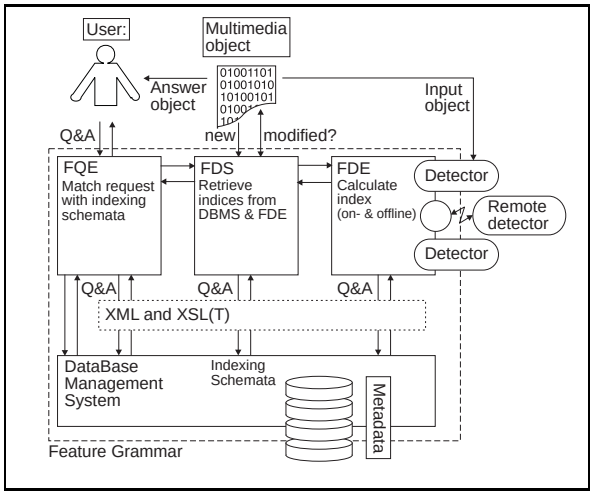

Fig. 4. Web-enabled advanced multimedia system: Acoi system using advanced multimedia categorization schemata.

based on a formal and grammatical framework for multimedia indexing schemata. The grammars describe the relationships between meta-data and detectors in terms of grammar rules. They are context-free grammars extended by feature detectors that can be executed on a multimedia object. This means that

a feature grammar represents an indexing schema and that it can consequently be used to find indices, as well as the original multimedia objects. In the Acoi system a feature grammar and the feature detectors are combined in a Feature Detector Engine, FDE (a parser). The FDE stores the meta-data, produced by the feature detectors, in a parse tree that in turn is put in a meta-index. In case of changes in the original multimedia or the feature detectors the Feature Detector Scheduler, FDS, uses knowledge of dependencies between feature detectors to update the meta-index. Finally, the meta-index database can be searched by means of the Feature Query Engine, FQE. Because the Acoi system does not specify the needed context for feature grammars and detectors, our physical, mathematical and logical framework of the previous section is required to derive the necessary multimedia semantics. Our framework provides hooks and contexts for logically, mathematically and physically sensible actions with respect to multimedia objects performed by the FDE, FDS and FQE of the Acoi system.

Our advanced multimedia system tailors to users, content providers and content producers. As depicted it has various operational modes, among which multimedia distributed storage, indexing, querying and retrieval. The indexing mode starts after a multimedia object is added to the database. The retrieval phase starts after the user has initiated the query mode on the system. Our advanced multimedia system governs categorization schemata underlying all these operational modes; our system provides therewith suitable feature grammars and related detectors that the Acoi system in turn uses.

As a demonstration of the added-values of our advanced multimedia system we apply web-enabled summarisation schemata to a multimedia presentation (see Fig. 5). Using the query instantiation that allows for textual formulation of the desired presentation searched for and for indicating the user's degree of occupancy, we derive two multimedia presentation summarisaties. In the summaries key multimedia objects preserving predominant audio-video and slide information are presented according to user availability. The multimedia summary consists of key video frames, grey audio intervals and slides. Key slide and audio transcript information steer the key video frame selection. We used categorical statistical relevance weighting and reinforcement to select the key multimedia objects (see section II). Such types of summaries appear in particular to be very useful in educational settings in which learners have limited time [10], [11].

IV. Conclusion

We presented a web-enabled adavanced multimedia system that enables a gauge consistent and robust categorization of multimedia objects. Sustainable retrieval, querying and indexing schemata have become web-enabled. Ambient aware multimedia web services for example in complex E-learning settings have come into sight.

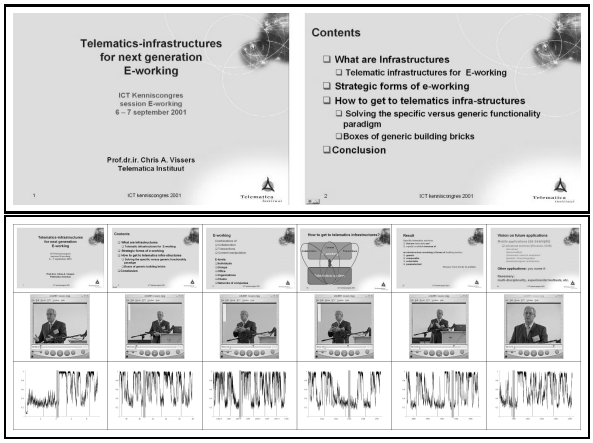

Fig. 5. Web-enabled summarisations of a multimedia presentation: a short slide summary (top; stressed user) versus a more fine-grained multimedia summary (bottom; relaxed user).

References

[1] M. Windhouwer, A. Schmidt and M. Kersten, "Acoi: A system for indexing multimedia objects," In *Proceedings of the first international workshop on information integration and web-based applications and services*, Yogyakarta, Indonesia, November 1999.

[2] A.H. Salden, "Multimedia system analysis and processing," In *Proceedings of 2001 IEEE International Conference on Multimedia and Expo*, ICME2001, August 22-25, 2001, Waseda University, Tokyo, Japan.

[3] F. Aldershoff and A. H. Salden,"Multiscale audio-video analysis and processing: segmentations and arrangements," In Proc. SPIE , Internet Multimedia Management Systems II, Vol. 4519, pp. 20-31, 2001.

[4] S. M. Iacob, R. L. Lagendijk, M. E. Iacob, "Video abstraction based on asymmetric similarity values," In *Proceedings SPIE Multimedia Storage and Archiving Systems IV*, Vol. 3846, pp. 181-191, 1999.

[5] C. Faloutsos, M. Flickner, W. Niblack, D. Petkovic, W. Wquitz, R. Barber, "Efficient and Effective Querying by Image Content," Research Report RJ 9203 (81511), IBM Almaden Research Center, San Jose, Aug. 1993.

[6] J. R. Smith and S. F. Chang, "VisualSEEK: A fully automated content-based image query system," In *ACM Multimedia*, 1996.

[7] A. Hamrapur, A. Gupta, B. Horowitz, C. F. Shu, C. Fuller, J. Bach, M. Gorkani and R. Jain, "VIRAGE Video Engine," In *SPIE Proceedings on Storage and Retrieval for Image and Video Databases V*, pp. 188-197, 1997.

[8] S. F. Chang, W. Chen, J. Meng, H. Sundaram and D. Zhong, "VideoQ: An automated content based video search system using visual cues," In *ACM Multimedia 1997*, pp. 313-324, 1997.

[9] Roman E. Maeder, *The Mathematica Programmer II*, Academic Press, 1996. http://www.wolfram.com/

[10] R. Brussee, A. Salden, H. van Vliet and P. Boekhoudt, "A web based architecture for e-learning," In *SSGRR 2001: International Conference on Advances in Infrastructure for Electronic Business, Science, and Education on the Internet*, L'Aquila, Italy, 2001.

[11] M. Grootveld and H. van Vliet (Eds.), "Engineering Educational Content," GigaCE Deliverable 6.1, TI/RS/2001/071, Enschede, The Netherlands, 2001.