CONTENT 2010

The Second International Conference on Creative Content Technologies

November 21-26, 2010 - Lisbon, Portugal

ComputationWorld 2010 Editors

Ali Beklen, IBM Turkey, Turkey
Jorge Ejarque, Barcelona Supercomputing Center, Spain
Wolfgang Gentzsch, EU Project DEISA, Board of Directors of OGF, Germany
Teemu Kanstren, VTT, Finland
Arne Koschel, Fachhochschule Hannover, Germany
Yong Woo Lee, University of Seoul, Korea
Li Li, Avaya Labs Research - Basking Ridge, USA
Michal Zemlicka, Charles University - Prague, Czech Republic

CONTENT 2010

Foreword

The Second International Conference on Creative Content Technologies [CONTENT 2010], held between November 21 and 26 in Lisbon, Portugal, targeted advanced concepts, solutions and applications in producing, transmitting and managing various forms of content and their combination. Multi-cast and uni-cast content distribution, content localization, on-demand or following customer profiles are common challenges for content producers and distributors. Special processing challenges occur when dealing with social content, graphic content, animation, speech, voice, image, audio, data, or image contents. Advanced producing and managing mechanisms and methodologies are now embedded in current and soon-to-be solutions.

We take here the opportunity to warmly thank all the members of the CONTENT 2010 Technical Program Committee, as well as the numerous reviewers. The creation of such a broad and high quality conference program would not have been possible without their involvement. We also kindly thank all the authors who dedicated much of their time and efforts to contribute to CONTENT 2010. We truly believe that, thanks to all these efforts, the final conference program consisted of top quality contributions.

Also, this event could not have been a reality without the support of many individuals, organizations, and sponsors. We are grateful to the members of the CONTENT 2010 organizing committee for their help in handling the logistics and for their work to make this professional meeting a success.

We hope that CONTENT 2010 was a successful international forum for the exchange of ideas and results between academia and industry and for the promotion of progress in the area of creative content technologies.

We are convinced that the participants found the event useful and communications very open. We also hope the attendees enjoyed the beautiful surroundings of Lisbon, Portugal.

CONTENT 2010 Chairs:

Ajith Abraham, Machine Intelligence Research Labs (MIR Labs), USA
Jalel Ben-Othman, Université de Versailles, France
Petre Dini, Concordia University, Canada/ IARIA, USA
Klaus Schmid, University of Hildesheim, Germany
Hans-Werner Sehring, T-Systems Multimedia Solutions GmbH, Germany

CONTENT 2010

Committee

CONTENT Advisory Chairs

Academia
Petre Dini, Concordia University, Canada/ IARIA, USA
Jalel Ben-Othman, Université de Versailles, France
Klaus Schmid, University of Hildesheim, Germany

Industry
Ajith Abraham, Machine Intelligence Research Labs (MIR Labs), USA
Hans-Werner Sehring, T-Systems Multimedia Solutions GmbH, Germany

CONTENT 2010 Technical Program Committee

Ajith Abraham, Machine Intelligence Research Labs (MIR Labs), USA
J. Enrique Agudo, University of Extremadura, Spain
Dana Al Kukhun, IRIT - University of Toulouse III, France
Marios C. Angelides, Brunel University - Uxbridge, UK
José Enrique Armendáriz-Iñigo, Universidad Pública de Navarra - Pamplona, Spain
Kambiz Badie, Iran Telecom Research Center, Iran
Christine Bauer, Vienna University of Economics and Business, Austria
Jalel Ben-Othman, Université de Versailles, France
Pradipta Biswas, University of Cambridge, UK
Letizia Bollini, Università degli Studi di Milano-Bicocca, Italy
Kadi Bouatouch, IRISA / University of Rennes 1, France
Laszlo Böszörmenyi, University Klagenfurt, Austria
Yiwei Cao, RWTH Aachen University, Germany
Eduard Céspedes Borràs, Universitat Autonoma de Barcelona (UAB), Spain
Eduardo Cerqueira, Federal University of Para, Brazil
Rafael del Vado Vírseda, Universidad Complutense de Madrid, Spain
Antonio Javier García Sánchez, Technical University of Cartagena, Spain
Patrick Gros, INRIA - Renes, France
Chih-Cheng Hung, Southern Polytechnic State University, USA
Harald Kosch, University of Passau, Germany
Narayanan Kulathuramaiyer, Universiti Malaysia Sarawak, Malaysia
Eugenijus Kurilovas, Institute of Mathematics and Informatics of Vilnius University / Vilnius Gediminas Technical University, Lithuania
Aggelos Lazaris, University of California - Riverside, USA
Maryam Tayefeh Mahmoudi, Iran Telecom Research Center, Iran
Vittorio Manetti, CRIAI Consortium / University of Napoli "Federico II", Italy
Michael Adeyeye Oluwasegun, University of Cape Town, South-Africa
Francisco Oteiza Lacalle, Telefonica R&D, Spain

Chris Poppe, Ghent University - IBBT - Ledeberg-Ghent, Belgium
Yonglin Ren, University of Ottawa, Canada
Joel Rodrigues, Instituto de Telecomunicações, University of Beira Interior, Portugal
Aitor Rodriguez Alsina, Universitat Autonoma de Barcelona, Spain
Héctor Sánchez, University of Extremadura, Spain
Klaus Schmid, University of Hildesheim, Germany
Hans-Werner Sehring, T-Systems Multimedia Solutions GmbH, Germany
Timothy K. Shih, Asia University - Wufeng, Taiwan
Emad Shihab, Queen's University - Kingston, Canada
Sean W. M. Sigueira, Federal University of the State of Rio de Janeiro (UNIRIO), Brazil
Robin JS Sloan, University of Abertay, UK
Bouchra Soukkarieh, IRIT - University of Toulouse III, France
Vladimir Stantchev, Berlin Institute of Technology, Germany
Toma Stefan-Adrian, The Military Technical Academy, Bucharest, Romania
Zhou Su, Waseda University, Japan
Daniel Thalmann, EPFL Vrlab - Lausanne, Switzerland
Božo Tomas, University of Mostar, Bosnia and Herzegovina
Tuan Tran, Oregon State University, USA
Honggang Wang, University of Massachusetts - Dartmouth, USA
Higang Yue, University of Lincoln, UK

Copyright Information

For your reference, this is the text governing the copyright release for material published by IARIA.

The copyright release is a transfer of publication rights, which allows IARIA and its partners to drive the dissemination of the published material. This allows IARIA to give articles increased visibility via distribution, inclusion in libraries, and arrangements for submission to indexes.

I, the undersigned, declare that the article is original, and that I represent the authors of this article in the copyright release matters. If this work has been done as work-for-hire, I have obtained all necessary clearances to execute a copyright release. I hereby irrevocably transfer exclusive copyright for this material to IARIA. I give IARIA permission or reproduce the work in any media format such as, but not limited to, print, digital, or electronic. I give IARIA permission to distribute the materials without restriction to any institutions or individuals. I give IARIA permission to submit the work for inclusion in article repositories as IARIA sees fit.

I, the undersigned, declare that to the best of my knowledge, the article is does not contain libelous or otherwise unlawful contents or invading the right of privacy or infringing on a proprietary right.

Following the copyright release, any circulated version of the article must bear the copyright notice and any header and footer information that IARIA applies to the published article.

IARIA grants royalty-free permission to the authors to disseminate the work, under the above provisions, for any academic, commercial, or industrial use. IARIA grants royalty-free permission to any individuals or institutions to make the article available electronically, online, or in print.

IARIA acknowledges that rights to any algorithm, process, procedure, apparatus, or articles of manufacture remain with the authors and their employers.

I, the undersigned, understand that IARIA will not be liable, in contract, tort (including, without limitation, negligence), pre-contract or other representations (other than fraudulent misrepresentations) or otherwise in connection with the publication of my work.

Exception to the above is made for work-for-hire performed while employed by the government. In that case, copyright to the material remains with the said government. The rightful owners (authors and government entity) grant unlimited and unrestricted permission to IARIA, IARIA's contractors, and IARIA's partners to further distribute the work.

Table of Contents

Classifying Content Mode of Organizational Texts Using Simple Neural and Neuro-Fuzzy Approaches

Maryam Tayefeh Mahmoudi[1,2], Babak N. Araabi[2], Kambiz Badie[1], Nafiseh Forouzideh[3]
1: Knowledge Engineering & Intelligent Systems Group
IT Research Faculty, Iran Telecom Research Center
Tehran, Iran
Emails: {Mahmodi, k_badie}@itrc.ac.ir
2: Control and Intelligent Processing Centre of Excellence,
School of Electrical and Computer Eng., University of Tehran,
Tehran, Iran
Emails: {tayefeh, araabi }@ut.ac.ir
3: Kish Intl. Campus, University of Tehran,
Kish, Iran
Email: n.forouzideh@gmail.com

***Abstract*—In this paper, we present simple neural and neuro-fuzzy approaches to classify the mode of a text's content which is organized for helping users with their organizational tasks. In this regard, 7 major features were chosen as inputs for our suggested approaches. 3 nominal values L, M, and H were used as the possible values for each feature. Results of experimentation on a dataset including 540 data show the fact that the Takagi-Sugeno as a neuro-fuzzy approach using lolimot learning algorithm, performs better compared to multi-layer perceptron and radial basis function as simple neural approaches. Due to the high performance of this approach, it is expected to be successfully applicable to a wide range of content mode classification issues in decision support environment.**

Keywords- text classification; neural network; neuro-fuzzy approach; organizational task; content mode.

I. INTRODUCTION

In recent years, text mining has been widely used to extract the significant information from a text, among which extracting facts or regularities as well as the focal points are mentionable [1, 2, 3, 4]. An important point in this concern is the type(s) or class (es) to which a text or parts of a text may belong to. This has made classification one of prime issues in text mining [5, 6]. One major aspect in text classification is to identify the type of a text's content, e.g., its mode/style, its peculiarities/ characteristics, the category it belongs to, as well as the peculiarities of the environment within which it has been prepared. Pattern recognition techniques have a wide range of applications in this issue. Due to the distributed characteristics of a text, i.e., the fact that its mode/style may exhibit itself in an aggregation of a variety of considerations in its different parts/ components, non-symbolic classification methods equipped with logic of uncertainty handling like probabilistic and fuzzy logic are expected to be particularly workable in this regard.

Based on the above point, in this paper, we present an approach for classifying the mode of a text's content using neuro-fuzzy techniques [7, 8]. Due to the significance of comprehensive contents in making efficient decisions in organizations, the content mode considered in our approach is the type of an organizational task with regard to which texts have been organized.

The structure of the paper is as follows. Section II represents the existing approaches to text classification systems, while the emphasis of Section III is on the proposed approach including "the architecture of the proposed approach", "feature selection", and "experimental results" as well. Concluding remarks is also presented in Section IV.

II. EXISTING APPROACHES TO TEXT CLASSIFICATION SYSTEMS

Text classification can be defined as assigning texts to a predefined set of categories, which is used in situations where classes of texts/contents are labeled and include specific features. In this regard, from the viewpoints of similarity and regularity in features, the input content/text is supposed to be finally classifiable in terms of some pre-defined classes that can be significant in some sense. Within this context, classification can be performed based on the type of content, subject/issue, qualification level and style of content/text and even its authors' specifications. Text classification can also ease the organization of increasing textual information, in particular Web pages and other electronic form of documents [9]. It usually consists of two parts: feature selector and text classifier.

 ISBN: 978-1-61208-110-6

Feature selector selects the features which are essential to classifying the text's content; in terms of a feature vector. The classifier then assigns the feature vector to the appropriate class (es). Researches indicate that many techniques can be used in feature selection to improve accuracy as well as to reduce the dimensions of the feature vector and thus reduce the time for computation. Feature selection mostly adopts various assessment functions such as document frequency, information gain, mutual information, and statistics (CHI) to calculate the weights [10]. Many classifiers have been applied to classify texts, including Naïve Bayes [11], decision trees [12], k-nearest neighborhood [13], linear discriminate analysis (LDA) [14], logistic regression [15], neural networks [6], support vector machines [16], rule learning algorithms [17], relevance feedbacks [18], etc. Several kinds of competitive networks are used in text classification, including learning vector quantization (LVQ) and self-organizing maps (SOM) network. These two are both variants of the basic unsupervised competitive network. Besides, back propagation (BP) and radial basis function (RBF) networks are two successful examples for classification. There also exist some other statistical approaches for modeling a document for text classification like LSA, pLSA, and LDA [19].

Some special classification methods are also available for specific purposes, like Rocchio, which is for text classification in information retrieval [5] and independent component analysis (ICA), which was developed for the blind signal decomposition and recently used for selecting the mutually independent features of a document [20]. Text representation may also have a significant role in classifying texts with several features [21]. A series of experiments on text classification using multi-word features have also been done [22]. Meanwhile, web text classification has also been introduced as one of the major activities in this field [23].

III. THE PROPOSED APPROACH

Due to the distributed characteristics of a text and the fact that its mode/style may exhibit itself in an aggregation of several considerations in its different parts/components, non-symbolic classification methods equipped with logic of uncertainty handling are expected to function more efficiently.

Based on the above point, in this paper, we present an approach for classifying the mode of a text's content using neuro-fuzzy techniques. The mode of text's content considered in our approach is the type of an organizational task with regard to which the text has been organized using the dataset that has been prepared on the basis of the existing technical reports at a research institute. It is interesting to see that these tasks are equally being used by a wide range of knowledge workers (researchers, innovators, developers, planners, analyzers, etc.) in an organization to disseminate results of their works in terms of appropriate contents. Some of the major tasks important for an organization are: Planning/Scheduling, Research, Innovation, Development/ optimization/ Improvement, Education/ Promotion, Analysis/ Assessment/ Assurance, and Guidance, Justification.

In this paper, six of these tasks Research, Development/Planning, General Learning, Justification, Innovation and Analysis/ Assessment, etc are considered as the output classes.

A. The Architecture of the Proposed Approach

In this paper, the focus is on classification of a text's content using neuro-fuzzy approach. Neuro–fuzzy approaches in general and neuro-fuzzy networks in particular are fuzzy models that are not solely designed by expert knowledge but are at least partly learned from experiential data. If no a-priori knowledge is available, the application of a fuzzy model does not make any sense from the model accuracy point of view. However, if accuracy is not the only ultimate goal and instead an understanding of the functioning of the process is desired, then fuzzy models are the best choice [7]. In this respect, features of each functionality of a text's content can be identified and valued. These functionalities are considered to be the 6 classes of organizational tasks discussed above. In this regard, 27 features defined, out of which 7 major features have been chosen as inputs for our neural and neuro- fuzzy approaches. The important features are: "General Background", "Existing viewpoints", "Key issue", "Proposed approach realization/ implementation", "Validation/Verification", "Comparative analysis & capability interpretation", "Conclusion & prospect anticipation". The values of each of the features have been determined by experts. For instance, in a general learning content, for each feature of "General Background", "Existing viewpoints", "Key issue", the nominal values of "L" (Low), "H" (High), and "M" (Medium) have been determined. Detailed information about the features, their values and output classes are represented comprehensively in the next section. Figure 1 illustrates the overview of our proposed classification system.

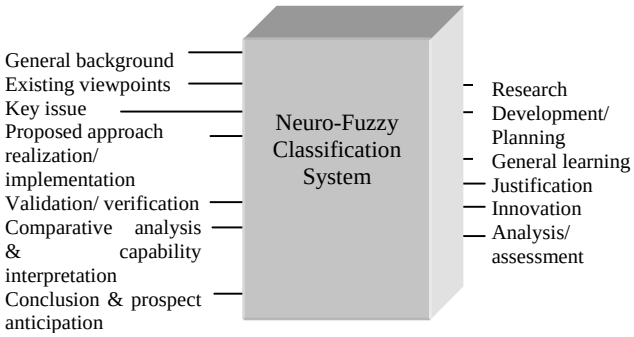

Figure 1. General view on the proposed classification system

In this paper, for classification purpose, both simple neural and neuro-fuzzy techniques have been considered. In this respect, multi-layer perceptron (MLP) and radial basis function (RBF) are implemented for simple neural and Takagi-Sugeno with Lolimot learning algorithm is implemented for neuro-fuzzy classification.

 ISBN: 978-1-61208-110-6

The experiments have been done on a dataset with 540 data, which have been prepared on the basis of the existing technical reports.

B. Feature Selection

With respect to mining issues, classifying the patterns existing in a database is of great importance, and due to this, selecting appropriate features for classification would also be significant. The high number of features in feature vector, makes in practice some difficulties when neural net is used as the classifier. In this respect, the major informative and uncorrelated features should be selected for classification [24].

In our approach, the appropriate features for classifying text's content are identified on the basis of the expert's idea and the existing approaches [25] as well. Within 27 previously identified features [25], 7 major features have been considered in this paper. Table 1 illustrates these features with their prospected values. It is obvious that, these features have been realized to be consistent for a wide range of contents which are to be created for helping users with their tasks in organizations, as discussed in the beginning of the section.

Obviously, based on the type of a task, a limited number of the labels and the corresponding sub-labels may be activated. Nominal values "L" (standing for Low), "M" (standing for Medium), and "H" (standing for High) associated with the labels of key segments indicate the extent according to which linguistically significant notions such as "What", "Who", "Whom", "Where", "Which", "When", "How", and "Why", can be addressed to create a petit content for each key segment in the content. This is done by the nominal values pre-agreed for each task, to show to what extent linguistically significant notions like "What", "Which", "Where", "When", "Whom", "Who", "Why", and "How" should be addressed [25].

Taking this point into account, the feature vector of input content is structured based on the afore-mentioned features and the nominal values (Table 1).

TABLE I. INPUT FEATURES AND OUTPUT CLASSES OF PROPOSED SYSTEM

Output Classes / Input Content Features	Research	Development / Planning	General Learning	Justification	Innovation	Analysis/ Assessment
General Background	H	M	L	L	M	L
Existing viewpoints	H	M	H	L	M	L
Key issue	H	M	M	M	M	M
Proposed approach realization/ implementation	H	M	L	M	L	M
Validation/ Verification	H	M	L	M	L	H
Comparative analysis & capability interpretation	H	M	L	L	L	L
Conclusion & prospect anticipation	H	H	L	L	L	L

Taking this point into account, the dataset used in this research would include the data from text/ content's labels that belong to 6 classes. It contains 540 samples with 7 attributes. After normalizing the input data and making the test and train data, classification would start.

C. Experimental Results

Simple neural approaches are used when no particular emphasis is made on the status of uncertainty in the related data, while neuro-fuzzy approaches are used to consider such a status of uncertainty. In the paper, we consider both of the approaches to show that uncertainty of the information in content is a matter which can not be disregarded [7, 18, 26].

1) Classification using MLP

In this respect, a feed forward MLP has been used with 1 hidden layer and a variation of hidden neurons. The optimal number of neurons in this respect was found to be 20. As we have 6 output classes, the binary forms of these classes would be as follows:

Output1/Class1 -> [0 0 1]
Output2/Class2 -> [0 1 0]
Output3/Class3 -> [0 1 1]
Output4/Class4 -> [1 0 0]
Output5/Class5 -> [1 0 1]
Output6/Class6 -> [1 1 0]

It is to be noted that if we divide the network into sub networks, the learning rate increases. In this regard, three networks of binary form of output classes are trained with normalized input data. The specifications of these three binary networks are as follows:

Number of neurons: 20; Train parameter epochs: 100; DivideParam.trainRatio = 0.7; DivideParam.testRatio = 0.15; DivideParam.valRatio = 0.15; Train Param. max_ fail = 30;

After training each network separately, the total network output is computed and is transformed from binary into decimal to have checked its status of belonging to the existing classes. Reconstructing test data for outputs and comparing the real classes with the network outputs yields realization of the whole classification process. The status of networks outputs are as follows:

1) First network: the best performance of validation, with least MSE is 0.029983 at epoch 6. The regression status shows 0.99, 0.94 and 0.89 learning respectively for the training, the test, and the validation data. Taking this point into account, 0.97 learning will be the result of the first experimentation.

2) Second network: the best performance of validation, with least MSE is 0.074266 at epoch 9. The regression status shows 0.94, 0.88 and 0.74 learning respectively for the training, the test, and the validation data. Taking this point into account, 0.90 learning will be the result of the second experimentation.

3) Third network: the best performance of validation, with least MSE is 0.001768 at epoch 99. The regression

 ISBN: 978-1-61208-110-6

status shows 0.999, 0.943 and 0.996 learning respectively for the training, the test, and the validation data. Taking this point into account, 0.9907 learning will be the result of the first experimentation.

As a result, it can be mentioned that the 3rd network learns totally better than the others networks with the rate of 99%, although it needs more epoch to reach the least MSE.

Results of the classification experiments on all the 540 data of the input dataset is illustrated in Figure 2.

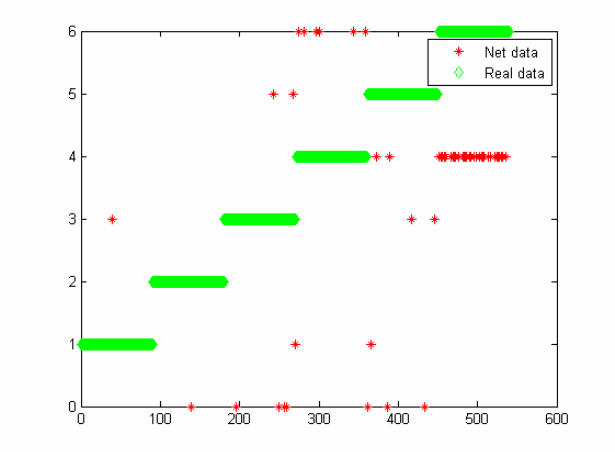

Figure 2. The classification result on all the input data

As it is seen, 57 inputs among 540 were false classified. The resultant total classification error is 10.55%, while the classification accuracy is 89.44%.

The experiments on the test data, reveals that 10 data were classified falsely. The classification error on the test data is 12.34% while the accuracy of the correct classification on test data is: 87.65%.

2) Classification using RBF

Another neural method for identification which has been used in this experiment is RBF whose basis function is Gaussian.

After normalizing all the inputs, the training and the test data are structured and all the three outputs are then computed. Based on the following conditions, the RBF is trained and tested on all data for the three networks:
goal = 0; spread = 1; MaxNeurons = 30; displayInterval = 2.
The status of outputs of the networks is as follows:

- First network: The best performance of NRBF is 0.0207203, considering Goal=0.
- Second network: The best performance of NRBF is 0.0585283 considering the Goal=0.
- Third network: The best performance of NRBF is 0.0136336, considering Goal=0.

As a result, it can be mentioned that the 3rd network totally learns better than the other networks with the rate of 85%.

Results of experiments on all the 540 data of input dataset are illustrated in figure 3.

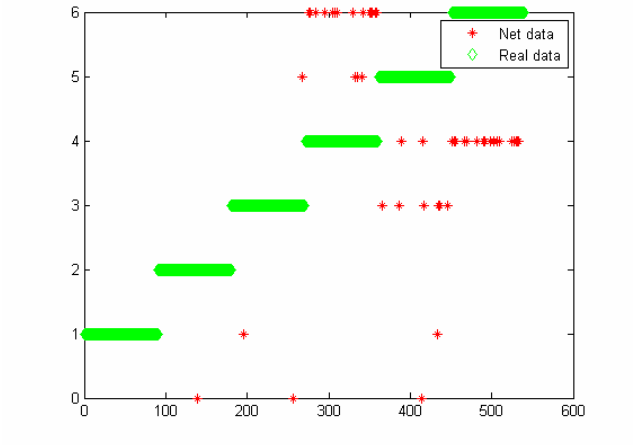

Figure 3. Classification result for all the data using NRBF

As it is seen, 48 inputs among 540 were false classified. The resultant total classification error is 8.88%, while the classification accuracy is 91.11%.

The experiments on the test data, reveals that 10 data were classified falsely. The classification error on the test data is 12.5% while the accuracy of the correct classification on test data is: 87.5%.

3) Classification Using Takagi-Sugeno with Lolimot

The neuro-fuzzy method applied for classification is Takagi-Sugeno with Lolimot learning algorithm [27]. As it is known, this method starts with an initial model, finds worst linear language model (LLM), checks all the divisions, finds best division and finally tests for convergence [7, 28].
Considerations for this experiment are as follows: smoothing factor (alpha) =1/3, mse_goal=1e-4 and reg_coef=0. Training the same three binary networks as previous parts, with max 30 neurons reveals that, the appropriate numbers of neurons for them respectively are: 6, 29, and 12.

As a classification result, it is to be noted that among 81 test data as input, 7 were classified falsely. Taking this point into account, the classification error was realized to be 8.64% and Lolimot was therefore able to distinguish 91.36% correct classes.

Figure 4 illustrates the classification status for both training and test data (540 input data). As it is seen, 23 data were classified falsely. In this regard, the total error of the network was realized to be 4, 26%.

 ISBN: 978-1-61208-110-6

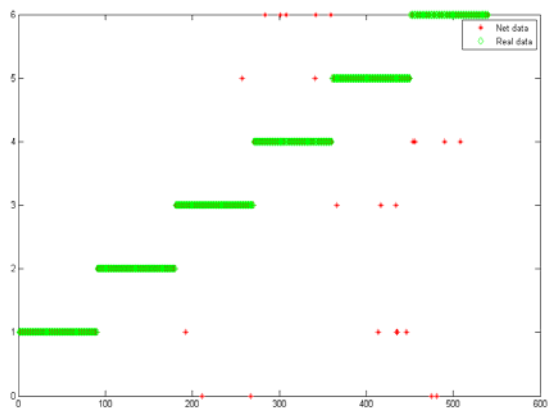

Figure 4. The classification of all the input data using TSK using lolimot

TABLE II. THE CLASSIFICATION RESULTS ON THE GIVEN DATASET FOR MLP, RBF AND TSK USING LOLIMOT

	MLP	RBF	Takagi-sugeno using lolimot
Appropriate No. of neurons for output1	20	30	6
Appropriate No. of neurons for output2	20	30	29
Appropriate No. of neurons for output3	20	30	12
Correct classification rate for Test data	% 87.65	% 87.5	% 91.36
False classification rate for Test data	% 12.34	% 12.5	%8.64
Correct classification rate for Whole data	% 89.44	% 91.11	%95.74
False classification rate for Whole data	% 10.55	% 8.88	%4.26
No. of corrected classified on Test data	71/81	71/81	74/81
No. of corrected classified on whole data	483/540	492/540	517/540

Figure 5 shows the comparison between the classification rates respectively belonging to MLP, RBF and Takagi-Sugeno using Lolimot. As it is seen from the experimental results, Takagi-Sugeno using Lolimot has classified better on test data compared to MLP and RBF.

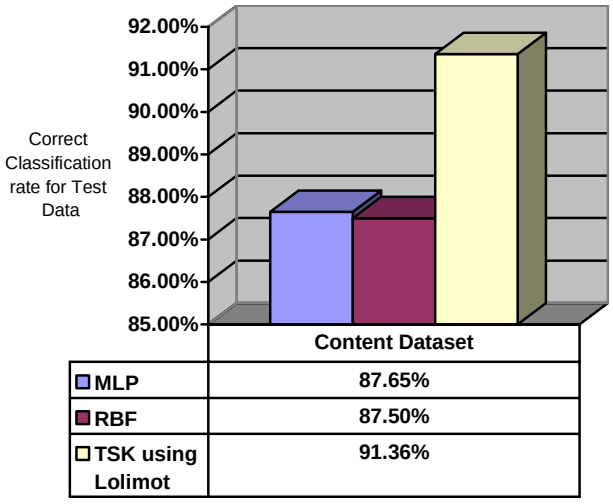

Figure 5. The comparison between Percent of corrected classification on Test data by MLP, RBF and TSK using Lolimot

The classification results on the whole data are illustrated in Figure 6. As it is seen, again TSK using Lolimot performs better than RBF, and RBF better than MLP.

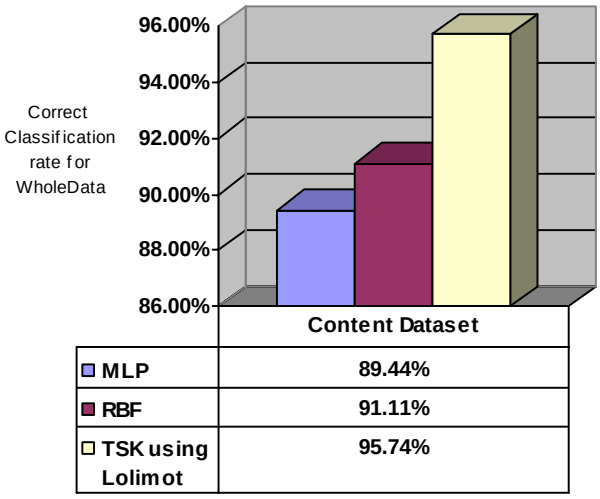

Figure 6. The comparison between Percent of corrected classification on whole data by MLP, RBF and TSK using Lolimot

As a conclusion, Takagi-Sugeno using lolimot learning algorithm, reveals better performance on the given dataset compared to the other mentioned algorithms. This is at first glance because a text's content has generally a multi-class or multi-modal nature, and thus due to its simultaneous affiliation to different classes (modes), classification approaches based on a sort of uncertainty handling logic can perform far better compared with those without such a basis. Moreover, the very peculiar ability of lolimot as a learning algorithm in speeding up the training procedure as well as incorporating with many kinds of prior knowledge (nominal

 ISBN: 978-1-61208-110-6

values for the input features in our case), and also its insensitivity toward curse of high dimensionality makes utilization of neuro-fuzzy approach more successful.

IV. CONCLUDING REMARKS

In this paper, the performance of multi-layer perceptron and radial basis function as simple neural approach, and Takagi-Sugeno with lolimot learning algorithm as neuro-fuzzy approach was evaluated for classifying the mode of a text's content, which is basically designed for helping users with their organizational tasks.

Experimental results on an initial dataset including the data belonging to 540 texts, demonstrate the fact that the Takagi-Sugeno with lolimot learning algorithm performs far better compared to simple neural approaches. This, as was discussed, is mainly due to ability of this approach in classifying the patterns of texts, which are somewhat multi-class or multi-modal in nature.

The approach presented in this paper can be particularly useful for organizing texts in decision support environments, where enriching the existing texts for supporting the human elements with their decisions (as the possible labels for content mode) is of particular significance.

REFERENCES

[1] D. S´anchez, M. J. Mart´ın-Bautista, I. Blanco, and C. Justicia de la Torre, "Text Knowledge Mining: An Alternative to Text Data Mining", 2008 IEEE Intl. Conf. on Data Mining , pp. 664-672.

[2] W. Wang, C. Wang, X. Cui, and A. Wang, "Fuzzy C-Means Text Clustering with Supervised Feature Selection," Fifth Intl. Conf. on Fuzzy Systems and Knowledge Discovery, 2008, Vol. 1, pp. 57-61.

[3] Y. Lu, S. Wang, S. Li, and C. Zhou, "Text Clustering via Particle Swarm Optimization", IEEE Conf. on Swarm Intelligence Symposium, (SIS '09), 2009, pp. 45-51.

[4] R. Li, J. Zheng and C. Pei, "Text Information Extraction Based on Genetic Algorithm and Hidden Markov Model" , First Intl. Workshop on Education Technology and Computer Science, Vol.1, 2009, pp.334-338.

[5] A. Danesh, B. Moshiri, and O. Fatemi, "Improve Text Classification Accuracy based on Classifier Fusion Methods", 10th Intl. Conf. on Information Fusion, 2007, pp. 1-6.

[6] Z. Wang, Y. He, and M. Jiang, "A Comparison among Three Neural Networks for Text Classification", IEEE Intl. Conf. on Signal Processing, 2006, pp. 1883-1886.

[7] O. Neles, "Nonlinear System Identification", Springer Pub., 2001.

[8] D. Kukolj, and E. mil Levi, "Identification of Complex Systems Based on Neural and Takagi–Sugeno Fuzzy Model", IEEE Trans. on Systems, Man, and Cybernetics, Part B: Cybernetics, Vol. 34, No. 1, Feb. 2004.

[9] M. R. Islam and M. R. Islam, "An Effective Term Weighting Method Using Random Walk Model for Text Classification", 11th Intl. Conf. on Computer and Information Technology (ICCIT 2008), 2008, pp. 411 - 414.

[10] Z. T. Yu , L. Han, C.L. Mao , J. Y. Guo, X. Y. Meng , and Z. K. Zhang, "Study on the Construction of Domain Text Classification Model with the Help of Domain Knowledge", Seventh Intl. Conf. on Machine Learning and Cybernetics, 2008, pp. 2612 - 2617.

[11] P. Frasconi, G. Soda, and A. Vullo, " Text categorization for multi-page documents: a hybrid naive Bayes HMM approach". In Proceedings of the 1st ACM/IEEE-CS Joint Conference on Digital Libraries (Roanoke, Virginia, United States). JCDL '01. ACM, New York, NY, 2001, pp. 11-20.

[12] R. E. Schapire and Y. Singer, "BoosTexter: a boosting-based system for text categorization", Machine Learning Journal, Vol. 39, No. 2/3, 2000, pp. 135–168.

[13] G. Guo, H. Wang, D. Bell, Y. Bi, and K. Greer, "Using *k*NN model for automatic text categorization" , Journal of Soft Computing - A Fusion of Foundations, Methodologies and Applications Vol. 10, No. 5, 2006, pp. 423-430.

[14] W. Li, L. Sun, Y. Feng and D. Zhang, "Information Retrieval Technology", Smoothing LDA Model for Text Categorization, Lecture Notes in Computer Science, Vol. 993, 2008, pp. 83-94.

[15] A. Genkin, D. Lewis, and D. Madigan, " Large-scale Bayesian logistic regression for text categorization", Technometrics Journal, Vol. 49, No. 3, 2007, pp. 291–304.

[16] F. Sebastiani, "Machine Learning in Automated Text Categorization", ACM Computing Surveys, Vol. 34, No. 1, 2002, pp. 1–47.

[17] C. Apt´e, F.J. Damerau, and S.M. Weiss, „Automated learning of decision rules for text categorization", ACM Trans. on Inforation Systems, *Vol. 12*, No. 3, 1994, pp. 233–251.

[18] M. R. Azimi-Sadjadi, J. Salazar, S. Srinivasan, S. Sheedvash, "An Adaptable Connectionist Text Retrieval System With Relevance Feedback, IEEE Trans. on Neural Networks, Vol. 18, No. 6, Nov. 2007.

[19] J. Wang, X. Geng, K. Gao, and L. Li, "Study on Topic Evolution Based on Text Mining", Fifth Intl. Conf. on Fuzzy Systems and Knowledge Discovery ,Vol. 2, 2008, pp. 509-513.

[20] M. Hu, S. Wang, A. Wang, and L. Wang, "Feature Extraction Based on the Independent Component Analysis for Text Classification", Fifth Intl. Conf. on Fuzzy Systems and Knowledge Discovery, 2008, Vol. 2, pp. 296-300.

[21] J. Wang and Y. Zhou, "A Novel Text Representation Model for Text Classification", First Intl. Conf. on Intelligent Networks and Intelligent Systems, 2008, pp. 702-705.

[22] W. Zhang, T. Yoshida, and X. Tang, "Text classification using multi-word features", IEEE Intl.. Conf. on Systems, Man and Cybernetics, 2007, pp. 3519-3524.

[23] S. Yin, Y. Qiu, and J. Ge, "Research and Realization of Text Mining Algorithm on Web", Intl. Conf. on Computational Intelligence and Security Workshops, (CISW 2007), 2007, pp. 413-416.

[24] M. Rezaei, "Input Variable Selection in System Identification Application in Time Series Prediction", M.SC Thesis, University of Tehran, Feb 2008.

[25] K. Badie, M. Kharrat, M. T. Mahmoudi, M. S. Mirian, S. Babazadeh, and T. M. Ghazi, "Creating Contents based on Inter-play Between the Ontologies of Content's Key Segments and Problem Context", The First Intl. Conf. on Creative Content Technologies (CONTENT 2009), 2009, pp. 626-631.

[26] R. D. Goyal, "Knowledge Based Neural Network for Text Classification", 2007 IEEE Intl. Conf. on Granular Computing, 2007.

[27] L. Germán , O. Massa, L. Corbalán, L. Lanzarini, and A. De Giusti, "Evolving Fuzzy Systems A new strategy for rule semantics preservation", http://citeseerx.ist.psu.edu/viewdoc/ summary, doi=10.1.1.87.9948.

[28] J. Rezaie, B. Moshiri, and B. N. Araabi, "Distributed Estimation Fusion with Global Track Feedback Using a Modified LOLIMOT Algorithm", SICE Annual Conf. 2007, pp. 2966-2973.

 ISBN: 978-1-61208-110-6

Using Virtual Agents to Cue Observer Attention

Assessment of the impact of agent animation

Santiago Martinez, Robin J.S. Sloan, Andrea Szymkowiak and Ken Scott-Brown
White Space Research
University of Abertay Dundee
Dundee, DD1 1HG. UK
s.martinez@abertay.ac.uk, r.sloan@abertay.ac.uk, a.szymkowiak@abertay.ac.uk, k.scott-brown@abertay.ac.uk

***Abstract*— This paper describes an experiment developed to study the performance of virtual agent motion cues within digital interfaces. Increasingly, agents are used in virtual environments as part of the branding process and to guide user interaction. However, the level of agent detail required to establish and enhance efficient allocation of attention remains unclear. Although complex agent motion is now possible, it is costly to implement and so should only be routinely implemented if a clear benefit can be shown. Previous methods of assessing the effect of gaze-cueing as a solution to scene complexity have relied principally on manual responses. The current study used an eye-movement recorder to directly assess the immediate overt allocation of attention by capturing the participant's eye-fixations following presentation of a cueing stimulus. We found that fully animated agents speed up user interaction with the interface. When user attention was directed using a fully animated agent cue, users responded 35% faster when compared with stepped 2-image agent cues, and 42% faster when compared with a static 1-image cue. These results inform techniques aimed at engaging users' attention in complex scenes such as computer games or digital transactions in social contexts by demonstrating the benefits of gaze cueing directly on the users eye movements, not just their manual responses.**

Keywords: agents, interfaces, computer animation, reaction time, eyetracking

I. INTRODUCTION

The allocation of attention by a human observer is a critical yet ubiquitous aspect of human behaviour. For the designer of human-computer interfaces, the efficient allocation of operator attention is critical to the uptake and continued use of their interface designs. Historically, many human-computer interfaces (HCI) have relied on static textual or pictorial cues, or a very limited sequence of frames loosely interconnected over time (for example, on automated teller device menus, or on websites). More recently, the increased power of computer graphics at more cost effective prices has allowed for the introduction of high resolution motion graphics in human computer interfaces. Until now, psychological insights on attention and the associated cognitive processes have mirrored the HCI reliance on either static or stepped pictorial stimuli, where stepped pictorial stimuli consist of a few static frames displayed over time to imply basic motion. Again, this legacy can be attributed to limitations in affordable and deployable computer graphics.

The reported study is centered on the evaluation of fully animated (25 frames per second) virtual agents, where both the head and eye-movements of the agent are animated to allocate user attention. In contrast to most previous studies that have relied on manual responses to agent gaze, the current study uses the captured eye-gaze of the participant as a response mechanism, following on from the work of Ware and Mikaelian [15].

Where observers look in any given scene is determined primarily by where information critical to the observer's next action is likely to be found. The visual system can easily be directed to guide and inform the motor system during the execution of information searching. Consequently, a record of the path observer gaze takes during a task provides researchers with what amounts to a running commentary on the changing information requirements of the motor system as the task unfolds [4]. This is the underlying principle of the reported experiment, which is an expansion of the cognitive ethology concept expressed by Smilek et al. [3] to virtual agents. The experiment is based on the deictic gaze cue – the concept that the gaze of others acts like a signal that is subconsciously interpreted by an observer's brain, and that it can transmit "information on the world" [10]. The gaze of another human agent is inherently difficulty to avoid, and it can be used as a specific pointer to direct an observer's attention [8]. The incorporation of this concept can be easily implemented into an agent-based interface.

The efficiency of such an interface can be assessed based on the speed of observer response to cues. In the case of the current study, the cues are presented as fully animated (dynamic) agents, stepped agents (two images), or static agent images. Coupled with appropriate software, a virtual agent can anticipate user's goals, and point (using gaze) to the area where the next action has to be performed. An agent with animated gaze may therefore be useful to adopt in digital interfaces to guide user attention and potentially increase the speed of attention allocation, or where the work space of human physical action may have many possible choices and the possibility of not selecting the right one is high.

In the following sections we will explain in detail the application of the virtual agent to cue observer attention. In

 ISBN: 978-1-61208-110-6

Section 2 we will describe the existing literature reviews from two different research fields. In Section 3 we will explain the method used to develop the experiment. In Section 4 we will present the results of that experiment. Finally, in Section 5, we will discuss the overall results, the effects of 3D compared with 2D agents and the impact on user engagement and agent animation.

II. LITERATURE REVIEW

Previous studies belong to two different but related research fields: namely cognitive psychology and computer interface design. Psychological studies have reviewed attention and its relationship with the cues. Posner [11] describes the process of orienting attention. Relative to neutral cue trials, participants were faster and/or more accurate at detecting a target given a valid cue, and they were slower and/or less accurate given an invalid cue. Friesen and Kingstone [5] worked with faces and lines drawn following the gaze direction towards the target area. They found that subjects were faster to respond when gaze was directed towards the intended target. This effect was reliable for three different types of target response: detection, localization and identification. Langton and Bruce [3], and more recently Langton et al. [9], investigated the case of attention in natural scene viewing. They concluded that facial stimuli which indicate direction by virtue of their head and eye position produce a reflexive orienting response in the observer. Eastwood et al. [3] produced experimental findings which led to the conclusion that facial stimuli are perceived even when observers are unaware of the stimuli. In 2006, Smilek et al. [3] focused on isolating specific processes underlying everyday cognitive failures. They developed a measure for attention-related cognitive failures with some success, and introduced the term of cognitive ethology.

Studies in HCI and computing are focused on proving the validity of eye-gaze as an input channel for machine control. Ware and Mikaelian [15] used an eye-tracker to compare the efficacy of gaze an as an input channel with other more usual inputs, such as manual input using physical devices. They found that the gaze input was faster with a sufficient size of target. Sibert and Jacob [12] studied the effectiveness of eye gaze in object selection using their own algorithm and compared gaze selection with a traditional input – a hand operated mouse. They found that gaze selection was 60% faster than mouse selection. They concluded that the eye-gaze interaction is convenient in workspaces where the hands are busy and another input channel is required.

The above research shows how eye-gaze can be used to assess the response of a user when accurate tracking is possible. In addition, it has been demonstrated that the eye-gaze of an agent can effectively allocate attention. However, the interplay between pictorial cues to gaze allocated attention (and subsequent assessment of allocated attention) is still to be fully explored. In particular, for the reported experiment, two goals were set by the authors; to assess the extent to which the gaze of the observer can be used to record their selection of targets and response time to agent cues, and to determine whether fully animated agents would offer an advantage over standard static (1-image) or stepped

Figure 1: The appearance of the virtual agent, surrounded by eight target squares, arranged on both the cardinal and oblique axes.

(2-image basic motion) agents when directing attention using gaze. By focusing on gaze as a means of target selection, this removes as much motor response as possible from the observer. Manual responses inevitably introduce uncertainty in establishing the true response time since they are an indirect response to the gaze cue (requiring over allocation of attention and eye-gaze, followed by translation of the response signal to the sensory modality of touch). Therefore, when it comes to assessing the effectiveness of animated versus static and stepped agent cues, directly recording the eye-movements of observers and using this data to determine the speed of their response and their selection of objects offers a significant advantage.

III. METHOD

A. *Task description*

In this experiment, participants were asked to perform an object selection task (using their gaze alone) on a series of twenty-four different agent animations, presented on a monitor at a resolution of 1024 x 768. Each of the videos showed a virtual agent's head in the centre of the screen surrounded by eight different possible target areas (see Fig. 1). Each agent was displayed on screen for 3000 ms. Over the course of the video, the agent would orient its head and eyes aim at a particular target square. The point at which the agent oriented its head and gaze (and the nature of the agent's movement) was determined by the type of agent cue (see below). Of the eight target areas in each video, only one was the right choice in each trial – the one that was specifically indicated by the agent. If the participants selected that specific area with their eye-gaze, it was counted as a success. If the participant selected any of the other seven areas, it was counted as incorrect. Fixations to areas outside the 8 target areas were coded as no target selected. The target areas were red squares approximately 150 x 150 pixels in size, and were all equidistant from the center of the screen.

 ISBN: 978-1-61208-110-6

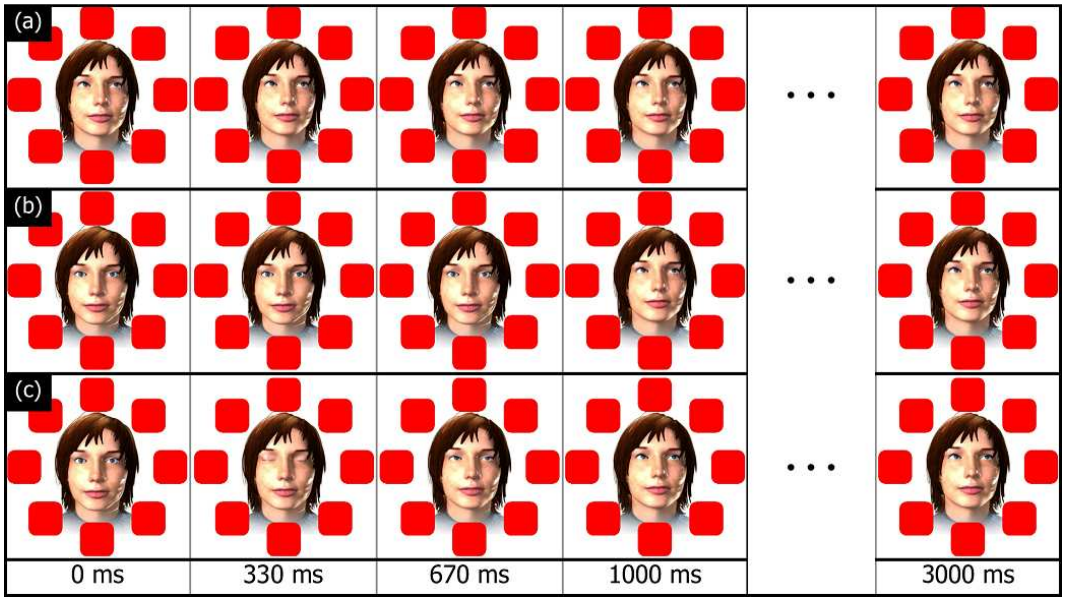

Figure 2. The appearance of the three types of helper agents over 1000 ms. Helper agents used head orientation and gaze to highlight one of eight targets. In the above example, three types of helper agent are shown highlighting the NE target. (a) shows a static (1-image) helper agent, which highlights the NE target from 0 ms onwards. (b) shows the stepped (2-image) helper agent, which looks towards the observer in frame 1 (from 0 ms) before changing to highlight the NE target in frame 2 (from 960 ms). (c) shows the dynamic (25-image, 25 fps) agent, which begins at 0 ms by looking at the observer, and is animated with natural movement so that the head and gaze shift towards the NE target at 960 ms. All helper agents are shown. to participants for a total of 3000 ms, so that the appearance of the agent at 1000 ms is held for two seconds.

B. *Agent Cues*

There were three different types of agent cues (see Fig. 2):

a) Static cue: A single image of an agent. The agent's head and eyes were aimed at the target area for the duration that the stimulus is displayed. The orientation cue was therefore presented from 0 ms till 3000 ms.

b) Stepped cue: Two images of an agent, sequenced to imply movement. The agent's head and eyes were looking straight forward from 0 ms, before the second image was displayed from 960 ms. In the second image, the agent's head and eyes were aimed at the target from 960 ms till 3000 ms.

c) Dynamic cue: A fully animated agent, showing naturalistic movement from 0 ms to 960 ms. The agent's head and eyes were pointing straight forward at 0 ms, before the agent moved (at 25 fps) to aim its head and eyes at the target area. The agent's gaze and head were aimed at the target at 960 ms. The full orientation cue was therefore presented from 960 ms till 3000 ms.

C. *Participants*

A total of sixteen participants were recruited from students and staff at the University of Abertay-Dundee. There was no compensation and all had normal or corrected-to-normal vision. During the experiment, two of them used contact lenses.

D. *Apparatus*

To capture participant gaze data, a modified (fixed position) SMI IView HED eye-movement recorder with two cameras was used. One camera recorded the environment (the target monitor) and the other tracked the participant's eye by an infrared light recording at a frequency of 50 Hz and accuracy of 0.5° of visual angle. Stimuli were presented on a TFT 19'' monitor with a 1024 x 768 resolution and 60Hz of frequency controlled by a separate PC. The monitor brightness and contrast were set up to 60% and 65% respectively to ease the cameras' recordings and avoid unnecessary reflections. Also, both devices were individually connected to two different computers. Viewing was conducted at a distance of 0.8 meters in a quiet experimental chamber.

Each participant underwent gaze calibration controlled by the experimenter prior to the start of data collection. The participant was sat down in a height adjustable chair with their chin on the chin rest and in front of the monitor at 0.9 meters distance. Firstly, the calibration of the eyetracker was completed by presenting a sequence of five separate screens with dots in the center and in the corners. The calibration covered the same surface occupied by the target areas.

A final image with the set of five points was shown to double check the calibration by the operator. The calibration

 ISBN: 978-1-61208-110-6

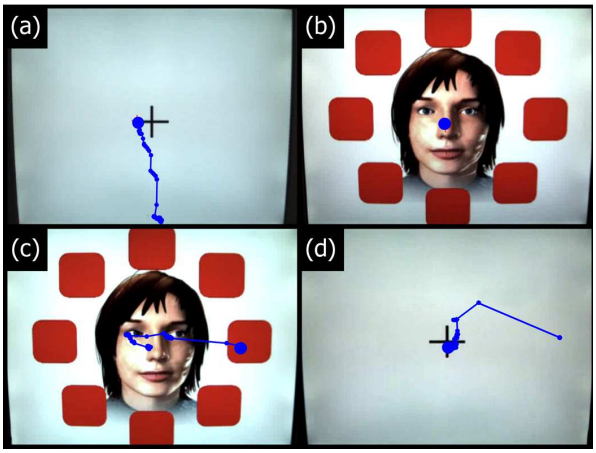

Figure 3: The eye tracking data of one participant, where the blue circles represent fixations. In image (a), the participant looks towards the cross before the agent appears in image (b). In image (c), the agent highlights the East target, at which point the participant looks towards this target, before fixating on the cross again in image (d).

was repeated if necessary following adjustments to the camera positions to ensure good calibration. The experiment started with a ten seconds countdown sequence. After that, the series of twenty-four videos (3 agent cue types x 8 target areas) were presented to participants in a randomized order. The duration of each task video was three seconds, and each video was shown one by one full screen. Before each task video, a central black cross over a white background was shown for two seconds to center the gaze of the participant. This ensured that the participant was looking at the centre of the screen at the start of each video. Fig. 3 shows sample screen captures from the eye-tracker.

E. Data analysis

The participant gaze data was analyzed using the software BeGaze 2.3. The data stored in BeGaze contained all the fixations' timestamps. Only trials where the participant's gaze started on the cross in the center of the screen were considered valid. Target selection was defined by the first full-gaze fixation occurring in the eight predefined areas of interest overlying the 8 target destinations. The fixation duration criterion for an observer response is defined in the light of previous literature. Ware and Mikaelian (1987) used 400 ms; Sibert and Jacob (2000) considered 150 ms. Because extended forced fixation (400 ms) can become laborious, we established a criterion for a successful cognitive response to fixation as equal or greater than to 250 ms, i.e., a fixation that locates on the target area at least for 250 ms.

Based on this concept, of the total number of possible cognitive responses, 92.18% were successfully tracked. Of the successfully tracked data, correct responses accounted for 95.2% of the total and mismatches accounted for 4.9%. The definition of a mismatch was when there was a fixation of 250 ms or more inside an incorrect target area. In 8.47% of the total mismatches, no clear target was selected – i.e., there was no fixation of 250 ms or more in any of the target areas.

TABLE I. PARTICIPANT SELECTION OF TARGETS

Type	Correct	Incorrect	No Target selected	Corrupt (Exclusions)
Static	95 %	5.7 %	0.8 %	7 / 128
Stepped	92.5 %	5.8 %	1.7 %	8 / 128
Dynamic	94.2 %	5 %	0.8 %	7 / 128

IV. RESULTS

Only one participant presented problems during the tracking because of the unexpected movement of her contact lens in the tracked eye. This resulted in four non-tracked responses in the same participant.

For each agent type a total of 128 eye tracking recordings were made. Recordings were then evaluated and allocated to one of four categories: Correct (where the observer clearly selected the intended target), Incorrect (where the observer clearly selected an unintended target), No Target (where it was not clear which target the observer had selected), and corrupted (where the eye tracking data had been disrupted resulting in lost data, for instance interference from reflections or other light sources). After excluding the corrupted recordings, it was clear that observers were able to accurately select the intended target regardless of whether the virtual agent was static (95%), stepped (92.5%), or dynamic (94.2%) (see Table I). This would suggest that, in general, the type of virtual agent (in terms whether it was static, stepped, or fully animated) did not substantially impact upon how effective it was at communicating what the intended target was.

A repeated measures analysis of variance (ANOVA) was used to determine whether agent type had an effect on how long it took participants to look at and select the intended target square. The response times for static agent cues - which contained agents which were oriented towards the target 960 ms earlier than both stepped and dynamic cues – were corrected to account for this difference. The analysis showed that the type of agent did have a significant effect on participant response time, $F(2, 30) = 52.73$, $p < .001$. Participants responded most quickly to the dynamic (fully animated) agent type ($M = 1220$, SE 95) than they did to either the stepped (2 frame) agent type ($M = 1874$, SE 61) or the static (1 frame) agent type ($M = 2091$, SE 59) (see Fig. 4).

Comparisons between agent types were assessed using the Bonferroni post-hoc test. The results showed that participants responded to the dynamic agent type significantly more quickly than both the static (Mean Deviation (MD) = 870, $p < .001$) and the stepped (MD = 654, $p < .005$) agent types. Furthermore, participants also responded to the stepped agent type significantly more quickly than the static agent type (MD = 217, $p < .005$) (see Table II). These results not only underline that static agent

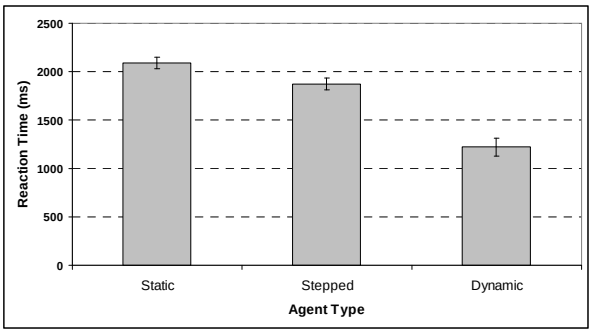

Figure 4: The mean response times for static, stepped, and dynamic agents indicate that participants reacted most quickly to the fully animated, dynamic agents

types are significantly less effective at cueing observer attention than either stepped or dynamic agents, but also that stepped agent types are significantly less effective than fully animated, dynamic agents.

V. DISCUSSION AND FUTURE WORK

Using a paradigm where the criterion for correct response to pictorial or animated agent gaze is the eye-gaze of the participant we found that the presence of full-motion in the gaze inducing agent drives the observer's attention the fastest. Gaze recorded responses for 25 frame stimuli were 35% faster than stepped and 42% faster than static stimuli. This result is consistent with previous research on gaze cueing [9]. The current paradigm provides the most direct route to the establishment of the overt allocation of gaze location since it subverts the need for a translation to a manual response. This confirms Ware and Mikaelian's [15] assertion that participants eye-gaze itself can be used to indicate responses.

The presence of movement in gaze cueing stimuli seems to drive the user's attention more quickly. One prediction arising from this is that, when compared with 2D agents, 3D agents make the expectations of more believable behaviour. The combination of additional pictorial cues and natural motion may make the appearance of the agent more akin to that of a human conversation partner. The additional realism possible with modern computer animation techniques may make agents more believable and engaging [14].

The present study indicates how the animation of an agent can be linked to the sequencing of the social 'script' or 'narrative' of a HCI interface experience. Previous investigators such as Kendon [6] observed a hierarchy of body movements in human speakers; while the head and hands tend to move during each sentence, shifts in the trunk and lower limbs occur primarily at topic shifts. They discovered the body and its movements as an additional part of the communication, participating in the timing and meaning of the dialogue. Argyle and Cook [1] discuss the use of deictic gaze in human conversation. They argued that during a conversation the gaze serves for information seeking, to send signals and to control the flow of the conversation. They explained how listeners look at the

TABLE II. MULTIPLE COMPARISONS BETWEEN AGENT TYPES

Type	Comparison	Mean Deviation	Std. Error	Sig.
Static	Stepped	217 ms	54.3	.004
	Dynamic	870 ms	85.8	.000
Stepped	Static	-217 ms	54.3	.004
	Dynamic	654 ms	114.3	.000
Dynamic	Static	-870 ms	85,8	.000
	Stepped	-654 ms	114.3	.000

speaker to supplement the auditory information. Speakers on the other hand spend much less time looking at the listener, partially because they need to attend to planning and do not want to load their senses while doing so. Preliminary work from our laboratory suggests that experience in the gaze task over time may lead to a learning effect whereby extended exposure to these stimuli leads to improved gaze allocation, this analysis will form part of a wider study including a sequence of guided navigation prompts in a naturalistic setting. Only by creating a natural sequence of user choices with a combination of gaze cues and items competing for attention (including distractors) can we fully confirm the efficacy of an agent-based cue in human computer transactions in the natural environment. The research here is consistent with the wider conclusions of other investigators [14], which indicate that vivid, animated emotional cues may be used as a tool to motivate and engage users of computers, when navigating complex interfaces. The results of this experiment provide guidance for agent design in consumer electronics such as computer games or animation. In order to avoid an unpleasant robotic awareness, natural motion and the correct presentation of the cue contribute to increase the deictic believability of the agent. Deictic believability in animated agents requires design that considers the physical properties of the environment where the transaction occurs. The agent design must take account of the positions of elements in and around the interface. The agent's relative location with respect to these objects, as well as social rules known from daily life, are critical to create deictic gestures, motions, and speech that are both effective, efficient and unambiguous. All these aspects have an effect in addition to the core the response time measure. They easily trigger natural and social interaction of human users, reaching the right level of expectations. Furthermore, they make the system errors, human mistakes and interaction barriers more acceptable and navigable to the user [2].

ACKNOWLEDGMENT

Martinez is grateful for support from the Alison Armstrong Studentship. Sloan is grateful for support from a University funded studentship. The authors gratefully acknowledge the supportive interdisciplinary research environment provided by Abertay's Whitespace Research Group, Institute for Arts Media and Computer Games and Centre for Psychology.

 ISBN: 978-1-61208-110-6

REFERENCES

[1] M. Argyle and M. Cook, “Gaze and mutual gaze”, New York: Cambridge University Press, 1976, 221 pages, ISBN-13: 978-0521208659.

[2] Diederiks, E. M, “Buddies in a box: animated characters in consumer electronics”,IUI '03, 2003, pp. 34-38, doi: http://doi.acm.org/10.1145/604045.604055

[3] J. D. Eastwood, D. Smilek, and P. M. Merikle, “Differential attentional guidance by unattended faces expressing positive and negative emotion”, Perception & Psychophysics, vol. 63, 2001, pp. 1004-1013.

[4] J. M. Findlay and I. D. Gilchrist, “Active vision: the psychology of looking and seeing”, Oxford University Press, Oxford. 2003.S. R. H.

[5] C. K. Friesen and A. Kingstone, “The eyes have it! reflexive orienting is triggered by nonpredictive gaze”, Psychonomic Bulletin and Review, vol. 5, 1998, pp. 490-495.

[6] A. Kendon, “Some relationships between body motion and speech: ana anlysis of an example”, In: A. Siegman and B. Pope, eds, Studies in Dyadic Communication, pp. 177-210, Elmsfor, NY: Pergamon Press, 1972.

[7] S. R. H. Langton and V. Bruce, “Reflexive visual orienting in response to the social attention of others”, Visual Cognition, vol. 6, 1999, pp. 541-567., doi:10.1080/135062899394939

[8] S. R. Langton, R. J. Watt, and V. Bruce, “Do the eyes have it? Cues to the direction of social attention”, Attention And Performance, vol. 4(2), 2000, pp. 50-59, ISSN 1364-6613, DOI: 10.1016/S1364-6613(99)01436-9

[9] S. R. Langton, C. O’Donnell, D. M. Riby, and C. J. Ballantyne, “Gaze cues influence the allocation of attention in natural scene viewing”, Experimental Psychology, vol. 59(12), 2006, pp. 2056-2064, doi: 10.1080/17470210600917884.

[10] I. Poggi and C. Pelachaud, “Signals and meanings of gaze in animated faces,”. In: S. Nuallain, C. Muhlvihill and P. McKevitt, eds, Language, Vision and Music. Amsterdam: John Benjamins, 2001

[11] M. I. Posner, “Orienting of attention”, Quarterly Journal of Experimental Psychology, vol. 32, 1980, pp. 3–25.

[12] L. E. Sibert and R. J. Jacob, “Evaluation of eye gaze interaction”, In Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI 00), ACM, 2000, pp. 281-288, doi: 10.1145/332040.332445.

[13] D. Smilek, E. Birmingham, D. Cameron, W. Bischof, and A. Kingstone, “Cognitive ethology and exploring attention in real world scenes,” Brain Research, vol. 1080, Issue 1, Attention, Awareness, and the Brain in Conscious Experience, 2006, pp. 101–119.

[14] T. Vanhala, V. Surakka, H. Siirtola, K. Raiha, B. Morel, and L. Ach, “Virtual proximity and facial expressions of computer agents regulate human emotions and attention”, Computer Animation And Virtual Worlds, vol 21(3-4), 2010, pp. 215-224, doi: 10.1002/cav.336.

[15] C. Ware and H. H. Mikaelian, “An evaluation of an eye tracker as a device for computer input”, SIGCHI Bull., 17, May. 1986, pp. 183-188, doi:10.1145/30851.275627.

 ISBN: 978-1-61208-110-6

The Community Network Game Project: Enriching Online Gamers Experience with User Generated Content

Shakeel Ahmad*, Christos Bouras†, Raouf Hamzaoui*, Andreas Papazois†, Erez Perelman‡, Alex Shani ‡,
Gwendal Simon §, George Tsichritzis †
*Department of Engineering
De Montfort University, Leicester, UK
sahmad@dmu.ac.uk, rhamzaoui@dmu.ac.uk
†Research Academic Computer Technology Institute
N. Kazantzaki, GR26500 Patras, Greece
bouras@cti.gr, papazois@ceid.upatras.gr, tsixritzis@cti.gr
‡Exent Technologies
Bazel 25, 49125 Petach-Tikva, Israel
eperelman@exent.com, ashani@exent.com
§Institut TELECOM
TELECOM Bretagne, France
gwendal.simon@telecom-bretagne.eu

***Abstract*—One of the most attractive features of Massively Multiplayer Online Games (MMOGs) is the possibility for users to interact with a large number of other users in a variety of collaborative and competitive situations. Gamers within an MMOG typically become members of active communities with mutual interests, shared adventures, and common objectives. We present the EU funded Community Network Game (CNG) project. The CNG project will provide tools to enhance collaborative activities between online gamers and will develop new tools for the generation, distribution and insertion of user generated content (UGC) into existing MMOGs. CNG will allow the addition of new engaging community services without changing the game code and without adding new processing or network loads to the MMOG central servers. The UGC considered by the CNG project includes 3D objects and graphics as well as video to be shared using peer-to-peer (P2P) technology. We describe the concept, innovations, and objectives of the project.**

***Keywords*-Massively Multiplayer Online Games; user generated content; P2P streaming; graphics insertion.**

I. INTRODUCTION

Massively Multiplayer Online Games (MMOGs) allow a large number of online users (in some cases millions) to inhabit the same virtual world and interact with each other in a variety of collaborative and competing scenarios. MMOGs are rapidly gaining in popularity. Data from [1] suggests that there were over 16 million active subscriptions to MMOGs by 2008, a figure that is growing fast and predicted to rise to at least 30 million by 2012.

MMOG gamers can build and become members of active communities with mutual interests, shared adventures, and common objectives. Players can play against other players (player versus player) or build groups (guilds) to compete against other groups (realm versus realm) or against computer-controlled enemies. This paper presents the Community Network Game (CNG) project [2], a recently EU funded project within the Seventh Framework Programme. The project, which started in February 2010 and has a duration of 30 months, aims at enhancing collaborative activities between MMOG gamers. This will be achieved by developing new tools for the generation, distribution and insertion of user-generated content (UGC) into existing MMOGs. This UGC may include items (textures, 3D objects) to be added to the game, live video captured from the game screen and streamed to other players, and videos showing walk-throughs, game tutorials, or changes in the virtual world to be watched on demand.

The main technologies proposed by the CNG project are the in-game graphical insertion technology (IGIT) and a peer-to-peer (P2P) system for the distribution of live video. IGIT is an innovative technology of replacing or inserting content into a game in real time without the need to change the game code in the client or server. For example, billboards can be inserted, tattoos can be added to in-game characters, an area on the screen can be assigned to display user information, and any type of window (browser, chat, etc.) can be inserted floating on or outside the game area. The technology can be implemented on multiple games, making it possible to create a community that is not limited to a specific game or publisher.

Enabling thousands of users to communicate UGC represents a significant challenge to networks already occupied by the MMOG client-server data. The CNG project intends to develop new techniques for UGC distribution that are friendly (supportive and not disruptive) to the MMOG client-

 ISBN: 978-1-61208-110-6

server traffic. The key innovation will be a P2P system that will allow MMOG gamers to stream live video of the game without interrupting the MMOG data flow and the need to upload the video data to a central server.

The remainder of the paper is organized as follows. Section II gives an overview of the state-of-the art in the areas related to the CNG project. Section III presents CNG's proposed innovations and technologies. Section IV concludes the paper by discussing CNG's benefits and expected impacts.

II. Related Work

A. UGC and Web collaboration tools

UGC refers to various kinds of media content that are produced by end-users. In the context of a game, this may refer to screen captures and video capture from within the video game. Another example of UGC may be the various mods created by the users. Furthermore, sharing and remixing UGC is a widespread online activity that crosses borders of age and gender. Avid players go to great lengths in their efforts to create shared content in which they reveal their mastery. Additional data layers are always included: narration, animation and primarily soundtrack. UGC sharing and remixing within game platforms, one of the most important goals of the CNG project, is currently not supported. Most MMOG-based UGC content is confined to dedicated player/game company sites as in World of Warcraft [3]. Many MMOG games also have their own community pages in social networking sites such as Facebook [4]. In April 2010, Facebook released significant updates to its API by allowing external websites to uniformly represent objects in the graph (e.g., people, photos, events, and community pages) and the connections between them (e.g., friend relationships, shared content, and photo tags). As a result, the Facebook API [5] can provide an unprecedented bridge between gamespaces and the social web. Additionally, many MMOG players use sites such as YouTube in order to share their game-based UGC. In 2008, Maxis incorporated YouTube APIs within their game, Spore, by enabling a player to upload video of their creations to their YouTube account with only two clicks [6].

Web 2.0 is a trend in the use of World Wide Web (WWW) technology and Web design that aims to facilitate creativity, information sharing, and, most notably, collaboration among users. These concepts have led to the development and evolution of Web-based communities and hosted services, such as social networking sites, wikis, blogs, and folksonomies. The Web 2.0 technologies are standardized by the WWW Consortium (W3C) [7]. Although the Web 2.0 term suggests a new version of the WWW, it does not refer to an update to any technical specifications, but to changes in the ways software developers and end-users use the web. The Web 2.0 based collaboration applications may include instant messaging, audio and video chat, file sharing and online voting and polling. For audio/video capturing and playback the Flash software platform [8] is commonly deployed. Other solutions are the Java Applet technology or standalone applications which run on Web browser and offer interoperability over different platforms. For instant messaging, online polling/voting and file sharing, Asynchronous JavaScript and XML (AJAX) [9] are commonly used. AJAX allows Web applications to retrieve data from the server asynchronously in the background without interfering with the display and behavior of the existing page. The use of AJAX techniques has led to an increase in interactive or dynamic interfaces on webpages. Finally, for WWW client-server communication, most of the Web 2.0 applications are based on Simple Object Access Protocol (SOAP) [10]. SOAP relies on XML as its message format, and usually relies on other Application Layer protocols, most notably the Remote Procedure Call (RPC) and HTTP.

Table I: Game adaptation technologies in FreeRideGames (FRG), Massive Incorporated (MI), PLAYXPERT (PX), Xfire (XF), Double Fusion (DF).

Product	FRG	MI	PX	XF	DF
In-game overlay	Yes	No	Yes	Yes	No
Game resize	Yes	No	No	No	No
Texture replacement	No	Yes	No	No	Yes
Need for SDK	No	Yes	No	No	Yes

B. Game adaptation technologies

In-game technologies have been used in the gaming market for several years. The gaming industry has adopted these technologies to increase its revenue by finding more financial sources and by attracting more users. In-game overlay allows to view and interact with windows outside the game, but without "Alt-Tabbing". It does so by rendering the window inside the game. Texture replacement enables to replace an original game texture with a different texture. In this way, the newly placed textures are seen as part of the original game content. This method is commonly used for dynamic in-game advertisement. Game size modification technology adapts the original game by decreasing its original size and surrounding it with an external content. The existing game adaptation products can be divided into two groups: products that require for the game developer to integrate the products software development kit (SDK) and products that do not impose this constraint (see Table I).

C. P2P live video systems

Traditional client-server video streaming systems have critical issues of high cost and low scalability on the server. P2P networking has been shown to be cost effective and easy to deploy. The main idea of P2P is to encourage users (peers) to act as both clients and servers. A peer in a P2P system not only downloads data, but also uploads it to serve other peers. The upload bandwidth, computing power and

 ISBN: 978-1-61208-110-6

storage space on the end user are exploited to reduce the burden on the servers.

Viewers of a live event wish to watch the video as soon as possible. That is, the time lag between the video source and end users is expected to be small. In a live streaming system, the live video content is diffused to all users in real time and video playback for all users is synchronized. Users that are watching the same live video can help each other to alleviate the load on the server. P2P live streaming systems allow viewers to delete the historic data after the playback, and hence have no requirement for any data storage and backup.

Based on the overlay network structure, the current approaches for P2P live streaming systems can be broadly classified into two categories: tree-based and mesh-based. In tree-based systems, peers form an overlay tree, and video data are pushed from the parent node to its children. However, a mesh-based system has no static streaming topology. Peers pull video data from each other for content delivery. Over the years, many tree-based systems have been proposed and evaluated, however, never took of commercially. Mesh-based P2P streaming systems achieve a large-scale deployment successfully, such as PPLive [11], PPStream [12], etc.

Most P2P live video systems rely on the transmission control protocol (TCP) (as in e.g., CoolStreaming, PPStream). TCP guarantees reliable transmission of the data by automatic retransmission of lost packets. However, as TCP requires in order delivery of the data and keeps on retransmitting a packet until an acknowledgement is received, significant delays may be introduced. Further delays are caused by the congestion control algorithm used by TCP, which reacts to packet loss by reducing the transmission rate, leading occasionally to service interruption. This presents a serious drawback for real-time video communication where the data must be available to the receiver at its playback time. Lost and delayed packets that miss their playback deadline not only are useless, they also consume the available bandwidth unnecessarily. An alternative to TCP is to use UDP as the transport protocol and apply application-layer error control. This includes UDP without error control (PPLive, TVAnts), UDP with FEC [13], ARQ [14], and Multiple Description Coding (MDC) [15].

III. Technologies and Innovations

To achieve its objectives, CNG will rely on innovative software technologies and a P2P live video system. While the MMOG architecture is not modified (the game content and the game data are still transferred through the MMOG servers), the following components will be added (Fig. 1): (i) Sandbox on the client side that is responsible for modifying the game environment; (ii) CNG Server for monitoring the P2P UGC communication. The CNG server acts as a tracker for the system in the sense that it is in charge of introducing

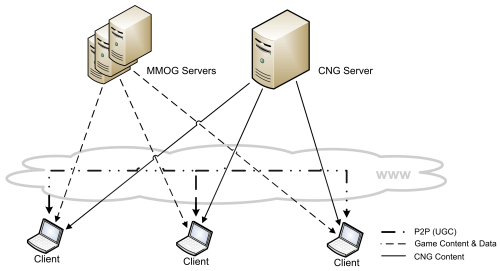

Figure 1. CNG architecture.

peers to other peers. It has persistent communication with the clients and manages the organization of the P2P exchanges.

A. UGC and Web collaboration tools

In CNG, the participation in community activities will not require closing or resizing the screen of the game and activating the tools' window. Instead, the CNG tools' window, will be integrated into the MMOG application environment. The CNG tools will use Web 2.0 technology to enable voice and video chat, instant messaging, polling, and file sharing. A flash-based collaborative video editor will be included in the CNG toolbox to allow users to edit videos and images. The system also includes tools to enable the upload of video files to social networking sites.

B. IGIT

The CNG project will enable to resize the game and surround it with external content, overlay the game, and replace an existing game texture with an external content. This will be done in a way that does not harm the game experience and without the need for SDK integration. Fig. 2 and Fig. 3 illustrate some of these features. Fig. 2 is a screenshot from the MMOG game "Roma Victor" [16] by RedBedlam. Fig. 3 shows the same game scene with a mock-up of CNG features. The modifications, which are numbered in Fig. refigit, are as follows: (1) The original resolution of the game was modified to enable an additional frame around the game to hold the in-frame objects. IGIT uses the GPU of the user's machine for changing the resolution of the game to avoid reduction in the image quality; (2) Instant messaging window as an example of active Web 2.0 application; (3) Web browser that presents online passive information (in this example, a leader board); (4) Another Web browser window that presents an updated advertisement; (5) MMOG specific chat to enable the users in a specific scene to cooperate; (6) In-game 3D UGC. In this example, a user added a note on a tree to publish an eBay auction; (7) Two windows of a video chat with casual friends or cooperative players.

The choices of which application to use and the applications' screen location are under the control of the user (player).

 ISBN: 978-1-61208-110-6

Figure 2. Original MMOG screenshot.

Figure 3. IGIT-modified MMOG screenshot.

C. P2P live video system

In existing MMOGs, a player can capture the video of the game and send it to a central server which broadcasts it live to other users [17]. However, this solution, which heavily relies on central servers has many drawbacks such as high costs for bandwidth, storage, and maintenance. Moreover, this solution is not easily scalable to increasing number of users. The CNG project intends to develop a P2P live video streaming system to address the limitations of server-based solutions. The CNG P2P live video system will allow every peer to become a source of a user-generated video stream for a potentially large set of receivers. While many P2P live video systems have been proposed, none of them has been specifically designed for the unique environment of MMOGs. In particular, none of the existing P2P live video systems addresses the following challenges:

- Any MMOG player should be able to multicast live video. The video can potentially be received by any other player in the P2P network.
- Live video streaming should not consume the upload and download bandwidth that is necessary for the smooth operation of the MMOG (MMOG client-server traffic).
- Live video should be delivered at about the same time to all peers at the same "level". This optional requirement can be useful in some situations. For example, a level can be defined as the set of all peers that are in the same region of the virtual space.

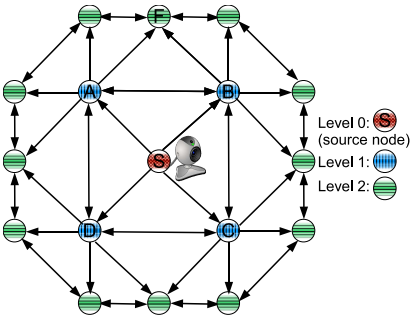

Figure 4. P2P topology. S is the source peer.

The CNG P2P live video system is designed as follows. A mesh-based topology is used for the P2P overlay. Peers are organized in different levels (Fig. 4). The source peer is placed at level 0. All peers connected directly to the source peer are at level. In general, a peer is considered to be at level j if its shortest route to the source peer consists of j intermediate links.

The video is captured in real time from the source screen, compressed, and partitioned into source blocks. Each block consists of one GOP (Group of Pictures) and is an independent unit of fixed playback duration (e.g., 1 s). The UDP protocol is used as the transport protocol. The source peer applies rateless coding on each source block and keeps on sending the resulting encoded symbols in encoded packets (packets of encoded symbols) to level-1 peers until it receives an acknowledgment. Level-1 peers forward the received packets to other level-1 peers immediately as instructed by the source peer. Level-1 peers also forward the received packets to the level-2 peers that are directly connected to them, etc. When a level-1 peer completes the decoding of a block, it sends an acknowledgment to all senders so that they stop sending it packets. Then it applies rateless coding on the decoded block to feed level-2 peers. Thus, each receiving peer has two phases: forwarding (before the decoding is successful) and encoding (after decoding the block). In the first phase, the receiving peer just forwards the received packets to the next level peers connected to it, while in the encoding phase, it generates encoded symbols from the decoded block and feeds the next-level peers.

The source peer computes a scheduling strategy for each source block. The strategy specifies the maximum number of encoded packets n that can be sent for this block, and

 ISBN: 978-1-61208-110-6

the time t_i at which packet i is sent with a hierarchical forwarding scheme F_i, $i = 1, 2, \ldots, n$. If the source peer receives an acknowledgment from a level-1 peer j before n packets are sent, it can update its scheduling strategy by, e.g., removing peer j from the forwarding schemes of the remaining packets. An example of a scheduling strategy for $n = 4$ and four level-1 peers A to D is as follows. $1 : t_1 : A \rightarrow B(\rightarrow D) + C, 2 : t_2 : B \rightarrow A + D(\rightarrow C), 3 : t_3 : C \rightarrow A(\rightarrow B) + D, 4 : t_4 : D \rightarrow C(\rightarrow A) + B$. For packet 1, the strategy says: transmit at time t_1 to A. A forwards the packet to B and C. B forwards it to D.

The complexity of the scheduling strategy depends on the neighbourhood relationships. In a clustered topology (where neighbours of a given peer are also neighbours with high probability), the scheduling strategy can become complex to decide. One of the challenges of the project consists of determining topologies, which allow efficient and simple computation of scheduling strategies.

Since a peer can have multiple neighbours, it can receive the same packet from multiple senders. To avoid this, a parent should know the other parents of its children. For example, peer B should know that F is its common child with A and not forward a packet to F if this packet has previously visited A. In the encoding phase, receiving duplicate packets can be avoided with high probability by forcing peers to use different seed values for the rateless code.

By having multiple senders, lost packets on one link can be compensated for by receiving more packets on other links. Players that are neighbours in the virtual world can be placed at the same level in the mesh, so that they can watch the video with approximately the same playback lag with respect to the original source.

Our system extends previous ideas proposed in [18], [19]. However there are many important differences between these works and our scheme. For example, the systems of [18], [19] do not have the notion of scheduling strategy and use a different approach to minimise the number of received duplicated packets. Also in [18], [19], there is no notion of levels within the mesh.

As UDP does not have a built-in congestion control mechanism, a pure UDP-based application may overwhelm the network. To address this problem, we aim to adapt the UDP sending rate according to receiver feedback. The feedback may consist of the average packet loss rate and the forward trip time (FTT). If the average FTT and loss rate are higher than threshold values, this is a strong indication of congestion. As a result, the sender has to adapt the sending rate accordingly.

Many peers can become a source of live content. However, a peer cannot participate in all overlays, because some resources are used in every overlay a peer belongs to. In practice, a user can decide whether to receive the stream from a given source. But an automatic management would be more suitable. We propose to continuously adjust the set of peers who are targeted to receive data from a source. The goal is to obtain a fair nearly congested system, where the peers that receive the stream are "close" to the user who generates the content.

We use the concept of Area of Interest (AoI) [20] for that purpose. An AoI is defined as the part of the virtual world around a user that generates content. When a peer is within the AoI of a user generating high-quality content, or when it belongs to many AoIs, it may experience congestion. The challenge is to design a mechanism for determining the best size of these AoIs, that is, a size such that the maximum amount of UGC is practically delivered while no user experiences congestion. The management of the AoI must then take into account the popularity of the virtual place and the capacity of the devices of the players that are located there. Such a management has been shown to be hard in wireless sensor networks [21], but some heuristics can perform well. For a player, the decision of increasing or decreasing the size of the AoI should be based on feedbacks from other nearby players in a collective manner.

Two strategies can be implemented. In the first one, one peer is congested, and not all peers in an AoI can be served. However, the capacity of peers in the surroundings of the congested peers makes that a new computation of AoI for all sources is not necessary. Instead, it is possible to "pass" one peer from one P2P overlay to another P2P overlay, so that the capacity provided by this peer can tackle the congestion issue. This strategy, which avoids heavy computations can solve local small congestion problems. The second strategy can be implemented when this first one fails. A process similar to the one that ensures fair resource sharing in TCP can be used. Every source periodically tries to increase (in an additive manner) the size of its AoI until congestion is detected. Then, the radius of the AoI is decreased in a multiplicative manner (see [22] for a similar technique).

In addition to designing an efficient P2P live video streaming system for a game environment, the CNG project proposes to contribute to a better understanding of the general problem of the diffusion of multiple video streams in a constrained environment. The goal is to maximize the amount of peers receiving content in an environment where not all peers can be reached because too few resources are available. If we assume tree overlays and consider only one video stream, the problem is to build a tree that spans the maximum number of peers with the constraint that every peer can only serve a limited number of other peers. In the context of many concurrent video streams, the problem becomes even harder with a constrained forest.

The building of degree-constrained trees is an NP-hard problem [23]. We propose to contribute to the analysis of the computational properties of this problem. In particular, the formulation of the problem into several Integer Programming models, and comprehensive benchmarks of these models

 ISBN: 978-1-61208-110-6

will enable the computation of optimal solutions on small instances of the problem. Besides, we aim at designing heuristic algorithms, which allow the computation of nearly optimal solutions for large problem instances, as well as approximate algorithms (algorithms that compute solutions that are proved to be never far to the optimal solutions).

IV. Conclusion

We presented the EU funded CNG project. CNG will support and enhance community activities between MMOG gamers by enabling them to create, share, and insert UGC. The UGC considered by the CNG project includes 3D objects, graphics, and video. CNG will develop in-game community activities using an in-game graphical insertion technology (IGIT). IGIT allows to replace or insert content in real time without the need to change the game's code in the client or server. CNG uses an architecture that efficiently combines the client-server infrastructure for the MMOG activities with a P2P overlay for the delivery of live video. The video traffic represents a real challenge to the network already occupied by the MMOG client-server data. The project will research and develop new techniques for P2P live video streaming that are friendly to the MMOG client-server traffic. Since video can be resource heavy, the network indirectly benefits from the increased locality of communication. CNG will also provide Web 2.0 tools for audio and video chat, instant messaging, in-game voting, reviewing, and polling. This will reduce the need for visiting forums outside the game and diluting the MMOG experience.

CNG has the potential to provide huge benefits to MMOG developers and operators. New community building tools will be offered cost-effectively and efficiently, without the need to redesign or recode the existing game offerings. The user experience will be enriched, and the needs of the end-users will be better addressed. The community will be brought into the content, and the game communities will become more engaged, reducing churn to other MMOGs. New income streams will be delivered with the help of in-game and around game advertising. Yet, MMOG developers and operators will be able to maintain control over how various commercial and UGC content is displayed, thus keeping editorial control of the look and feel of their MMOG.

Acknowledgment

The research leading to these results has received funding from the European Commission's Seventh Framework Programme (FP7, 2007-2013) under the grant agreement no. ICT-248175 (CNG project).

References

[1] [Online]. Available at: http://www.mmogchart.com. Last accessed: 24/8/2010.

[2] [Online]. Available at: http://www.cng-project.eu/. Last accessed: 24/8/2010.

[3] [Online]. Available at:: http://www.worldofwarcraft.com/splash-sc2date.htm. Last accessed: 25/5/2010.

[4] [Online]. Available at: http://www.Facebook.com/group.php?gid=2210867052. Last accessed: 25/5/2010.

[5] Facebook API. [Online]. Available at: http://graph.Facebook.com. Last accessed: 25/5/2010.

[6] Youtube API. [Online]. Available at: http://code.google.com/apis/youtube/casestudies/ea.html. Last accessed: 30/6/2010.

[7] W3C. [Online]. Available at: http://www.w3.org/. Last accessed: 25/5/2010.

[8] Adobe Flash. [Online]. Available at: http://www.adobe.com/support/flash/downloads.html. Last accessed: 26/8/2010.

[9] XMLHttpRequest, W3C Working Draft, 2009. [Online]. Available at: http://www.w3.org/TR/2009/WD-XMLHttpRequest-20091119/. Last accessed: 26/8/2010.

[10] SOAP Version 1.2 Part 1: Messaging Framework (Second Edition), W3C Recommendation, 2007. [Online]. Available at: http://www.w3.org/TR/soap12-part1. Last accessed: 26/8/2010.

[11] [Online]. Available at: http://www.PPlive.com. Last accessed: 26/8/2010.

[12] [Online]. Available at: http://www.PPstream.com. Last accessed: 26/8/2010.

[13] V.N. Padmanabhan, H.J. Wang, and P.A. Chou, "Resilient peer-to-peer streaming," Proc. IEEE ICNP, pp. 16–27, Atlanta, GA, Nov. 2003.

[14] E. Setton, J. Noh, and B. Girod, "Rate-distortion optimized video peer-to-peer multicast streaming," Proc. P2PMMS'05, pp. 39–48, Singapore, Nov. 2005.

[15] E. Akyol, A.M. Tekalp, and M.R. Civanlar, "A flexible multiple description coding framework for adaptive peer-to-peer video streaming," IEEE Journal of Selected Topics in Signal Processing, vol. 1, no. 2, pp. 231–245, Aug. 2007.

[16] [Online]. Available at: http://www.roma-victor.com. Last accessed 23/8/10.

[17] [Online]. Available at: http://xfire.com. Last accessed 25/8/10.

[18] C. Wu and B. Li, "rStream: resilient and optimal peer-to-peer streaming with rateless codes," IEEE Transactions on Parallel and Distributed Systems, vol. 19, pp. 77–92, Jan. 2008.

[19] M. Grangetto, R. Gaeta, and M. Sereno, "Reducing content distribution time in P2P-based multicast using rateless codes," Proc. Italian Networking Workshop, Cortina, pp. 1–12, Jan. 2009.

[20] J.R. Jiang, Y.-L. Huang, and S.-Y. Hu, "Scalable AOI-cast for peer-to-peer networked virtual environments," Journal of Internet Technology, vol. 10, no. 2, pp. 119–126, 2009.

[21] B. Han and J. Leblet, and G. Simon, "Query range problem in wireless sensor networks," IEEE Comm. Letters, vol. 13, no. 1, pp. 55–57, 2009.

[22] B. Han and G. Simon, "Fair capacity sharing among multiple sinks in wireless sensor networks," Proc. IEEE MASS Conference, Pisa, pp. 1–9, Oct. 2007.

[23] J. Könemann and R. Ravi, "Primal-dual meets local search: approximating MST's with nonuniform degree bounds," Proc. 35th Annual ACM Symposium on Theory of Computing, San Diego, CA, pp. 389–395, 2003.

 ISBN: 978-1-61208-110-6

Internet Business Intelligence

Hao Tan
Service Oriented Computing (SOC)
LIRIS, University Lyon 1
LYON, FRANCE
Email: ferdinandfly@gmail.com

Parisa Ghodous
SOC
LIRIS, University Lyon 1
LYON, FRANCE
Email: parisa.ghodous@recherche.univ-lyon1.fr

Jacky Montiel
ALTERNANCE Soft
LYON, FRANCE
Email: jmontiel@ntsys.fr

***Abstract*—Business Intelligence (BI) refers to computer-based techniques used in spotting, digging-out, and analyzing business data. It is mainly focused on how to dig out business data. This type of business data is a on-line web database which can be searched through their Web query interfaces. Deep Web (often called hidden web or invisible web) is composed of all the web databases. With the evolution of the "deep web", more and more researchers pay attention to the "integration" of the web database. However, to achieve this goal, it needs a complex system and many applications to work together. We are interested in an automatic extracting system to get the formulas or the lists of the results from those websites in specific domain of government procurement. To tackle this challenge, we propose a solution to create a unified interface and to inquire resources in a predefined domain. In this paper, we will discuss the automatic extracting system in several steps. First of all, the web query interfaces crawler which can execute JavaScript guarantees the coverage of the web database. Secondly, the query interface extractor and the interface integrator can allow us query all these founded web databases through a global query interface. Thirdly, the result page extractor and the result integrator can give a unified presentation. Lastly, a feedback method is developed to gather the result accuracy. A statistical model is built to improve the performance of the step 2 and 3. We assume our system is a dynamic system, which means the more we use it, the more precise results we will get.**

Keywords-schema matching; web-database integration;

I. INTRODUCTION

Deep web (often called invisible or hidden Web) is made up of massive web databases. The traditional search engines which use static URL based crawler are not able to access most of the data on the Internet. The reason is that this kind of information is hidden behind the query interfaces and does not contain a unique URL link. The recent study [20] shows that the top 3 famous search engine: Google, Yahoo and MSN only cover respectively 32%,32% and 11% of the result pages of the sample Web databases. More importantly, even if we combine the three search engines, we can only get 37% coverage. Because of this, users often have difficulties to find the sources of web database and query them to get the results. On the other hand,the interfaces are built to be queried and generate the dynamic result pages. As traditional access methods cannot accomplish the work of searching the deep web, it is imperative to find a new way to index the hidden result pages.

Deploying an automatic system to understand and extract the deep web information from the entire Internet is still impracticable based on the existing technologies. Recent research [10], [13] and [17] decomposed this into different domain-based web database integration problems. For each domain, a global interface can be built to integrate all the web databases. To enable effective access to web databases, our works focus on building a domain based, automatic web database integration system that is independent of the domain style. We choose the domain $Government\ Procurement$ to analyze and to test the performance of this system. The main purpose of such technology is to be more efficient to get the business information from the Internet as we called $Internet\ Business\ Intelligence$. This application covers all the sites found in the domain predefined and makes a unified query form to list all the analyzed data. According to this objective, we need to develop a system that has the following three features. First, it should dynamically find the data sources in a specific domain. The only input is some keywords of this domain. Second, it needs to integrate these data sources automatically, including query interfaces and the query results. This integration should not have human intervention such as predefined training samples, and query interface extractor interpreted by programmer, etc. Third, the performance should be measured and it needs a real-time feedback system to gather information to adjust the integrator. That means this system is a self-training system.

The remainder of the paper is organized as follows. In Section 2, the related works are introduced. In Section 3, the whole web-database integration system is explained and the relationships between the subsystems are clarified by diagrams. In Section 4, the Interface Extraction subsystem is well developed and in Section 5, we focused on the Interfaces Integration subsystem. A feedback method is introduced in Section 6 and we conclude in Section 7.

II. RELATED WORK

The purpose of Internet Business Intelligence is not limited to domain-based automatic web database integration. It can be easily extended to a large scaled integration system by supplying semantic ontology for additional domains. This system considers the new problems come with the developments of technology, such as JavaScript embedded form extraction. With the minimum query conditions matching and the feedback modules, we can build a high accuracy, high efficiency matching system.

The Information Integration has been studied for a long time. The early works focused on the traditional Database Integration. Li and Clifton [21] developed a semi-automatic semantic integration procedure (SEMINT) which can find the corresponding attributes. Batini and Lenzerini [2] introduced the conceptual foundation to the problem of schema integration, which integrates the different individual methodologies and gave a general guideline for future improvement. The

 ISBN: 978-1-61208-110-6

traditional database integration is often divided into two steps, the View Integration, which produces a global conceptual description of a database, and the Database Integration, which produces a global schema of the databases.

Since the late nineties, the $deep\ web$ information integration(or Web Database Integration) is coming into the view of the researchers. Early stage research focused on small-scale, reconfigured, and manual intervened systems. The recent works focus on dynamic, large-scaled systems. As the web databases usually provide a integrated query interface to let users query the databases, we may consider that a global conceptual description for each database already existed. We do not need to make such "View Integration" for Web Databases. On the other hand, in comparison with the traditional databases integration problem, we do need to "understand" or "get" the attribute descriptions from the query interfaces and query result pages. Suppose a global schema is built correctly, the search results should be also correct. Therefore the Web database Integration can be viewed in two parts, the Query Interface Integration(QII) and the Query Result Integration(QRI). Rahm and Bernstein [6] gave a survey of the approaches of traditional automatic schema matching.

The research QII is mainly based on the visual elements identification. Golin and Reissa [8] gave a specification of visual language. Zhang ([19] and [22]) gave us an approach to understand query interfaces: Best-Effort parsing with Hidden Syntax. Liu and Meng [20] presented another approach $ViDE$ based on calculating the $distance$ among visual elements. Chuang and Chang [17] introduced a context aware wrapping system which contains the peer sources to facilitate the subsequent matching and to improve the extraction accuracy. The matching between different interfaces is often determined by calculating the semantic matching times (in results of search). He and Chang [13] introduced an approach named DCM framework. This approach tried to build a complex matching which matches a set of m attributes to another set of n attributes. Madhavan and Bernstein [7] introduced a new algorithm, Cupid Matchers in comparison with the traditional schema matching, such as Linguistic based Matchers, Constraint based Matchers, Individual matchers and Combinational matchers. He and Meng [10] concerned with the E-commerce interfaces and proposed a weight based matcher. His approach tried to automatically construct a global interface and the global attribute from the query interfaces. Wang and Wen [12] proposed an approach based on query probing and instance-based schema matching techniques. He separated the schema-matching into two areas: intra-site and inter-site. In other words, the matching between results and query interfaces in the same source and the matching between different sources.

The Web data extraction is a little bit different from interface extraction. Laender and Silva [1] gave a survey of traditional web data extraction tools. These tools are often based on declarative languages, HTML structure analysis, natural languages process, machine learning, data modeling and ontology. All of these tools focused on the treatment of HTML script. Chang and Hsu [4] introduced a method to trait the HTML code and to find the repeated patterns in the query result pages. Hu and Meng [5] introduced a semantic blocks, section and data items identification system. The data items which have the same role were mapped.

Our work has several main differences from the other works. First, the new web-page parser can execute Javascript and analyze the $true$ web-page by visual relationships. Second, we are not trying to match every attributes or query conditions in every query-interface. Only the most closed query conditions will be matched. In other words, through all the query-interfaces, there exists a minimum query-condition group which permits to construct a query and to get all the records from the back-end databases. At last, a feedback system based on automation theory of Nonlinear Discrete Systems was never discussed before.

III. Automatic Web-database Integration System Architecture

In order to build such a domain based, automatic Web-database integration system, we decompose it into several parts:

1) Web database crawler: finds the web database of specific domain.
2) Interface Extraction: parses the query form to the group or element trees.
3) Interface Integration: applies to a domain specific semantic ontology and builds a mapping from the query form contents to the semantic model.
4) Query result integration: parses and matches the different result lists from the different web query interfaces.
5) Collection system: collects the responses of end-users by detecting the records rank of the search results. A click is a user action when he clicks the hyper link in the result page. We construct a table to save the times of clicks for each record of search result. The record rank can be built from this table. The mismatching query conditions and query results normally should have a record rank much lower than the correct ones.
6) Feedback and Ranking System: includes the search results ranking, element group matches ranking, interface parsers ranking, and the web database sites ranking. The accuracy and the integrity will be gathered from the clients and send to the Ranking system. The ranking system will then $FeedBack$ to the formal subsystems, like $Interfaces\ Extraction$ subsystem, $Interfaces\ Integration$ subsystem and $Query\ result\ Integration$ subsystem.

Web database crawler may be viewed as a traditional search engine with some predefined characters. This crawler can travel through the web and identify the query interfaces. With the study of [14], the depth of a web database is often very limited. According to their work, 94% Web database have a depth within 3, these database are found from 1,000,000 randomly selected IPs. Our approach to this subsection is divided into 2 steps: First, a traditional crawler finds the site root pages which contain the web database. Secondly, a JavaScript concerned crawler find out all the query interfaces in every site. $JavaScript\ concerned$ is a conception in comparison with the traditional site crawler. The traditional crawler is designed to get the static HTML code from the static links of pages. With the evolution of JavaScript and the new technology, the original HTML code does not contain all

 ISBN: 978-1-61208-110-6

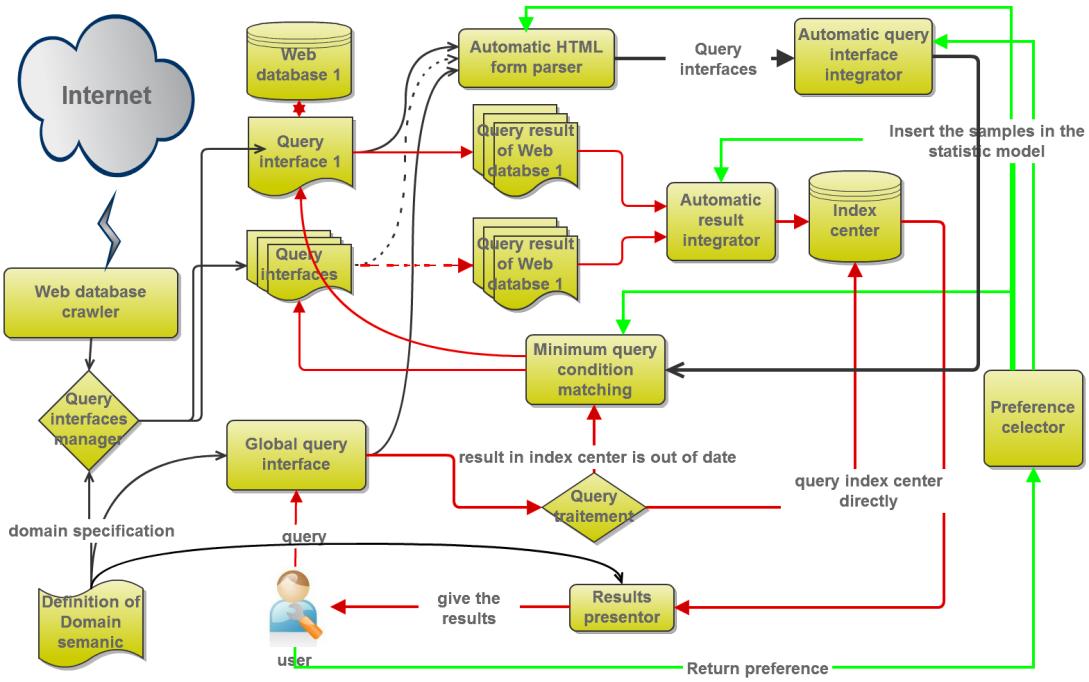

Figure 1. Web Query Interface Integration System

the information that we really see from the navigator. The final HTML code is often treated by the initialization Javascript and could be modified by the end-user's action.

A layout engine, which can execute JavaScript from a URL and simulate the end-user's action, is necessary. The most popular layout engines are Trident, WebKit and Gecko. Trident is used by Internet Explorer, WebKit is the core of the Safari and Chrome and Gecko is used in Mozilla Firefox. All the layout engines share the same idea:

- Get resources from a URL include CSS, HTML, JavaScript, image, video, etc.
- Construct the DOM tree and Render Tree. DOM tree contains all the nodes of which should be showed in the navigator. Render tree contains the visual definition of the nodes in DOM tree.
- Draw the elements of DOM Tree and take the description from Render Tree.

The new crawler embeds a layout engine to get resources from a URL and to generate the DOM tree and the Render tree. The last step, which draws the DOM elements, and consumes most of the memory and time for a layout engine, will not be executed. The DOM tree will be gathered by our crawler and query interfaces could be detected after the analysis with the semantic ontology. The render tree could also be used for further analysis, like the position of elements, size, etc.

The Interface Extraction system will focus on *understanding* the web query interfaces. The Interfaces Integration system will focus on matching the groups of query conditions from the different query interfaces. The traditional schema matching process has been focusing on identifying semantic relationships between two attributes. Why do we talk about the query conditions matching instead of the elements/attributes matching? We should take into consideration the purpose of interface integration, which is to build a mapping between query interfaces, to enable a query condition pass through all of them. Not only the query capabilities of query interfaces may not be equivalent, but also the matched attributes/elements may play different roles in the different web databases. Therefore the objective is to find equivalent query conditions but not equivalent *attributes*. Furthermore, how could we guarantee the precision and integrity of the matching, not only for the query conditions in all the interfaces, but also for the query records from the different result pages? How could we associate the responses of end-users to the interface extraction system, interfaces integration, and query-result integration?

Figure 1 describes the relationships between the function modules. The advantage of this structure is if we change the domain, the only thing that we need to modify is the manually collected Semantic Ontology. The process in Figure 1 can be simply denoted as four parts:

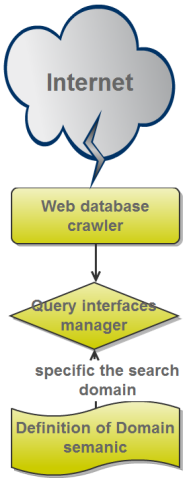

Figure 2. Web interface crawler

- Query interface crawler find the site root that contains web databases. It is showed in Figure 2
- Analyzes the structure of a known formula and then realize a semantics association between the parameters of the semantics model and the interface elements to simulate the client's query request. This part is indicated

 ISBN: 978-1-61208-110-6

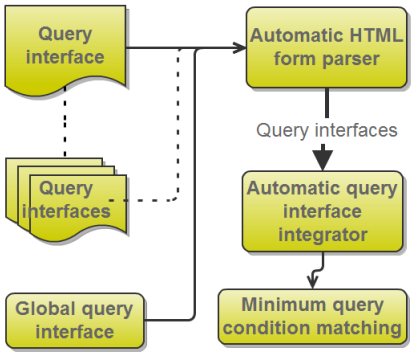

Figure 3. Interface integration

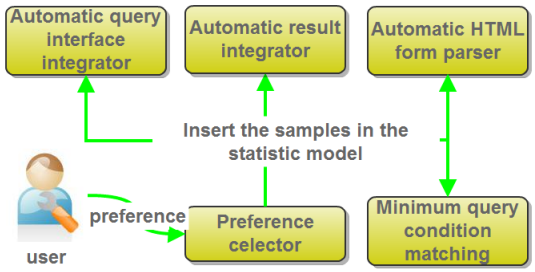

Figure 4. Feed-back system

by black arrow line and showed in Figure 3. It shows how we treat and merge the interfaces with the global interface by the definition of semantic clusters. This interfaces integration system focus on a so called ”Minimum Query Condition Matching”. This matching does not attempt to match every attribute pair found. It tries to find a $minimum$ query set across all the interfaces, which has the highest accuracy and allows our system to get all information from the back-end databases by the web query interfaces.

- Analyzes and parse the results page: it defines the structure of a result page and then identifies the relevant result fields and their associations with the semantics model parameters. This part is indicated by red arrow line.
- Analyzes the feedback from end-users and modifies the preferences of query interface matchers and query result matchers. It involves users into the matching system. Such kind of feedback mechanism can correct and improve the likelihood of successful mapping. This part is indicated by green arrow line and is showed in Figure 4.

IV. Interface Extraction

Query interfaces hide the data behind them from direct access. A form extractor is a prerequisite for the mapping works. Traditionally, we use a wrapper induction program which is supplemented by a GUI for users to click and highlight strings on a rendered Web page to produce a training example. These training examples help the program build up the patterns of the forms. If we can build all the patterns of the forms and give them relationships correspondingly, we can translate the queries from one interface to another. This kind of work is time consuming. With the development of the network, the ”manual” analysis is become impracticable. An $automatic$ extractor is thus required to a true web database query system. As a general automatic extraction (in all different domains) is almost impossible for an all-in-one integration, the recent works of automatic extraction is always ”domain-based”. We enrich the study of Zhang [19] which proposed a ”Best-effort” parsing with ”Hidden Syntax”. It treats the web interface in a ”visual” way and introduces a ”Hidden Syntax” between the positions and the relationships. The rules, named ”patterns specification”, are given and the way of extraction, named ”pattern recognition”, is specified. From the human point of view, this approach is much better than HTML pattern extractors because all the interfaces are designed for visual effect. The HTML code often contains the clues for programmers. As a result, the related tags may be placed scattered or the attributes and the corresponding elements may not have the same name. However, the visual presentation is always the main purpose of form design. So if we can make a system which extracts the visual elements, the accuracy will be guaranteed. In fact, the syntax or grammar of visual language is not a huge set and can be defined properly. On the other hand, the ”pattern recognition” is more flexible. A more general and precise method will be proposed in the following paragraph.

A. Hidden syntax

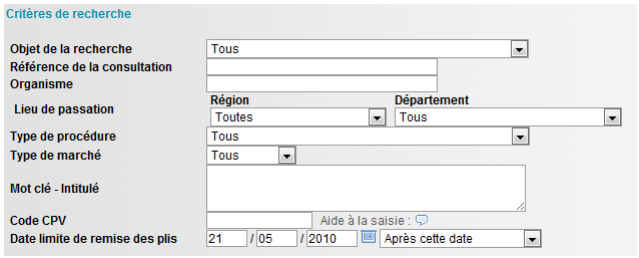

Figure 5. Query interface:Achatpublic.net

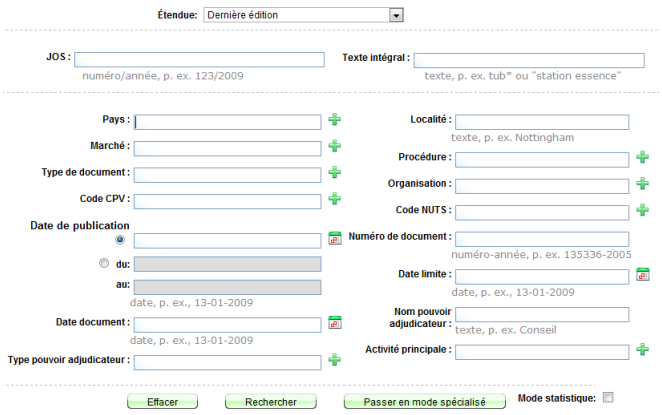

Figure 6. Query interface:TED

Suppose we have two different query interfaces as Figure 5 and Figure 6 in the domain of $Government\ Procurement$. We could easily find out some simple rules: the attribute is always on the left or above the input element. They are often in the same line. Normally, this kind of formula is built for human to read and understand. There exist many thorough studies in the domain of diagram analyzing. As Zhang [22] discussed, it assumed that there is a relationship between the semantics and the presentations behind the query-form creation, which is guided by some hidden syntax based on topology and proximity. So we can easily get a set of $pattern\ specification$,see [8]. On the other side, the form design also has some hidden syntax in the style of the HTML code for the query-form. He and Meng [9] has discussed some of the

 ISBN: 978-1-61208-110-6

rules of the HTML code. In fact, the $pattern\ discovery$ has been extensively studied, like [4] and [18]. But the main idea of their works is to find out the maximum repeated patterns which are mostly used in the query result. Normally a query-form does not contain any $repeated\ patterns$, so the method of $find\ repeat$ cannot be used directly. However, on the other hand, the definition of the pattern is useful for us if we consider the visual analysis at the same time. Therefore the **Hidden Syntax** of us is defined by a combination of the $pattern\ specification$ of HTML container and semantic relationships of the visual elements. The definition of hidden syntax could be defined as following:

The raw map of production rules: A production P in V/S grammar Σ, N, s, P_d, P_f is a four-tuple H, M, C, F. Head $H \epsilon N$ is a non-terminal symbol. Components $M \subseteq \Sigma \bigcup N$ is a multi set of symbols. Constraint C is a Boolean function defined on M. Constructor F is a function defined on M, returning an instance of H.

The rules definition are well studied in the work of [19] and [11]. The differences of the V/S grammar are that:

- P_d does not only contains the rules of visual elements, but also contains some rules of HTML patterns.
- P_f does not only contains the preference beyond the grammar, but also contains the preference by a method of the result of searching and matching.

The definition of symbol: The smallest visual elements $\bigcap$ the smallest HTML tag elements. For example, Figure 7 and Figure 8 represent the two different symbol groups for the same query condition, i.e., date range. For the 7, we find the first three input elements that are connected by '/'. The fourth input element does not close to any attribute. So we can associate the first three input elements as a date input and the fourth is a selection. These four elements with the attribute build a group.

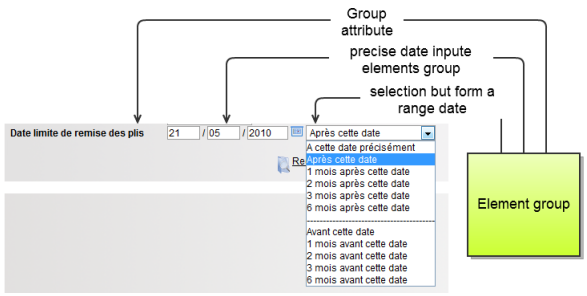

Figure 7. Date analysis: Achat public.net

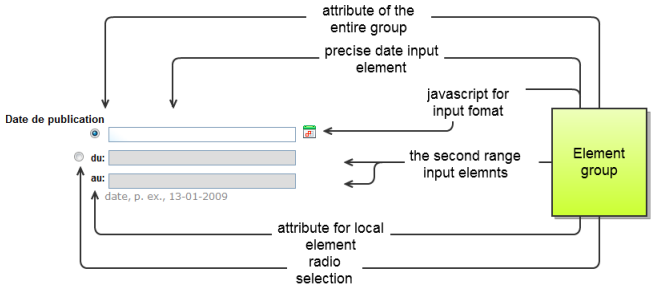

Figure 8. Date analysis: TED

B. Preferences

A preference of Σ, N, s, P_d, P_f is defined in I, U, W

- conflicting instances $I = A, B$ where $A, B \epsilon M$
- conflicting condition: a Boolean expression on A, B
- probability criteria W: a probability function on A, B that specifies the winner or both of them.

W should be considered as two branches. A threshold τ should be considered:

- $W > \tau$ and $W < (1 - \tau)$: A and B should be taken as 2 branches.
- $W < \tau$ or $W > \tau$: A and B should be taken as a confect and take A if $W < \tau$, take B if $W > \tau$

For example, in the Figure 8, the attribute "du" is very close to the ratio button, while it is also very close to the input element on its right hand. Which one should be chosen? Considering the semantic relations, it should connect to the attribute "au", and next to "au" there is only one input element on its right hand. So we prefer to link the "du" attribute with the input element on its right hand. Furthermore, we could even link the attribute "du" with the ratio button. When we try to match a range date, the attribute "du" will associate with a value 1/0, which cannot make any valuable query condition in date search. Formally, the preferences are constructed by three parts.

1) preference from the topology
2) preference from the constraint of HTML TAG
3) probability function from the responses of end-users

With the preferences, we need a parser to generate an original parse forest which eliminates all the obvious conflicts and ambiguities. In this parser, a probability function, which gathers the information from end-users, will influence the parse tree. Suppose we get $n\ pair$ different parse trees after the parse process. Each pair is decided by the three constraints. If the parse tree still contains some different parse results after we have pruned all the obvious conflicts and ambiguities, how could we decide the preferences? To answer this question, we need to know only the end-users can decide which query result is what they want. Hence, the question is how can we allow the end-user influence the preferences of the Parser. We build a subsystem to gather the feedback from end-users and build preferences for each combination of nodes. If we can choose a proper threshold of the preference, then we can choose the parse condition and resolve the conflicts or ambiguities.

C. Dynamic page

There is a problem that has never been well considered in Interface Extraction, i.e., the dynamic query interfaces which contain java script. The query interfaces often have some java script or even Ajax to modify the form structures and the form contents. For example, Figure 10 is the result of a selection from 9. Before we select a value in the selection tag, we only have a form which contains one element: select tag. After we chose something in the selection tag, we get a totally different form to extract.

Figure 9. Before choosing the option

Take account of the JavaScript, we need a layout engine which allows us to execute of JavaScript. Also, we need this layout engine to simulate the actions of a navigator such as click, drag and drop etc.. After each step of the action that we simulate, the whole parse tree should be rebuilt with grammar

 ISBN: 978-1-61208-110-6

Figure 10. After choosing the option

Σ, N, s, P_d, P_f. By comparing with the original parse tree, we got a new one. The new nods construct a new sub tree which can be pasted to the main tree. If there are several different results of JavaScript, we consider that the query interface has several different equivalent parse tree. The selection value should be considered as a search condition such as in Figure 9 and Figure 10.

V. Interfaces Integration

Query interfaces integration is a fundamental problem of Web Database Integration system. After the studies of related works, the types of matching can be described as: Query interfaces matching, Query results matching, Multi interfaces complex matching [13], Query interfaces and Query results matching. We can consider these kinds of matching is a type of $flat$ matching because they do not pay attention to the feedback of end-users. Meanwhile, if we also consider the research presentation, there is a big misunderstanding in Query Interfaces matching. The Web database integration does not really need to **integrate** or **match** every attribute in every Web database Interface. Logically it is impossible if we concern about the differences of the query abilities among the Query interfaces. Due to the purpose of our system, forming a unified result to end-users, we only need to integrate/match the main attributes/query conditions or the $clue$ of elements/attributes groups which enable our query system to query through all the Web databases and get all the information behind the interface. The purpose of interface integration is not to find the entire attribute pairs or group pairs and match them. On the contrary, the purpose is to find a minimum matching set to query them. Hence our Web database integration system should solve the following two problems:

- how can we find the minimum set of attributes which allow us to query all the information of the database. This attribute set could be the most easily matched set.
- How to match them.

To answer these two questions, we assume every pair of matching between global interface and web query interfaces have preferences, and the groups of the matching of all interfaces have preferences. A very high threshold of preference is defined as default and a small subset of matches can be selected. The selected matches are used to be the minimum matching set to query the web databases. The results-queries matching will be executed. For instance, if the results do not have matching with the queries for one of the Web Database, we consider our minimum subset is not suitable for it. The system will cut down the threshold and rebuild the minimum subset of matches.

A. Interface Analysis

First, elements of the query interface have generally the following types of format: **text box, ratio button, check box, and selection list**. Text box allows an input ”infinite”. A selection list provides a pre-determined value which is ”finite”. Radio box is like a single-select list. Check box is like a multi-select list. So we can say there are only two types of format: String S/ array()K, where S is infinite and array K is finite. Secondly, the elements have some types of relationships that need to be grouped together:

- Range type: it refers to the situation where two or more elements are used to specify the range.
- constraint type: like radio box: last name / first name should be grouped with element which has an attribute ”name”. An element could be used as a constraint for another element. The constraint could be ”or” if they are radio box, and could be ”and” if they are check box.

B. Semantic relationships for elements matching

We identify three semantic relationships among elements or the groups of elements. Suppose they have some kind of $matching$. If they do not have any relationships or exclusive relationships, it is not discussed here.

- **Synonymy** $T1$ is a synonym of $T2$
- **Hypernymy** $T1$ is more generic than $T2$,denote $H(T1, T2)$
- **Meronymy** $T1$ is a part of $T2$, denote $M(T1, T2)$

There is a case that should be noticed: T1 is a group of two text boxes, attribute is ”price”. T2 is an element of select list, attribute is price, the list is (0-100,100-200,200-400,¿ 400) we can say they are Synonymy because they play the same role in the query system. But all the option of T2 is a Meronymy of T1. For example, T1 condition is 0-300, the correspondence of T2 should be 3 query combined together: (0-100) $\bigcup$ (100-200) $\bigcup$ (200-400). More generally, this phenomenon deduces another important problem: Query translation.

C. Query translation

To translate Q_s from T1 to T2 in the above example, there are 3 query heterogeneities defined in [22]:

- **Attribute level**: Two sources may not support query the same concept or may query the same concept using different attribute names.
- **Predicate level**: Two sources may use different predicate templates for the same concept. For instance, price predicate template in T2 is a different set of value range from T1.
- **Query level**: Two sources may have different capabilities of querying valid combinations of predicate templates. In this case, the query translation is almost $impossible$.

According to the three different query heterogeneities, we can consider our minimum attribute set problem as: a group of attribute which can be matched from all the web database interfaces with no query heterogeneities, e.g., Figure 7 and Figure 8. The capability of Figure 7 can only define a period

 ISBN: 978-1-61208-110-6

of several months before or after a certain date which is given by a formal input value. If we need a period from 01-01-2008 to 15-02-2008, the closest condition is 2 month after 01-01-2008. For the interface of Figure 8, we need to choose the second option of the radio and just give the exact time interval. Moreover, in the query interface of Figure 5 we do not have the condition $pays$ but in the second interface of Figure 6, it has. So the query condition of $pays$ can never be matched between the two interfaces. There are two possibilities: Figure 5 contains only the information of one country, or Figure 5 do not distinguish the property $pays$. We cannot solve the matching problem by just analyzing the semantic means, a statically method based on the feedback system will be introduced in sector 5.

D. Matching discovery

1) Preparation of matching discovery: We take query interfaces as flat schema with sets of attribute entities. Due to domain-specific integration, we need to define a semantic ontology for the domain. Relying on the semantic ontology, a set of groups of attribute can be built from the web query interfaces. According to our study, the projects which are being carried out from other's studies always have a global interface and some kind of groups of attribute. The matching discovery problem is to find out the attribute entities matching to the global interface.

The matching discovery needs another precondition, which is data processing step. The data processing for our system here is attributes normalization. We consider the text of attributes like the natural languages. There are many studies about natural language normalization or attribute normalization such as [13]. The data preprocessing step consists of

- standard normalization [16] is a process of removing the commoner morphological and the flexional endings from words in English. It is mainly used as part of term normalization process that usually done when setting up Information Retrieval systems.
- normalize irregular nouns and verbs
- removes common stop words

2) Merging: All kinds of merging intends to construct the semantic relationship groups. This process will reduce the quantity of the final semantic relationship groups. The less we have the semantic relationships groups, the easier we establish the matching.

Type recognition Sometimes the same attribute name could have different meanings if they are in different types.

Syntactic Merging

- $name\ base\ merging$: It is based on the occurrence frequency of attributes. The statistics result of occurrence will give some preferences of merging rules.
- $domain-based\ merging$: It is based on the occurrence frequency of attribute groups. If some groups are similar and have a high occurrence, even they denote to different meanings, we can merge them together.

E. Matching construction

The approach of [13] is much more convincing than the works of [3] and [10]. He and Chang [13] introduce a ranking system who considers the samples of $N : M$ matching among the query interfaces. It does not consider any influence of the feedback from end-users. The accuracy and the integrity of matching can only be judged by the end-users. We will introduce the matching construction system which includes two steps, $Rank\ and\ selection$.

Given a set of discovered matching candidates, $R = \{M_1, M_2, ..., M_V\}$, the ranked matching is $R_C = \{M_{t_1}, ..., M_{t_V}\}$. For a n-array complex matching M_j in $R : G_{j_1} = G_{j_2} = ... = G_{j_\omega}$, C_{max} the maximal m_n value among pairs of groups in a matching is : $C_{max}(M_j, m_n) = max m_n(G_{j_r}, G_{j_t}), \forall G_{j_r}, G_{j_t}, j_r \neq j_t$.

We rank matching with the following rules:

1) if $s(M_j, m_n) > s(M_k, m_n), M_j$ is ranked higher than M_k
2) if $s(M_j, m_n) = s(M_k, m_n)$, and $M_j \succeq M_k, M_j$ is ranked higher than M_k
3) the $\succeq$ is a semantically subsumption which is described as a "top-k" approach.

In the theory of [13], the ranked matching R_C only considers the relationship between interfaces, the match ratio does not depend on any feedback of the end-users. Our consideration is based on this fact: in the statistical point of view, a correct matching will get more click ratio than a mismatching. Now the problem for us is how to decide the feedback coefficient.

F. Matching Selection

The selection rules are:

1) among the remaining matching in RC, choose the highest ranked matching Mt.
2) Remove matching conflicting with Mt in RC.
3) If RC is empty, stop; otherwise, go to step 1.

The main purpose of these rules is to remove all the conflicts and keep the highest ranked items.

VI. Feedback ranking integration

As the Figure 4 shows, the click ratio of end-users will be gathered by the system. We take the two examples of Figure 5 and Figure 6. The query condition pay is not contained in interface 1 but in interface 2. Suppose end-users execute queries by our global interface of government procurement. They want to find out all the procurement contracts. Suppose Figure 5 contains only the information of French government and Figure 6 contains the information of Europe. The matching between the two interfaces does not contain the condition "pay". Our system will not find any difference between the two web databases at the beginning. After this system has ran for a period of time, the feedback from the end-users will be distinguished. When condition "pay=French" is selected in the global interface, suppose the condition "pay" exist in our global interface, interface of Figure 5 and Figure 6 will give the same result. The feedback system will get the same click ratio of the results from the two databases. When condition "pay!=French" is selected, the click ratio of the result from query interface of Figure 5 get 0 and the click ratio for Figure 6 will be very high. The feedback system will then give the query interface in Figure 5 a precondition "pay=French".

Proposition: Consider every web database is an input, denote $x_1, x_2, ..., x_n$, suppose that we have find n web database sources need to be searched. The request query is an input denote u. After the interface extraction and interface

 ISBN: 978-1-61208-110-6

matcher, each database will give us a search result, denote $y_1, y_2, ..., y_n$. The process of extraction, match denote as $k_{11}, k_{21}, ..., k_{nn}$. The transfer function could be noted as:

$$\begin{bmatrix} y_1 \\ y_2 \\ \dots \\ y_n \end{bmatrix} = u \times \begin{bmatrix} k_{11} & k_{12} & \dots & k_{1n} \\ k_{21} & k_{22} & \dots & k_{1n} \\ \dots & \dots & \dots & \dots \\ k_{n1} & k_{n2} & \dots & k_{nn} \end{bmatrix} \times \begin{bmatrix} x_1 \\ x_2 \\ \dots \\ x_n \end{bmatrix}$$

This is the transfer function of open-loop. For the close-loop function, it denotes:

$$\begin{bmatrix} y_1^n \\ y_2^n \\ \dots \\ y_n^n \end{bmatrix} = u \times \begin{bmatrix} k_{11} & k_{12} & \dots & k_{1n} \\ k_{21} & k_{22} & \dots & k_{1n} \\ \dots & \dots & \dots & \dots \\ k_{n1} & k_{n2} & \dots & k_{nn} \end{bmatrix} \times \begin{bmatrix} x_1 \\ x_2 \\ \dots \\ x_n \end{bmatrix}$$
$$+ \sum_{i=1}^{n} (\begin{bmatrix} c_{11} & c_{12} & \dots & c_{1n} \\ c_{21} & c_{22} & \dots & c_{1n} \\ \dots & \dots & \dots & \dots \\ c_{n1} & c_{n2} & \dots & c_{nn} \end{bmatrix} \times \begin{bmatrix} y_1^i \\ y_2^i \\ \dots \\ y_n^i \end{bmatrix})$$

Where $c_{11}, c_{12}, ..., c_{nn}$ are the coefficients of feedback function. Specially, $\forall n \epsilon N, \forall j \epsilon n, \sum_{i=1}^{n}(C_{ij}) = 1$. It means the sum of the probability distribution is 1. The difficulty here is how to define the presentation of x and y. If we can find the presentation function of x and y, the stability and the convergence speed can be calculated by the theory of automatic control.

VII. Conclusion

In this article, we have shown some associated mapping approaches for understanding the web interfaces and their limitations. The proposed method needs improvements and tests of performance in real conditions. Several elements are to be assessed:

- Improve the quality of the model: the user interaction allows it to quickly converge to a reliable mapping, which is the speed of convergence and stability.
- Quality of mapping techniques, configuration and optimization.

In addition, unsolved problems have been identified as:

- how can we assure the minimum matching attributes groups are sufficient for querying the web database?
- how can we find out the best parameters to assure the convergence speed and stability of the feedback system?
- how can we specify the transfer function with the known attribute matcher theory?

Further work is to evaluate this approach through a prototype of our case study in the domain of government procurement.

References

[1] A.H.F. Laender, A.S. da Silva, B.A. Ribeiro-Neto, and J.S. Teixeira *A brief Survey of Web Data Extraction Tools* ACM SIGMOD Record, vol. 31, pp.80-93,Jun 2002

[2] C. Batini ,M. Lenzerini, and S.B. Navathe *A Comparative Analysis of Methodologies for Database Schema Integration* ACM Computing Surveys, vol. 18, pp.323-364, Dec 1986

[3] B. He and K.C. Chang *Statistical Schema Matching across Web Query Interfaces* Proc. ACM SIGMOD international conference on Management of data, pp.217-228, 2003

[4] C. Chang, C. Hsu, and S. Lui *Automatic information extraction from semi-structured Web pages by pattern discovery*Decision Support Systems,vol. 35, pp.129-147, Apr 2003

[5] D. Hu and X. Meng *Automatic Data Extraction from Data-rich Web Pages* DASFAA , LNCS 3453, pp.828-839, 2005

[6] E. Rahm and P. Bernstein *A survey of approaches to automatic schema matching* VLDB, vol. 10, pp.334-350, Dec 2001

[7] J. Madhavan ,P. Bernstein, and E. Rahm *Generic Schema Matching with Cupid* Proc. of the 27th International Conference on VLDB, pp.49-58, 2001

[8] E.J. Golin and S.P. Reissa *The specification of Visual Language Syntax* Journal of Visual Languages and Computing, vol. 1, pp.141-157, Jun 1990

[9] H. He, W. Meng, C. Yu and Z. Wu *Constructing Interface schemas for Search Interfaces of Web Databases* WISE, LNCS 3806, pp.29-42, 2005

[10] H. He, W. Meng, C. Yu and Z. Wu *WISE-Integrator: An Automatic Integrator of Web Search Interfaces for E-Commerce* Proc. of the 29th international conference on VLDB, vol. 29, pp.357-368, 2003

[11] G. Costagliola and V. Deufemia *Visual language editors based on lr parsing techniques* 8Th International Workshop On Parsing Technologies, IWPT'03, pp.79-90, 2003

[12] J. Wang, J. Wen, F. Lochovsky and W. Ma *Instance-based Schema Matching for web databases by Domain-specific Query Probing* Proc. of the 13th international conference on VLDB, vol. 30, pp.408-419, 2004

[13] B. He and K.C.C. Chang *Automatic complex Schema Matching across web Query Interfaces: A Correlation Mining Approach* ACM Transactions on Database Systems, vol. 31, pp.346-395, Mar 2006

[14] K.C.C Chang, B. He, C. Li, M. Patel and Z. Zhang *Structured databases on the web: Observations and Implications* ACM SIGMOD Record, vol. 33, pp.61-70, Sep 2004

[15] M.R. Genesereth and A.M. Keller *Infomaster: An Information Integration System* Proceedings of the 1997 ACM SIGMOD international Conference on Management of Data, pp.539-542, May 1997

[16] Martin Poter *The porter stemming algorithm* http://tartarus.org/\~martin/PorterStemmer/ unpublished

[17] S. Chuang and K.C.C Chang *Context-Aware Wrapping: Synchronized Data Extraction* Proc. of the 33rd international Conference on VLDB, pp.699-710, Sep 2007

[18] Y. Yang and H.J. Zhang *HTML Page Analysis Based on Visual Clues* Proc. 6th International Conference on Document Analysis and Recognition, pp.859-864, Sept 2001

[19] Z. Zhang, B. He and K.C.C Chang *Understanding Web Query Interfaces: Best-Effort Parsing with Hidden Syntax* Proc. of the 2004 ACM SIGMOD, pp.107-118, June 2004

[20] W. Liu, X.F. Meng, and W. Meng *ViDE: A Vision-based Approach for Deep Web Data Extraction* IEEE Trans. on Knowl. and Data Eng, vol. 22, pp.447-460, Mar 2010

[21] W.S. Li and C. Clifton *SEMINT: A tool for identifying attribute correspondences in heterogeneous databases using neural networks* Data & Knowledge Engineering, vol. 33, pp.49-84, 2000

[22] Z. Zhang *Large scale information integration on the web: finding, understanding and querying web databases* unpublished, PhD thesis, 2007

 ISBN: 978-1-61208-110-6

Drafting 2D Characters with Primitive Shape Scaffolds

Golam Ashraf
School of Computing
National University of Singapore
gashraf@nus.edu.sg

Kaiser Md. Nahiduzzaman
School of Computing
National University of Singapore
kaisernahid@gmail.com

Nguyen Kim Hai Le
School of Computing
National University of Singapore
dcslnkh@nus.edu.sg

Li Mo
School of Computing
National University of Singapore
limo1985@gmail.com

Abstract - **Primitive shapes, namely circle, triangle and square, form the basis of design and cognition of a large number of objects we find around us. Inspired by this, we have recently proposed a primitive shape field representation that simplifies detailed convex shapes. In this paper, we apply this representation in image annotation, character drafting, and repurposing of 2D artwork. We implement a character design system that allows artists to sketch rough drafts using body-part scaffolding and then retarget a pre-annotated library image onto the draft mannequin. We also compare this algorithm with Free Form Deformation lattice deformers. This allows them to quickly create an estimate of the desired character without spending effort in inking and painting. The key technical algorithms presented here have a wide range of applications, such as structured shape design (humanoid cartoons, vehicles, consumer products, etc.), content retrieval, morphing, cartoonification/stylization, and procedural detailing (textures/shapes). The key technical contributions of this paper are vector fitting of strokes and a novel primitive cage field image-warping algorithm to warp articulated characters.**

Keywords - character design, shape scaffold, deformation, image warping.

I. INTRODUCTION

One of the first things children learn is to manipulate and express with primitive shapes. Even as adults, we naturally tend to decompose complex compositions into primitives. Basic shapes like triangles, circles and squares are so well understood, that even a textual/verbal description of structures in terms of these shapes elicits a natural visualization in our brain. Basic shapes play an important role in design drafts [2, 4, 7, 10, 11, 12, 19, 21]. For example, artists use shape scaffolding to pre-visualize the final form, using basic shapes to represent each component or part. Apart from establishing the volume and mass distribution of the figure, these shapes may also help portray a certain personality, as is widely seen in stylized cartoon drawings. For example, in Pixar's recent animated feature titled "UP", the main protagonist had distinctively square features to highlight his "cooped-in" life. The square features were amplified by contrasting with a large round nose, as well as distinctly rounded supporting characters. Depending on the art style, primitive shapes may become less apparent with the addition of details; e.g. clothes, accessories, and hair for humanoid figures.

The main contribution of this paper is the formulation of a continuous primitive-shape field to address art creation, manipulation and understanding. It is inspired by three strong potentials: a) Complex shapes are often cognitively processed as groups of primitives; b) Complex shape design often begins as a scaffolding of primitive shapes; c) Though details may obscure component shape cues in different degrees, and it may be hard to synthesize these details from bare scaffold versions, the underlying scaffold could provide useful anchors for manipulation. We feel that primitive shape fields add strong intuition to the manipulation and understanding of organic shapes. This paper provides implementation details on vector shape fitting of input-strokes (for intuitive art interface), and shape-field based image warping (for retargeting existing character art to the input drawing). The proposed algorithms produce compelling results with minimal setup and an uncomplicated interface.

Without losing generality, we limit the scope of the problem of designing complex shapes with a known structure to humanoid character drawings. The paper addresses practical challenges of implementing a 2D character visualizing system for pre-production artists. Artists usually create several draft drawings for each character, first with rough shape scaffolds, and then with details for a few promising ones. Our proposal can substantially speed up this process by allowing them to draw a rough shape scaffold, and quickly map appropriate detail templates from the image library. It also allows for filter and stroke based shape refinement on the drawn scaffold. This gives artists a lot of freedom, and takes away the tedium of manually detailing every prospective sample, most of which will be probably rejected by the art director anyways.

The rest of the paper is organized as followed. Section 2 has the related work to our paper. In Section 3, we will explain our approach and system in details. Section 4 will have some comparisons of our results with the Free Form Deformation (FFD) algorithm to prove our better performance in warping method regarding to discontinuity artifact caused by FFD.

 ISBN: 978-1-61208-110-6

II. RELATED WORK

Shape description and shaper representation is a well studied field because of its tremendous importance in pattern recognition and computer vision [5, 17, 23, 25]. These methods can be classified according to several criteria [14]. The first classification is based on the use of shape boundary points as opposed to the interior of the shape. These two approaches are known as external and internal, respectively. Another classification can be made according to whether the result is numeric or non-numeric. The scalar transform techniques map the image into an attribute vector description, while the space-domain techniques transform the input image into an alternative spatial domain representation. The third classification can be made on the basis of whether a transformation is information preserving or information losing. There is also an approach called mathematical morphology which is a geometrical based approach for image analysis [14]. It provides a potential tool for extracting geometrical structures and representing shapes in many applications. Inspired by all these developments and from the fact that primitive shapes like circle, triangle and squares play a central role in human perception we developed the shape descriptor with a scaled/rotated/blended combination of these three primitive shapes [15]. Our descriptor can approximate any convex shape with a mixture of these three primitives. Every arbitrary shape is represented as a vector of height, width, rotation, centroid-position and three weight values for circle, triangle, and rectangle.

Sketching: Schmidt et al. [25] explain the importance of the scaffolding technique in their review of sketching and inking techniques used by artists. In this method, artists construct characters from basic blocks representing different body parts. Our paper addresses this need for rapid abstraction of these basic blocks from rough strokes. Thorne et al. [35] proposed the concept of sketching for character animation, but do not include shape modeling. Orzan et al. [22] propose "Diffusion Curve" primitives for the creation of soft color-gradients from input strokes, along with an image analysis method to automatically extract Diffusion Curves from photographs. Schmidt et al. [26] propose “ShapeShop”, a 3D sketch authoring system generating implicit surfaces, with non-linear editing via a construction history tree. Although these curve-based methods are intuitive, they require a fair amount of detailing. Thus they are inappropriate for rapid drafting. Our primitive blocks are a lossy abstraction of detailed convex shapes, and thus are easier to represent, construct and perceive.

Deformation: Laplacian deformation allows user-specified tweaks to one or a few points on the deformable surface, to be smoothly propagated to the vicinity. The tweaks are treated as hard constraints and the aim is to find an optimal deformation to satisfy them [3, 14, 31, 32]. Igarashi et al. [14] first proposed an interactive system that lets user deform a two-dimensional shape using a variant of constrained Laplacian deformation. In this system the shape is represented by a triangle mesh and the user moves several vertices of the mesh as constrained handles. The system then computes the positions of the remaining free vertices by minimizing the distortion of each triangle. A two-step closed algorithm is used instead of physically based simulation in order to achieve real-time interaction.

By combining locally optimal block matching with as-rigid-as-possible shape regularization, Sykora et al. [33] proposed a geometrically motivated iterative scheme to register images undergoing large FFD and appearance variations. They also demonstrated the usability of their scheme in tasks required for the cartoon animation production pipeline including unsupervised tweening, example-based shape deformation, auto-painting, editing and motion retargeting. The embedding lattice in this system consists of several connected squares. In this case local rigid transformations are computed individually for each square and then the global smoothing step is used to ensure consistency. This extension is performed to enable more flexible deformation and to preserve local rigidity of the original shape. The algorithm produces similar results to [24] but allows smooth control over shape rigidity.

In FFD methods [27], the displacement of a cage control-point influences the entire space inside the lattice. However, specifying mesh deformations this way is both cumbersome and counterintuitive. Griessmair and Purgathofer [9] extended this technique to employ a trivariate B-spline basis. Though these methods are simple, efficient and popular in use, they suffer from the drawback of a restrictive original volume shape. Parallelepiped volumes rarely bear any visual correlation to the objects they deform and typically have a globally uniform lattice point structure that is larger than is required for the deformations to which they are applied. EFFD [5] is an improvement as it allows user-specified base-shapes, but manual lattice creation and deformation are still cumbersome [6].

MacCracken and Joy [18] use a volume equivalent of the Catmull-Clark subdivision scheme for surfaces to iteratively define a volume of space based on a control point structure of arbitrary topology. This is a significant step in increasing the admissible set of control lattice shapes. The technique is powerful and its only real shortcoming is the potential continuity problems of the mapping function (a combination of subdivision and interpolation) of points within the volume. The approach also suffers from the same discontinuity problems as Catmull-Clark surfaces at extraordinary vertices [30].

Exposing mathematical parameters for indirect manipulation via a GUI interface has two major disadvantages. Firstly, there is no intuitive connection between these parameters and the user-desired manipulation. Secondly, deformations defined using the handles of a specific representation cannot be trivially applied to other shape representations or even different instances of the same shape representation [1]. Integrated bone and cage deformation systems avoid potential artifacts that may arise

 ISBN: 978-1-61208-110-6

in case of independent localized cages [34]. An interactive system that lets users move and deform a two-dimensional shape without manually establishing a skeleton or FFD domain beforehand was presented by [14].

Image Warping: Previous works use FFDs [8] or point features [16] to warp images. They typically use PCA to reduce the feature space, but the final deformation proceeds in the detailed triangulated mesh space. We propose a continuous pixel-based warp algorithm that does not need triangulation and hence avoids sampling and fold-over problems. We present some details for completeness here, and discuss additional details on resolving ambiguous pixels shared by multiple body parts. This algorithm provides a stable, fully automated and economical alternative to linear blend skinning based methods. Furthermore, we also compare its performance with FFD lattice deformation.

III. SYSTEM AND IMPLEMENTATION

We describe a system that allows body-part annotation of pre-existing orthographic character images, and then correctly retargets them to any valid draft scaffold sketched by the artist.

We briefly describe our prior work in shape representation outline in [15], and then describe the novel image warping algorithm, as we will develop on these to implement character image retargeting to draft scaffold drawings. The key technical contributions of this paper are vector fitting of outline strokes, and shape-field based image warping to resolve the artifact on human body warp.

A. Algorithms

Since we currently implement only front view drawings and images, we propose the following blocks to address these challenges: a) Vector shape representation that generates artifact-free continuous transitions between circle, triangle and square; b) Fitting a set of mouse/stylus generated strokes to the most representative vector shape for that body part; c) Vector field image warp that retargets library images to draft scaffolds.

As shown in Fig. 1, we store each of the three normalized primitive shapes as a set of eight quadratic Bezier curves. The solid points represent segment boundaries and the ragged blotches represent mid-segment control points. Note how a null segment (1-2) had to be created for the apex of the triangle. The reason why our piecewise curve segments work so well is that we were able to carefully identify the corresponding segments for the diverse topologies of circle, triangle and square. As a result, even under simple linear interpolation, we do not notice any tears or inconsistent shapes.

1) Vector Shape Representation

The normalized shapes can be affine transformed to any location, scale and rotation. Finally, the shape weights are applied to blend the corresponding Bezier control points, to yield an in-between shape. Note that start-end-mid control points of only corresponding segments are interpolated, as shown in Eqns. 1 and 2.

$$p'_j = \sum_{i=1}^{3} (w_i . p_{i,j}) \quad \text{.....(1)}$$

$$m'_j = \sum_{i=1}^{3} (w_i . m_{i,j}) \quad \text{.....(2)}$$

$$\text{where, } \sum_{i=0}^{2} w_i = 1$$

$$\text{And, } j \in \{1,2,3,4,5,6,7,8\}$$

In the above equations, p'_j and m'_j represent the *j-th* blended segment boundary and midpoints respectively, while $p_{i,j}$, and $m_{i,j}$ represent the corresponding control points in the *i-th* primitive shape (circle, triangle, square). w_i is the weight contribution from the *i-th* primitive shape. Results of some blend operations are shown in Fig. 1. The cross hair under the shapes indicates the shape weights.

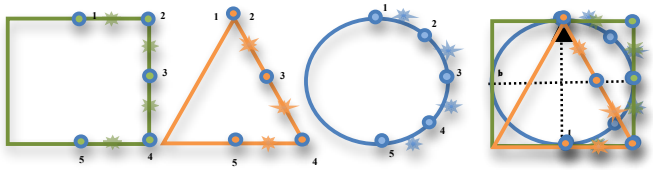

Figure 1. Consistent interpolation of circle, triangle, and square [15]

Results of some blend operations are shown in Fig. 2. The cross hairs under the shapes indicate the shape weights. With this background information about our primitive representation, we are now ready to describe vector fitting of stroked body-part line drawings. We assume that the input shapes are roughly symmetric about their medial axis, and generally convex.

2) Vector Fitting

A closed input stroke can be treated as a set of connected points, where the first and last points are fairly close to each other. We first resample the stroke at fixed angular intervals about the centroid of the input points. This helps avoid any bias due to variances in stylus pressure and stroke timing. A standard projection variance maximization algorithm, commonly employed to compute Oriented Bounding Boxes, is used to find the medial axis. In this algorithm, a ray is cast through the centroid, then all the boundary points are projected onto the ray, and the variance of the projected point distances from the centroid is noted. The ray that produces maximum variance is estimated to be the medial

axis. Once the medial axis is noted, the axial-length and lateral-breadth of the shape can be easily calculated.

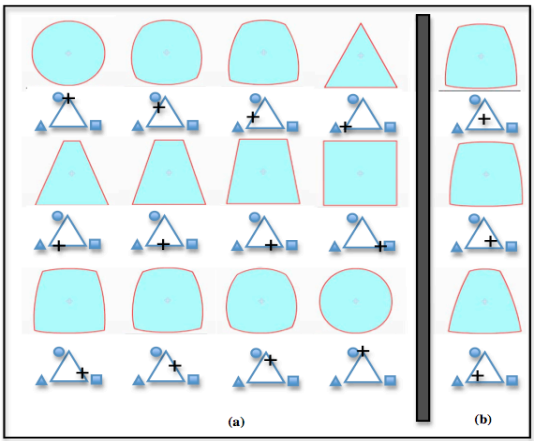

Figure 2. Blended shapes after consistent interpolation (shape weights indicated by cursor positions) [15]

We then perform a normalization affine transform to align the input shape to the Y-axis and scale it into a unit square. This simplifies shape error checking while ensuring rotation/translation/scale invariance during the fitting process. Lastly, we compute the best primitive shape combination, by minimizing boundary distance errors between our template shape combinations and the input points. In practice, this is a simple 2-level for-loop, incrementing shape weights by a fixed small value, and measuring the accumulated shape error. The shape error is calculated by accumulating slice-width errors over 40 lateral segments (along the medial axis). We have achieved decent fitting results for most cases. However, there are some cases where shapes computed with boundary distance errors do not match with human perception. We are currently working to improve the qualitative results through a perception regression model.

3) Space Parameterization

As shown in Fig. 3, we use a data structure {s,t} for parameterizing the cage and performing image warping, where t is a floating point number whose integral part holds the bezier segment number of the curve and s is the measurement of of distance along the line joining the center of a cage and the point on the bezier-segment-curve. Each pixel in Cartesian coordinates {x,y} can be easily converted into polar shape coordinates {s,t} and vice versa.

4) Image Warping

The challenge for articulated characters is that the limbs may be rotated significantly, and deformation along joints thus is a prime concern. We first explain the basic algorithm and then discuss how influences from multiple cages are resolved at overlapping and boundary regions.

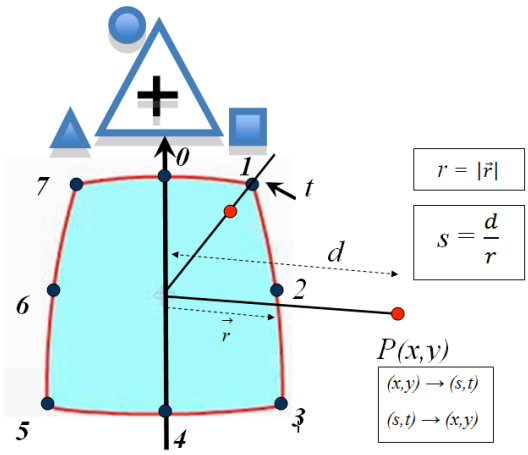

Figure 3. Polar Coordinate Parameterization of a Cage

The image based warp idea proceeds in a scan line manner, closely resembling texture fetches in the graphics pipeline. For each pixel p in the final image, an $\{s,t\}$ coordinate is first calculated with respect to the corresponding vector cage. The source pixel p_0 can be extracted by converting the same polar coordinates in the corresponding source cage to Cartesian coordinates. The color for the morphed pixel can then be fetched from this source pixel.

Since a morphed pixel can be expressed as a member of multiple cages (due to the overlap among body parts, i.e. neck, torso and head), pixel p in the final image might end up being sourced from a few different pixels in the source image (due to different $\{s,t\}$ coordinates in different cages). To resolve this issue, we perform distance-weighted influence blending of the different source pixel positions contributed by different cages. We avoid pixel color blending to prevent texture artifacts. The blending weights are derived as inversely proportional to pixel p's distance to the nearest boundary point on the associated cage, as shown in Eqn. 3, where d_i is the distance between p and the center of $cage_i$, and r_i is the corresponding center-boundary distance.

$$b_i = \frac{1}{d_i - r_i} \quad \text{.....(3)}$$

By doing this, we can effectively reduce undesirable stretching artifacts at boundaries between body parts and at joints undergoing large rotations. This simple weighting scheme does away with the need for manually specified deformation weights (e.g. in Linear Blend Skinning), and saves the artist a lot of extra manual work.

B. System Implementation

The system should be able to let artist annotate body parts of character images, and then to transform them into any hand-drawn scaffolds. Our system consists of three

 ISBN: 978-1-61208-110-6

main modules: a) Body part annotation; b) Scaffold sketching c) Image Retargeting. The interface has been designed for browsers to enable remote markup and drawings, with a server-side image-library.

1) Annotation

Fig. 3 shows the stroke annotation tool that can be used to trace out body parts. Each part is vector-fitted using least square error minimization after rotation and scale normalization, as described in Sec. A.2. The annotation process takes only a few minutes per image. We currently mark-up the following parts: head, neck, torso, upper arms, lower arms, hands, abdomen/hip, upper legs, lower legs and feet. We also construct a body silhouette to mark the whole body shape (blue outline in Fig. 4). The character image and its set of body cages are then stored into our image library, ready to be retargeted onto new drawings.

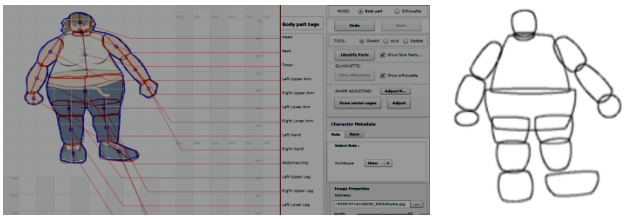

Figure 4. Manual Shape Annotation and Vector Fitting

2) Scaffold Sketching

Our system allows artist to design whole body scaffold that contains the same number of body parts. Based on the position of each part in relation with the others, it can support automatic deduction of body parts. However, the accuracy depends on the eccentricity of the proportions. The artist can easily correct wrongly annotated parts with a few mouse clicks in the drawing interface. A completed scaffold drawing is also automatically vector-fitted using the algorithm mentioned in Sec. A.2. Newly drawn scaffolds can be saved into a library for review and modification purposes. The whole process of scaffold design and image mark-up typically takes only a few minutes to complete.

3) Image Retargeting

Once the scaffold design is complete, the artist can then choose the character in the image library to retarget onto the scaffold. They can directly tweak source annotations to create deformations (as shown in Fig. 5), or they can retarget a selected (pre-annotated) image onto a newly drawn scaffold (as shown in Fig. 6). Fig. 7 illustrates a variety of scaffold retargeting results for two source character images. Results are obtained in real time, given the economical performance of the warping algorithm. We have achieved substantial acceleration for GPU implementation only for images above HD resolution.

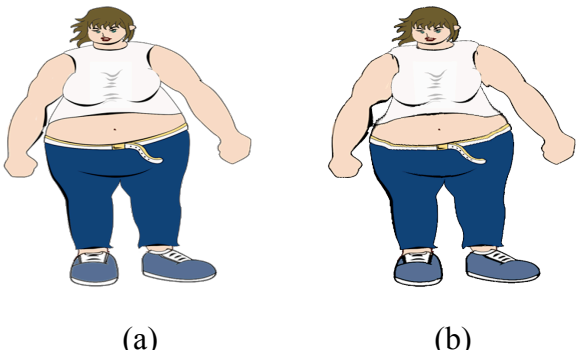

Figure 5. Interactive Deformation Mode: Original vector cage for torso in (a) made slimmer in (b) by tweaking the torso cage directly in vector space. (a) Shape(tri:36, sqr:12, cir:52); (b) Shape(tri:16, sqr:62, cir:22).

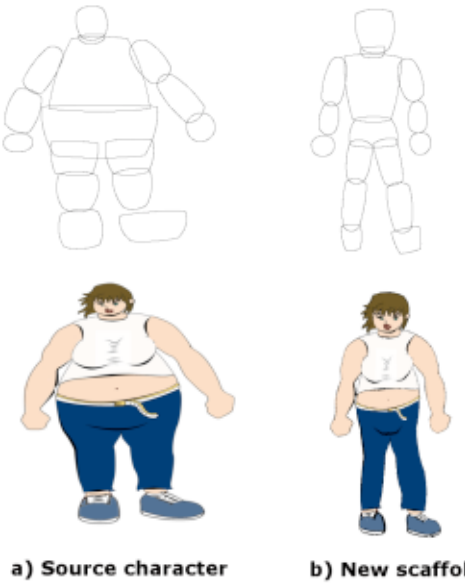

Figure 6. Warped output with a pair of source character image and new body scaffold. a) Source character image and its vector shape representation of body cages. b) Vector shape representation of artist's new scaffold and result produced by the system

As evident from the results, our mesh-less algorithm performs decently for retargeting between different shaped/proportioned characters. It can handle minor change in postures as well. We are currently working on improvements to the warping algorithm that allow drastic changes in postures, and also drawings from different views that will enable texture retargeting from 3D models.

IV. COMPARISON WITH FREE FORM DEFORMATION

We now cite several desirable properties from deformation survey papers [1,6], and then do a head-to-head comparison with a well known deformation method.

 ISBN: 978-1-61208-110-6

Users expect deformation operations to preserve the shape of the object locally. They also wish that the deformation could be produced with intuitive operations, and with a minimal number of control handles. The resulting deformation should be smooth (differentiable) and predictable (can arrive at the same result for the given configuration, irrespective of the previous state). The algorithm should be efficient allowing real-time interactivity for at least a low-resolution model. Lastly, it is desirable to have shape operators to aide deformation tools for personality-driven, stylized character design.

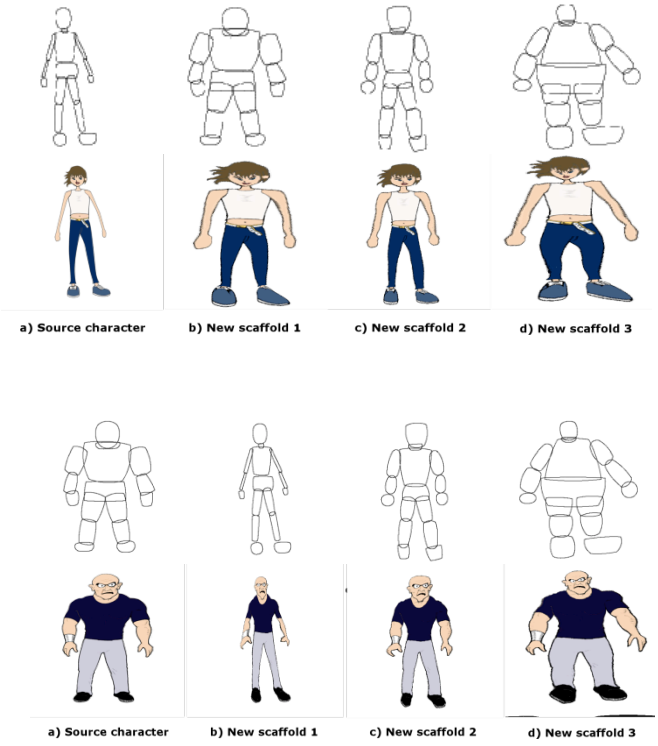

Figure 7. More examples of retargeting characters to draft scaffolds. a) Source image and its vector shape representation. b) c) and d) Result image with new scaffold 1, 2 and 3 from the artist.

In order to compare our results with FFD lattice deformers, we have implemented a shape-driven setup of FFD cage control points using Maya™. Since FFDs are setup as regular XY grids for 2D deformation, we can regularly sample our vector cage in *s*-*t* coordinate space and place the control points in corresponding Cartesian coordinate locations. For example, all boundary FFD points must lie on locations where *s=1* for a given cage. One problem with regular FFDs is that the base deformer points must follow a regular grid pattern, causing accuracy problems when morphing from non-rectangular body shapes. An alternative would be to use more flexible representations like EFFD [1], but these come at the cost of multiple iterative computations.

Our warp algorithm gave better results than the FFD deformers, especially at cage boundaries. For example, in Fig. 8, the torso and upper arms are drawn unnaturally towards each other at the joints in the FFD case, for a triangular shape morph. Our warp algorithm better preserves the original area, as every pixel transformation is a weighted influence between all shape fields (as opposed to isolated FFD deformation with distance fall-off influences). This property is also evident in the Fig. 5b and Fig 7 (second row, last column), where abrupt discontinuities in the underlying cages are smoothly reflected in the warped image.

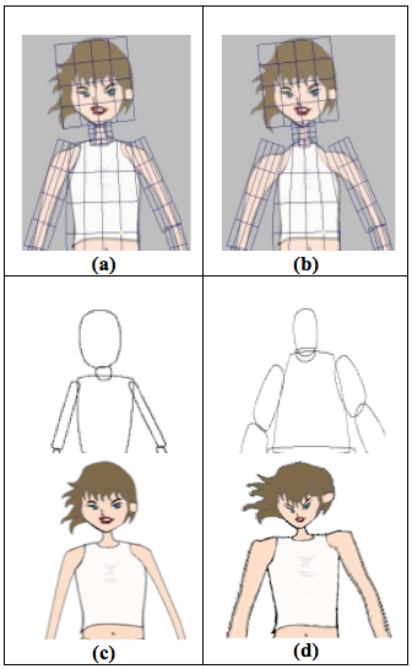

Figure 8. Comparison of FFD with our Primitive Cage Field deformation. a) Source image before applying FFD. b) Result image after apply FFD with artifact among torso and upper arms. c) Source image and its vector shape representation. d) Result image and its vector shape representation of artist's new scaffold with smooth connection among torso and upper arms.

V. CONCLUSION AND FUTURE WORK

We have successfully demonstrated a system that allows artists to retarget a pre-annotated character image onto the draft scaffold. This allows them to quickly create an estimate of the desired character without spending effort in mapping detail templates. We have demonstrated decent quality results for a variety of scaffold proportions and shapes for three different humanoid characters. We have also shown how our warping algorithm yields more stable results than an equivalent FFD setup. Our method does not require any expensive mesh calculations, and is free from fold-over or hole artifacts, typical of mesh deformation methods.

We are currently working on a number of improvements: support for multi-stroked outlines, auto-segmentation of

 ISBN: 978-1-61208-110-6

body part cages from a full-body outline drawing, and better warp robustness to large posture and view changes. We hope that these contributions will open up paths for more intuitive tools and easier character design.

ACKNOWLEDGMENT

This research is funded by MDA GAMBIT fund (WBS: R-252-000-357-490), sponsored by Media Development Authority of Singapore.

REFERENCES

[1] A. Angelidis , K. Singh, "Space deformations and their application to shape modeling," ACM SIGGRAPH 2006 Courses, Boston, 2006.

[2] N. Beiman, "Prepare to Board! : Creating Story and Characters for Animated feature," Focal Press, 2007.

[3] M. Botsch, M. Pauly, M. Wicke, and M. H. Gross, Adaptive space deformations based on rigid cells. Computer Graphics Forum 26, 3, 339–347, 2007.

[4] S. Camara, "All About Techniques in Drawing for Animation Production," 1st ed. 2006, Barron's Education Series, Inc.

[5] S. Coquillart, "Extended free-form deformation: A sculpturing tool for 3D geometric modeling," Comput. Graph. 24, 4, 187–196.

[6] J. Gain and D. Bechmann, "A survey of spatial deformation from a user-centered perspective," ACM Transactions on Graphics (TOG), v.27 n.4, p.1-21, October 2008.

[7] L. Garrett, "Visual design : A Problem-Solving Approach," Huntington, N.Y., R. E. Krieger Pub. Co., 1975.

[8] B. Gooch, E. Reinhard and A. Gooch, "Human facial illustrations: Creation and psychophysical evaluation," ACM Trans. Graph.23, 1 (2004), 27–44.

[9] J. Griessmair and W. Purgathofer, Deformation of solids with trivariate B-splines, Eurographics 89, 137–148.

[10] J. Hamm, "Cartooning the Head & Figure".

[11] C. Hart, "Cartoon cool : How to Draw New Retro-Style Characters," Watson-Guptill, 2005.

[12] C. Hart, "Simplified Anatomy for the Comic Book Artist : How to Draw the New Streamlined Look of Action-adventure Comics," Watson-Guptill, 2007.

[13] L. N. K. Hai, W. Y. Peng and G. Ashraf, "Shape Stylized Face Caricatures," unpublished.

[14] T. Igarashi, T. Moscovich, and J. F. Hughes, "As-rigid-aspossible shape manipulation," ACM Trans. Graphics 24(3), 1134–1141 (2005).

[15] M. T. Islam, K. M. Nahiduzzaman, Y. P. Why and G. Ashraf, "Learning from Humanoid Cartoon Designs," *Advances in Data Mining, Applications and Theoretical Aspects (LNCS Springer),* 6171:606-616, Berlin, 2010..

[16] J. Liu, Y. Chen and W. Gao, "Mapping Learning in Eigenspace for Harmonious Caricature Generation," ACM Multimedia 2006: 683-686

[17] S. Loncaric, "A survey of shape analysis techniques," Pattern Recognition 31 (1998), pp. 983–1001.

[18] R. MacCracken and K. Joy, "Free-form deformations with lattices of arbitrary topology," In: SIGGRAPH 96 Conference Proceedings, pp. 181–188 (1996).

[19] M. D. Mattesi, "Force : Dynamic Life Drawing for Animators".

[20] L. Moccozet and N. M. Thalmann, Dirichlet free-form deformations and their application to hand simulation. Computer Animation, 93–102, 1997.

[21] J. A. Mugnaini, "Drawing: A Search for Form".

[22] A. Orzan, A. Bousseau, H. Winnemöller, P. Barla, J. Thollot, and D. Salesin, "Diffusion Curves: A Vector Representation for Smooth-Shaded Images," ACM Transactions on Graphics (Proceedings of SIGGRAPH 2008), Volume 27 – 2008.

[23] T. Pavlidis, "A review of algorithms for shape analysis," Comput. Graphics Image Process. 7 (1978) (2), pp. 243–258.

[24] A. R. Rivers, and D. L. James, FastLSM: Fast lattice shape matching for robust real-time deformation. ACM Transactions on Graphics 26, 3, 82, 2007.

[25] R. Schmidt, T. Isenberg, P. Jepp, K. Singh, and B. Wyvill, "Sketching, Scaffolding, and Inking: A Visual History for Interactive," 3D Modeling, 2007.

[26] R. Schmidt, B. Wyvill, M.C. Sousa, and J.A. Jorge, "ShapeShop: Sketch-Based Solid Modeling with BlobTrees," 2nd Eurographics Workshop on Sketch-Based Interfaces and Modeling, pp. 53-62, 2005.

[27] T. W. Sederberg and S. R. Parry, "Free-form deformation of solid geometric models," Comput. Graph. 20, 4, 151–160.

[28] J. Serra, Image Analysis and Mathematical Morphology, Academic, New York (1982).

[29] A. Sheffer and V. Kraevoy, "Pyramid coordinates for morphing and deformation," In: Proceedings of 3DPVT (2004).

[30] K. Singh, E. Kokkevis, Skinning Characters using Surface Oriented Free-Form Deformations. Graphics Interface 2000: 35-42. 2000.

[31] O. Sorkine, and M. ALEXA, As-rigid-as-possible surface modeling. In Proceedings of Eurographics/ACM SIGGRAPH Symposium on Geometry Processing, 109–116. 2007.

[32] R. W. Sumner, J. Schmid, and M. Pauly, Embedded deformation for shape manipulation. ACM Transactions on Graphics 26, 3, 80. 2007.

[33] D. Sýkora, J. Dingliana and S Collins, As-rigid-as-possible image registration for hand-drawn cartoon animations. NPAR 2009: 25-33. 2009.

[34] J. Tao, Z. Qian-Yi, P. Michiel, D. Cohen-Or, U. Neumann, "Reusable Skinning Templates Using Cage-based Deformations," Dec 2008, Proceedings of ACM SIGGRAPH Asia 2008.

[35] M. Thorne and D. Burke, "Motion Doodles: An Interface for Sketching Character," Motion University of British Columbia, 2004.

[36] R. Wang, "Image Understanding," China. 1995.

[37] Y. Wang, K. Xu, Y. Xiong and Z.-Q. Cheng, 2D shape deformation based on rigid square matching. Computer Animation and Virtual Worlds 19, 3–4, 411–420, 2008.

[38] Y. Weng, W. Xu, Y. Wu, K. Zhou, and B. Guo, "2D shape deformation using nonlinear least squares optimization," The Visual Computer 22(9), 653–660, 2006.

 ISBN: 978-1-61208-110-6

Non-linear Video

A cross-platform interactive video experience

Robert Seeliger
Future Applications and Media
Fraunhofer Institute FOKUS
Berlin, Germany
robert.seeliger@fokus.fraunhofer.de

Christian Räck
Future Applications and Media
Fraunhofer Institute FOKUS
Berlin, Germany
christian.raeck@fokus.fraunhofer.de

Dr. Stefan Arbanowski
Future Applications and Media
Fraunhofer Institute FOKUS
Berlin, Germany
stefan.arbanowski@fokus.fraunhofer.de

***Abstract* - Non-linear video is an approach, which makes video content an interactive experience. Non-linear video gives the viewer the opportunity to interact with objects that are part of the video and access supplemental information. On demand, multimedia content is linked with related information. Interactive, time independent navigation opens a new ways to experience video content. This paper shows how such an system could be built upon IPTV and web technology in a cross platform manner.**

Interactive video; non-linear content; user interaction; object description; personalization; IPTV

I. INTRODUCTION

Internet Protocol (IP) based media services, such as Internet Protocol Television (IPTV) [1], are different from conventional TV technology. Via IPTV, television content will be viewed and delivered through technologies used for computer networks. This opens to a wide range of new media services and asks for innovative advertisements and content patterns to satisfy the ever increasing demand of the advertisement industry as well as content production for predefined interfaces to place product and content related information and advertisements in close vicinity to particular objects within the video.

Advertisements in the form of pre-rolls, post-rolls, and overlays are becoming increasingly ineffective as more and more consumers decide to skip these commercials due to the unimportance of the offered information and low individual involvement and interest in the offered product. This fact is boosted by the technological and economic evolutions in digital Television (TV) and media business areas. There is a strong need for new marketing strategies, innovative advertisements and finally a need for new kinds of interactive video content to fit these requirements [2].

Non-linear video concerns the delivery of personalized interactive value added information and related videos to end-users. Such an interactive item could be a video clip that can be paused at any time. Thus the introduced technology enables customers to decide when and which related information, interactive items and advertisements are displayed. In those environments customers can pause each video content at any time.

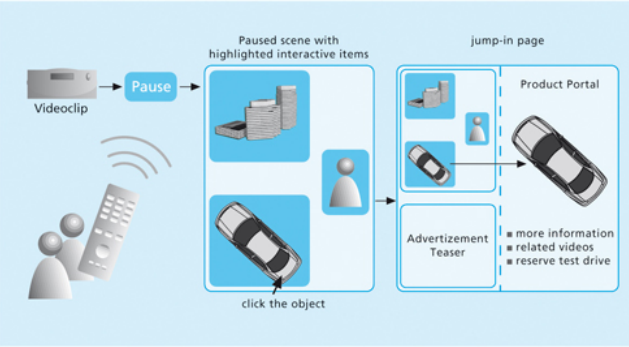

Figure 1. Non-linear video content interaction

Once paused, several objects in that current scene are automatically highlighted. Each highlighted object displays additional information, such as a detailed description and links to other objects or related content, when clicked. The displayed objects, descriptions and links are personalized based on previously learnt end-user profiles. The Non-linear video system identifies which objects are relevant to a particular customer based on his or her current situation and then only displays these objects for a personalized and interactive experience. Figure 1 shows, how different elements, such as a piece of content, a description of the object shown there, and content recommendations are linked to create a new interactive content experience.

II. OVERVIEW

Non-linear video offers the platform for realizing interactive and personalized multimedia content. Videos are linked to related information thus making time independent navigation possible. The linear character of traditional moving pictures and video formats are enhanced with non-linear video towards multiple ways of interaction. The user can navigate at any time through objects, which are contained in the video content such as TV programs and films. As soon as the viewer clicks on an object that interests them, supplemental interactive information or video is displayed. This can be any multimedia content (image, text, animation, video, pdf), websites, or alternative communication methods such as telephone, chat, email as

 ISBN: 978-1-61208-110-6

well as Web 2.0, community and social media services, which can be shown on a time basis or in subject to the user. Non-linear video supplements the previous available technology of pre and post rolls, commercial breaks, product placement, and overlays. Videos become clickable using links to content and additional information which are brought up parallel to the existing moving pictures. Metadata describe the objects in the video as well as possible ways to interact with it.

III. RELATED WORK

As the history of the last ten years and recent developments on interactive multimedia content and related TV or Web-based solutions have descriptive shown, interactive video content pines for easer solutions. Whether previous systems [3] seem to provide all technical issues and apparently fulfill all requirements for those interactive multimedia experiences, in fact the bulk of them disappear due to lack of usability. This includes both, the user side to experience value added services and real interactivity, and the service providers modality to serve such interactive content offers. Those solutions were mostly based on large and complex metadata descriptions as MPEG7 [4], related interactive TV technologies as MHP [5] or more theoretical approaches on top of MPEG4 [6] video scene and object descriptions. Unfortunately these approaches do not affect any products and services on the media market till now. In contrast to the depicted history of interactive video, the envisaged Non-linear video solution tackles a more applicable and practicable route to enable interactive content based on a media platform to serve videos, and the related metadata to provide object descriptions and the associated information as well.

The realization of the above-mentioned vision in section I require the exploitation of a number of ideas, which have been developed in different technology domains. Ideas taken from interactive and non-linear video utilization build the basis for the definition of the new interactive content features as described above. Results taken from previous projects and developments, which dealt with the realization of a recommendation system [7], are used to provide the required personalization features. A central component of the described platform is the Interactive Video Player that utilizes results from different technology domains, such as recommendations, next-generation video platforms [8] as well as standardized IPTV infrastructures [9] to provide the content including the interactive items to end-users.

IV. USE CASE

Following, a usage scenario for interactive content-based on the Non-linear video technology will be introduced. The 'Berlin Tourism' usage scenario can be categorized as a scenario for an interactive tourism information system. In this scenario a local tourism centre wants to promote their special offer for a weekend trip to Berlin. The campaign includes a series of short video clips about the most famous sights of Berlin, which are delivered through our envisioned

Figure 2. Browser based player for interactive content

new interactive non-linear video platform. The tourism centre links detailed tourist information to the sights and identifies a number of additional objects in the videos. Affiliates of the tourism centre are allowed to place their own information and related content on the previously identified objects.

Affiliates of the tourism centre select the content that is provided by the tourism centre to feature their own products and services on the tourism centre website. The *traveller* is an end-user, who visits the tourism centre website looking for a place to go on his vacation. He will take advantage of the interactive service in terms of being able to plan his trip to Berlin and get informed in detail by watching and navigating through the interactive videos using the non-linear video technology on his mobile phone, Laptop or TV at home.

V. HOW CONTENT BECOMES INTERACTIVE

Traditionally, videos and TV content have been created to be consumed passively. Using non-linear video, TV and multimedia content is made to be experienced interactively. This technology creates a seamless transition between additional information, valued added services and video content. The resulting interactive content functions just like a website. Individual sections can be annotated and linked to continuous content such as text, images, video and links. This information is represented by XML based metadata and is made available using an interactive video player. Web-Standards are employed as the access control technology. The FOKUS Tagging-Tool enables the annotation of the raw video data. This intelligent software supports the editor by identifying relevant objects and sections of a video-scene, placing information and by highlighting. The tagging tool is designed to be used as a web-based solution and can be conveniently utilized within the browser. In addition to the time and spatial data, you can also add keywords to describe objects such as the type, category or kind of interaction.

VI. CROSS PLATFORM INTERACTIVE VIDEO UTILIZATION

Non-linear video works with many of the user devices available today. The technology of Fraunhofer FOKUS enables the convergent use of interactive video content on TV, the web as well as mobile phone. This unique

 ISBN: 978-1-61208-110-6

environment makes interactive media available regardless of

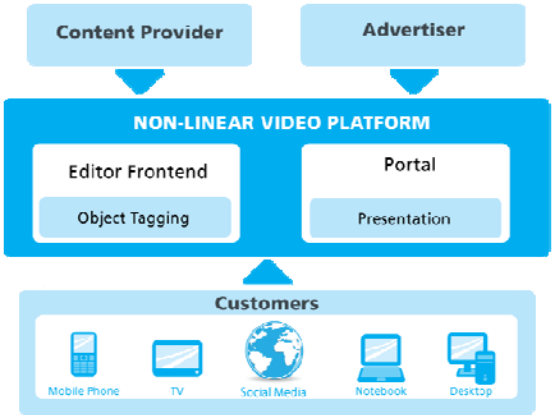

Figure 3. Cross platform approach

the end device and access platform.
The different ways for the viewer to interact, the graphic presentation of the supplemental information as well as the content itself all come in the best possible format for any type of user device. Non-linear video thus offers diverse variations of both the visual aspects of and interaction with the content. As a result, it can offer interactive content, relevant additional information and communication channels custom-made for both the user and the device.

VII. ARCHITECTURE

This section provides a brief overview on the overall architecture of the envisioned interactive media platform and its downstream marketplace. The envisioned architecture is a three-tier architecture that consists of a rich content "media player component" and a video tagging and annotation toolkit both in the presentation tier, a set of backend components in the application tier and one or more servers in the data tier.

A. *Building blocks*

The high-level architecture of the platform consists of six different building blocks including sub modules:

- Interactive Media Player (client side)
- Object Tracking and Linking Tool (client-side)
- Media Server (server-side)
- Recommender Server (server-side)
- Marketplace (server-side)
- Advertising Server (server-side

Fig.4 shows the high-level architecture and the above-mentioned building blocks and their functionality.

B. *Workflow description*

The depicted technology for user-initiated content interaction based on our Non-linear video technology refers to the following set of features:

- The raw video is enriched by metadata.
- Users can initiate a content interaction session at any time.
- Prepared objects will be highlighted for interaction.
- Consumers can select several objects to get further information (dive into the content).

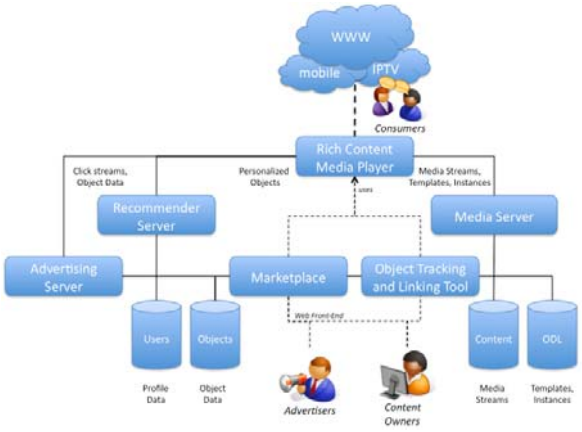

Figure 4. Overall system architecture

The additional information, which is displayed when a highlighted object is selected, provides an added value to the user. In this way, he is capable of getting a description of the highlighted object as well as links to related content of any type. Navigating the hyperlinked content spawns a hierarchical object tree, which represents the user's particular interests at a given moment in time. In contrast to traditional content, which has been produced for linear and non-interactive TV, the user is free to choose what content he wants to consume now and what will be next – the user goes interactive. In a first step, the raw video-content has to be analyzed and annotated. This results in an identification and description of the objects in the scenes (e.g., a car, watch, jacket, sight). The outcome of this procedure is metadata information describing the video content. Advertisers and content producers can place product related information and interactivity by annotating the predefined objects. After completion of the annotation and aggregation step, the raw video content is enriched with interactive items for user interaction. If the user initiates a pause while watching the video, the rich media player highlights all objects within the current scene, which have been pre-annotated for user interaction. The user is now skilled to interact directly with the interactive items. Based on the given metadata information, various possibilities are conceivable, e.g., link to product portals, direct shopping, product related add-on information and nonlinear video scene navigation.

 ISBN: 978-1-61208-110-6

VIII. REFERENCE IMPLEMENTATION

A. Interactive Media Player

The Interactive Video Player component displays interactive videos, which are delivered in the format that has been defined as Non-linear video. The media player highlights all objects, which have been previously identified and which are described in the corresponding metadata. When an end-user clicks on an object a new user interaction session is started and all actions that are linked to the selected object are triggered. To ensure a high attention rate among the target audience objects can be filtered or differently coloured based on their predicted relevance. For that, a recommender system is used to calculate the relevance of all objects in this scene for a particular end-user. Fig. 5 shows the implementation of these interactive media player component in terms of our mobile solution running on Apples iPhone. Other implementations for Web-Browser and Hybrid TV sets are also available. The system provides multiple interaction layers, which enables the user to leave the current video by watching the next clip about an object in a hierarchical way, he can navigate throw the media in a non-linear manner.

Figure 5. Non-linear video implementation for smartphones

B. Object identification, tracking and linkage tool

The Object Tracking and Linking Tool is used to identify all objects in a video. This is a three-step process: At first an object template has to be defined for each object that should be tracked in a video. Next a new instance has to be created for each scene that contains this particular object. At last the object has to be tracked in each scene. The object template contains the definition of the tracking shape, a link to an HTML page that provides further information about the object and a list of related Web links. Each object instance contains information about a particular scene in which the object is visible. Single frames numbers and the position of the tracking shape in the video frames are recorded.

The output of the tracking and linking process is the object template definition and the list of object instances for all objects in a video. This metadata is described in an XML-based *Object Definition Language (ODL)*, which has been adopted on the experience of XML based metadata as TV Anytime [10] [11]. The information, which is described in the ODL, is used by the Interactive Video Player to enable the end-user interaction as described in the previous paragraph.

In the following example, implementations of metadata separate from the video data are given, i.e., metadata and video data exist as separate files and database entries which are combined in the video data player. This has the advantage that the video data can remain unchanged. The video data player has to load the video as shown in "example.obj" in the sample code of the ODL given below. Further, the video player has to load the description of the identified objects in the video data and the respective linking.

C. Object Definition Language

The following sample XML-code describes the object identification. For each object that a user can interact with, a

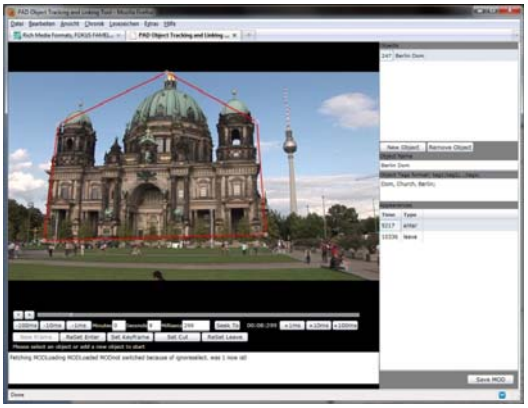

Figure 6. Web-based user interface of tagging toolkit

set of bounding boxes is defined. Each bounding box is

```
<?xml version="1.0" encoding="utf-16"?>
<MediaObjectDescription xmlns:xsi="http://www.w3.org/2001/XMLSchema-
instance" xmlns:xsd="http://www.w3.org/2001/XMLSchema">
 <ObjectAppearances>
  <ObjectFrame>
   <ShapeTimeFrames>
    <ShapeInTime>
     <TimeFrame>3769</TimeFrame>
     <Points>
      <RelPoint>
       <X>0.48125</X>
       <Y>0.26379310344827589</Y>
      </RelPoint>
     </Points>
   </ShapeTimeFrames>
   <Object>
    <ObjectID>94</ObjectID>
    <Name>Berlin Cathedral</Name>
    <Tags>Berlin;Point of Interest;Cathedral;</Tags>
    <PredictedRelevance>0</PredictedRelevance>
    <LinkedAdvertisement>
     <AdvertisementID>38</AdvertisementID>
     <Name>Berlin Cathedral</Name>
<Text>The Berlin Cathedral…</Text>
<PictureURI>http://sampleURL/samplepic.jpg</PictureURI>
      <URI>http://samplevideo.com;feature=related</URI>
    </LinkedAdvertisement>
   </Object>
  </ObjectFrame>
 </ObjectAppearances>
 <MediaID>15</MediaID>
 <MediaURI>http://sampleuri/samplevideo</MediaURI>
 <Owner>
  <CoID>5</CoID>
  <Name>tester</Name>
  <CoTags />
 </Owner>
</MediaObjectDescription>
```

 ISBN: 978-1-61208-110-6

specified with its type (e.g., “a rectangular box”) and size. Then the set of frames where a particular object is shown is identified. Last, but not least transitions of the bounding boxes between individual frames are defined.

IX. CONCLUSION

As opposed to traditional video, which limits interaction and use due to its linear nature, interactive video opens up new and diverse possibilities. FOKUS non-linear video supports multiple levels of interaction in the content itself as well as between the interactive objects and the supplemental information available. The viewer can access interactive video content using three basic levels of interaction:

- Moving picture content (video material)
- Interactive objects in video (text, audio, video, image, web)
- Communication channels (telephone, chat, email, web 2.0, social networking)

Whereas the level of moving picture content sparks someone’s interest in a topic or product, the newly created levels enable customization of content and make detailed information about the objects in the video available. In combination the interactive possibilities, that are carried out to produce a non-linear story line that, can be freely chosen and interacted with by the viewer.

The linking of information using non-linear video can be applied ideally for multistep information access and business models: from free use, which grants access to general information to registration sites as well as premium content access through a pay system. Since interaction in today’s video platforms is limited to commenting on posts and placing advertising banners, the non-linear video technology offers new possibilities for interactive video use. Multimedia data, video objects and additional information are linked to interactive content and enriched by customized object interaction. Related information, communication channels and content are all customized for the user and his end device. Supplemental information enables advanced applications such as interactive advertising, video portals and edutainment.

Fostered by the innovative technology and simultaneously upcoming opportunities, manufacturers and CE-Industry will be empowered to open up new markets. Interactive media services become reality and will boost the user experience to a higher level. The current solution provides interactive video content via rich internet application technology as MS Silverlight [12]. The integration to our ETSI TISPAN [12] and Open IPTV Forum standards compliant Open IPTV Ecosystem [14] has already been done. It shows how interactive content may be used in future IP based TV, and Hybrid TV environments [15]. Next steps will introduce more differentiated interaction models for identified interactive object items within the videos. This will imply enhancements on the object description language, the tagging toolkit and the associated interactive media players for the depicted various target platforms.

References

[1] O. Friedrich, R. Seeliger, and A. Al-Hezmi, „Prototyping interactive and personalized IPTV-services on top of open IMS infrastructures” 6. EuroiTV2008, Salzburg Austria

[2] T. Haaker, N. Ito, S. Smit, J. Vester, M.G lbahar, K. Shepherd, and J. Zoric, Business Models for Networked Media Services, 7th European Interactive TV Conference (EuroITV 2009), Leuven, Belgium, June 2009

[3] Hang Nguyen, Julien Royer, Durgesh O Mishra, Olivier Martinot, and Françoise Preteux, “Atomatic generation of explicitly embedded advertisement for interactive TV: concept and system architecture”

[4] MPEG-7 Overview (v.10), ISO/IEC JTC1/SC29/WG11 N6828 Palma de Mallorca, October 2004

[5] ETSI ES 201 812 V1.1.2 (2006-08) Digital Video Broadcasting (DVB);Multimedia Home Platform (MHP) Specification 1.0.3

[6] MPEG-4 BIFS tutorial http://gpac.sourceforge.net/tutorial/bifs_intro.htm

[7] Christian Räck, Stefan Arbanowski, and Stephan Steglich, “A Generic Multipurpose Recommender System for Contextual Recommendations”, 8th Intern. Symposium on Autonomous Decentralized Systems (ISADS 2007), Sedona, Arizona, March 2007

[8] Draft ETSI RTS 182 027 V3.1.0 (2008-11). Telecommunications and Internet converged Services and Protocols for Advanced Networking (TISPAN); IPTV Architecture: IPTV functions supported by the IMS Subsystem

[9] Open IPTV Forum Release 1 Specification: Volume1 Overview V1.0 available at http://www.oipf.tv

[10] TV-Anytime Forum, http://www.tv-anytime.org/, 2010-06-10

[11] ETSI TV Anytime XML Schema: ETSI TS 102 822

[12] MS Silverlight Rich Internet Application Technology http://silverlight.net/

[13] ETSI ES 282 001: "Telecommunications and Internet converged Services and Protocols for Advanced Networking (TISPAN); NGN Functional Architecture Release 2

[14] Fraunhofer FOKUS Open IPTV Ecosystem, http://www.fokus.fraunhofer.de/en/fame/_pdf/FOKUS-IPTV-Ecosystem.pdf, 2010-06-09

[15] Philips Net TV, press information http://www.digitalnewsroom.philips.com/press/nettv-pr.doc, 2010-06-11

 ISBN: 978-1-61208-110-6

Constraints in Course Design Using Web 2.0 Tools: A Croatian Case

Nikola Vlahovic, Zeljka Pozgaj
Faculty of Economics and Business
University of Zagreb
Trg J. F. Kennedyja 6, Zagreb, Croatia
{nvlahovic, zpozgaj}@efzg.hr

***Abstract*— Even though Internet and Web have always supported some form of social interaction, Web 2.0 shifts this paradigm to a new level. As social networking that rises from Web 2.0 applications, gains acceptance within the Internet community and the general public, a concept of enabling e-learning using Web 2.0 tools and services becomes more and more recognized. Goal of this paper is to evaluate empirically some of the constraints that are crucial for course design using social networking tools. A survey was carried out among current students at the Faculty of Economics and Business in Zagreb in order to evaluate issues connected with student constraints on course design. A cluster analysis was used in order to identify different student groups and their expectations from Web 2.0 tools that would be used to support e-learning. The results show that there are four typical groups of student attitudes towards implementation of Web 2.0 in e-learning. A successful course design should take into account expectations and demands for each of these groups. According to these results comparison of identified needs and currently available practices was conducted showing the advantages Web 2.0 contributes to e-learning while also uncovering weak points that can be further improved.**

Keywords- e-learning, Web 2.0, social networking, Course design, cluster analysis, teaching

I. INTRODUCTION

Development of the Internet has influenced the way people communicate, work and learn, enabling new means and possibilities in distance learning. Distance learning implemented through the use of Internet services has evolved into e-learning paradigm. The possibilities of synchronous communication was introduced into learning processes, while the quality of asynchronous communication was improved using Internet services such as email, chat, forums and newsgroups. Also videoconferencing and teleconferencing made distance learning more accessible and cheaper for both the institutions but also for students. These services also allowed for new communication possibilities and new types of information exchange. Even though e-learning is seen as a tremendous enhancement of distance learning, there is still only passive approach to acquiring and using educational materials [1].

Web 2.0 paradigm has introduced further change in the way e-learning can be implemented. Web 2.0 is a platform that enables interaction, collaboration and information exchange between various users resulting in participative creation of rich new content [2]. In terms of e-learning interaction between teachers and students is improved by emphasizing the role of students. Also interaction of students with each other is introduced into the learning process as a new significant factor that improves the results of learning.

The goal of this paper is to examine the possibilities of employing Web 2.0 services and social networking services as an additional platform for design and creation of e-learning course content. Based on the current attitudes of students a number of constraints over course design can be identified in order to customize this learning platform to suite needs of students. In order to identify these requirements a survey was carried out among the students at the Faculty of Economics and Business. The results were used to estimate further steps in developing current e-learning practices and tools used, and also to develop an approach to the introduction of Web 2.0 as a platform for e-learning 2.0.

The rest of the paper is organized as follows: In Section 2, e-learning and Web 2.0 are defined and explained. In Section 3, issues in course design are presented as seen by the students, teacher and institutions. In Section 4, research methodology is described along with the questionnaire and basic statistics of the targeted student sample. Also, cluster analysis which was used for the analysis of student groups is described. Results are discussed in the following Section 5, along with a few guidelines for the introduction of the Web 2.0 platform. Finally, Section 6 contains conclusions and final remarks.

II. E-LEANING AND WEB 2.0

A. Development of e-learning

The term e-learning pertains on a very complex and dynamic process regarding and connecting learning processes and developments in digitalisation and ICT. There is a multitude of definitions of e-learning in literature. Most of these definitions refer to certain aspects of e-learning such as simple adoption of electronic media [17, 18], or possibility of achieving distance learning [20, 21] and so on. Each definition describes subtle differences that are associated with the term itself, but different authors emphasise different aspects of it depending on the context. A more general definition of the term can be found in [19] where Tavangarin et al. define e-learning as "*all forms of electronically supported learning and teaching, which are procedural in character and aim to effect the construction of knowledge with reference to individual experience, practice*

 ISBN: 978-1-61208-110-6

and knowledge of the learner. Information and communication systems, whether networked or not, serve as specific media […] to implement the learning process". In other words e-learning is an umbrella concept, which comprises almost anything related to learning in combination with information and communication technology [4]. This definition includes a type of education where students work on their own at home and communicate with teachers and other students via e-mail, electronic forum, videoconferencing, chat rooms, bulletin boards … and other computer-based communication [3]. As Dichtanz [4] points out the time and space component of e-learning, the same definition recognizes e-learning as a collection of teaching and information packages in continuing education which is available at any time and any place and is delivered to learners electronically.

E-learning development is strongly associated with the organisation of the distance learning process. The process of this evolution from distance learning to e-learning in its present form was carried out in three different stages: (1) stage before digital era; (2) digital era of the Internet and World Wide Web; (3) Web 2.0 stage and E-learning 2.0.

At the very beginning of distance learning organisation, learning process was goal oriented, available to a particular audience, learning materials were in the form of print-outs, and the communication between students and teachers was based on traditional types of communication. This type of learning process was known as "correspondence study" or "correspondence education". In his way formal education was made available to working people who were able to complete courses with minimum time spent at the education institution. Next advancement in distance learning process organisation was the application of analogue technologies in the learning process (radio teaching, radio-television teaching). Even though this technology introduced additional benefits for overall learning process, learning still remained passive in its nature, limited to a particular audience and with limited communication possibilities. It allowed for introduction of learning in areas where learning process was undeliverable such as remote geographical locations (i.e., smaller islands of mainland, or scarcely populated mainland areas). More radical advancement of the learning process was introduced with the appearance of the Internet and World Wide Web. Information technology is still rapidly developing; teaching and learning materials are digitized, stored in databases. Due to the usage of modern ICT, collaboration of participants involved in learning process is highly facilitated. The progression of the Internet has set the ground for the rapid development of e-learning based on the Web [6]. E-learning was established as a new form of learning offering new possibilities. Some of these possibilities include: availability of a variety of learning materials in variety of presentation types, learning at one's own convenience – in terms of time and place of learning, unlimited possibility of repetitive learning (re-learning). Benefits from e-learning improve both distance learning implementations but, unlike earlier enhancements, they also significantly improve traditional learning processes. Implementation of e-learning is based on a learning management system (LMS) that is used to organize and deliver online courses [14]. Most important functions of an LMS include management of the course information, tacking of student progress and cataloguing of reusable learning objects. Learning process is enriched with simplified, user-friendly communication possibilities, sharing of information and opinions among students and teachers through e-mail, chat, instant messaging, file sharing, etc., but still it is a type of passive learning. Internet was a rich source of information, but it didn't allow its users to participate in the creation process, it didn't allow interactivity [7]. The most recent enhancement of the e-learning process is achieved through Web 2.0. Web 2.0 refers to a change in the way the Internet is used, representing its innovative collaborative nature. Flexibility, pervasive access, user-friendliness, interactivity, social interaction and collaboration and information sharing are just some of the advantages Web 2.0 brings to E-learning. All of these increase student motivation and foster student reflection [8] giving the students better control over their learning results.

B. Web 2.0 paradigme and its influence on social networking

Web 2.0 has made a shift in the way Internets users perceive and use Web content. Most of the developments of Internet and Web services have been technical in nature until Web 2.0. Most of the Web 1.0 content was published and edited by information owners and professionals. Even content that originated from the common users was first screened and approved or edited by web site editors or at least Internet service providers before it was made available online. Web 2.0 introduced a sociological change of the paradigm by excluding the middleman between Internet users. In this new paradigm users share their information directly. Internet service providers only provide the appropriate platform but do not interfere with the content that is published by the users for other users on a peer basis. This change enabled Internet users to become active users of the Web and realize potentials that were otherwise hardly achievable or ineffective, like collective intelligence, massive collaborative efforts, non-mainstream news content and niches, folksonomies, etc.

Possible disadvantage or threats Web 2.0 may lead up to is the creation of the digital narcissism and amateurism which can undermine expertise and safety of available information. Some of the critics already warn that the Web is filled with mistakes, half-truths and misunderstandings that make navigating and using Web difficult and exhaustive.

Nevertheless, applying Web 2.0 into the learning process can result in more positive change of the learning results than generating negative outcomes. For example, using online social networking service can be a good supplement to existing e-learning platform as it enables additional possibilities Web 2.0 generally provides. This is because all of the efforts of the students and teachers are readily publicly available to the whole learning group, which motivates students to be original. Being aware that all of the work they dedicate to mastering a course can be seen and appreciated by whole group, which can additionally stimulate students to

 ISBN: 978-1-61208-110-6

make their best effort. Along with the increase of student interaction this is one of the most important advantage e-learning 2.0 can provide.

III. COURSE DESIGN INCORPORATING WEB 2.0 TOOLS

Some of the most significant potentials of using Web 2.0 tools in learning are: (1) ability to provide anytime, anywhere learning, (2) give access to vast amounts of content, (3) increase students' opportunities to interact with other students, teachers and experts, (4) extend learning to the traditionally excluded, to the disabled and to the global community [9]. The actualization of these potentials, though, can be questionable since it largely depends on the properties of the course design. There is a number of issues during the course design using Web 2.0 services that need to be taken into consideration. We can divide these issues into three groups: issues for students, issues for teachers, and issues for Institutions.

Some of the most important issues for the students include (1) the inadequate online access, (2) the need to provide training for those not skilled in the use of a range of software used (3) tendency to become uncritical about the material they get from using Web 2.0 tools and (4) the blurring of the distinction between full-time and part-time study. In the first two cases course design can be adjusted so that the final implementation is not overdependent upon the ICT [10]. The problem of underdeveloped criticism and assessment of obtained materials is connected with the students' inability to reflect on their learning. This is a skill students should acquire during their earlier education as a permanent process that is developed continually throughout their study period. In order to support these processes course design should include additional tools to allow for formal reflection on lessons learned. In this way e-learning 2.0 can promote student self-awareness and self-criticism. Finally, in order to cancel the effect of blurring the distinction of full-time and part-time study course design should be able to incorporate flexible study patterns so that part-time study can be achieved as a lifelong learning opportunity [16].

Most important issues for the teachers are (1) the increase of the workload, (2) requirement of acquiring new technical skills, (3) prejudice towards e-learning 2.0 and (4) unclear intellectual property rights. In the first two issues teachers need to dedicate more of their time in order to establish a course and invest even some additional time into acquiring new skills to be able to design a course. In order to minimize additional workload the course design should be implemented so that is requires minimum of maintenance during the running of a course. In this way teachers can manage their time better and successfully balance between gaining new skills and maintaining content of the course. The last two issues can be resolved by informing the staff about the advantages of e-learning that can promote teachers' roles in the learning process and not demote them as they might fear. Intellectual property rights should be backed up by the educational institution and the LMS should allow for some mechanism of authentication of the users before they can download the learning materials [13].

The educational institutions should concentrate on resolving the following issues in order to facilitate the learning environment for both students and teachers: (1) establishing a customizable LMS of the institution [15], (2) support and encourage staff development and (3) define policies and practices in assessment processes [12].

IV. RESEARCH METHODOLOGY AND RESULTS

For the purpose of this research a survey was carried out among the students of the first year of undergraduate study in Business Economics at the University of Zagreb.

The goal of the survey was twofold. First goal was to investigate student attitude towards social networking services and their involvement and habits in using some of these services. Within this part of the survey students were asked weather they use some of the social networking services, and if they do why did they join and what do they use these services for. Additionally students were asked if they can see social networks as relevant tool for e-learning. Second part of the survey contained questions about their concerns about using social networks. They were asked to rank dangers from using social networks and more generally the Internet and in the same context to asses on average how much time they spend using Internet. Using results from the first part of the survey it is possible to deduce the advantages from using social networks in e-learning courses, while from results of the second part of the survey it is possible to assess threats and challenges in implementation of social networks as e-learning tool.

Collected data sample contains 184 answers to survey questions. Majority of students were female, while only 26% were male students, which reflects the structure of the whole population of students at the Faculty of Business and Economics (Fig. 1).

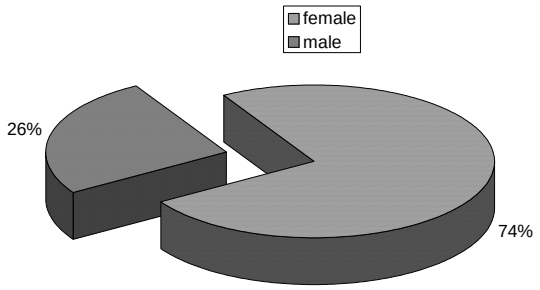

Figure 1. Surveyed students by sex.

When it comes to using particular Web 2.0 services Facebook is the most popular since 98% of students have an open profile. Other Web 2.0 services are barely represented since only 11% of students have an open profile with the second most popular service MySpace (Fig. 2).

 ISBN: 978-1-61208-110-6

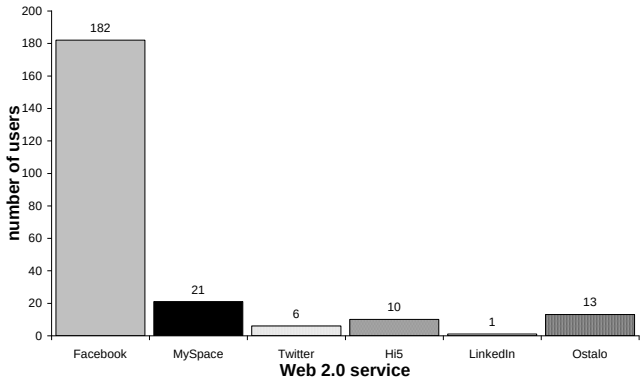

Figure 2. Number of students with open profiles of most common Web 2.0 services.

Finally, 61% of students believe that the usage of Web 2.0 services would be useful to them as an e-learning tool (Fig. 3).

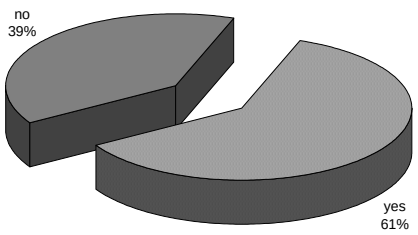

Figure 3. Number of students that believe Web 2.0 services can be used in e-learning successfully.

According to these data an in-depth analysis was carried out using undirected data mining methodology. The goal of the analysis was to identify typical behaviours of students in terms of their usage of Web 2.0 services and attitudes towards e-learning and the possible combination of the two. Cluster analysis was applied because it can provide basis for generating stereotypical properties that final implementation of course should contain in order to appeal and become useful to majority of the student population.

Cluster analysis, also called segmentation analysis or taxonomy analysis, is used to identify homogeneous subgroups of cases in data set. These groups, also known as clusters, are formed so that they both minimize within-group variation and maximize between-group variation. For each cluster a typical value across predictor variables is identified called centroid. Centroid is the value that has the minimal average distance for all members of each cluster, but that has maximum distance to centorids of other clusters [11]. By comparing centroids and statistics of different clusters differences can be determined between clusters and stereotypes can be created. For each stereotype a number of measures can be taken into account in order to customize final course design to targeted students.

The analysis showed that four distinctive student groups can be described. All of the student groups have some attitudes in common. ID theft is the biggest concern for all of the students, while disinformation on the Web or becoming addicted to Web is not perceived as a threat.

In Table 1, we can see typical values for some of the most important questions from the survey that make distinctive differences between groups.

First cluster consists of students that spend more time online where they participate in forum discussions (4-e) communicating with their "real world" friends (3-3), but also meeting new people (4-b). Due to intensive online communication these students see the main threat for online security in ID theft and possibility of false introductions from other Internet users. This group of students has good experience with the online world and they are willing to use Web. 2.0 services in education promptly. This is why this cluster can be called the cluster of "skilled active online users" which generally represent early adopters of technology.

Second cluster contains students that also spend more time online, but they do not actively participate in content creation in terms of posting on forums and discussion groups. They usually open profiles with online services under influence of their friends who have profiles already opened. They are more concerned with becoming addicted to Web then the other groups, but they do see a possibility of using Web 2.0 in the learning process but they need a bit more encouragement since they have a more passive approach. This is why we can call this group a cluster of

TABLE I. SUMMARY OF DETECTED CLUSTERS

	Sex	3-1*	3-3*	4-b*	4-e*	6*	8-1*	8-2*	8-3*	8-4*	8-5*	9 hour per week online	n	%
1	male	n	y	y	y	y	1	4	2	3	5	7hpw to 14hpw	42	23
2	female	y	n	n	n	y	1	5	4	2	3	7hpw to 14hpw	62	33
3	female	y	n	n	n	y	1	5	3	4	4	less than 7hwp	36	20
4	female	n	n	n	n	n	2	4	3	1	5	less than 7hpw	44	24

3-1 – Reason for opening a social network profile: friends' suggestion
3-3 – Reason for opening a social network profile: reconnect with old friends
4-b – Using social network for meeting new people
4-e – Using social network for discussions of leisure and private topics
6 – Do you think Web 2.0 can be used in learning process?
8-1 ID theft as a threat on soc. networks (1-high concern; 5-least concern)
8-2 false information on soc. networks (1-high concern; 5-least concern)
8-3 false representation of other users (1-high concern; 5-least concern)
8-4 lack of privacy on soc. networks (1-high concern; 5-least concern)
8-5 addiction to soc. network and Internet (1-high concern; 5-least concern)

 ISBN: 978-1-61208-110-6

"cautious passive online users" which can also be treated as trend followers in terms of technology adoption.

Third cluster can be viewed as a subgroup of the second cluster since students from this group also tend to use online services on recommendation of their friends, but are rather cautious when using interactive tools these services provide. These users on the other hand spend less time online and have much less interest of online communication and Internet. This is why they are the group least concerned with online security issues. This cluster can be also referred to as the cluster of "online beginners" or late adopters.

Finally, the fourth cluster contains students that are highly concerned with privacy on the Internet (8-4 in Table 1), and this is why they are highly sceptical and unwilling about using Web 2.0 services as e-learning tool (6 in Table 1). This is why this cluster can be referred to as the cluster of "sceptical non-users" or refusers.

V. DISCUSSION

As we have shown earlier there are four distinctive groups of students that have different habits and attitudes towards the Internet, but also the possibilities in employing Internet services in the learning process. These results correspond with the expectations based on the theory of diffusion of innovation. What is crucial though is that during course design considerations of specific needs that can encourage each of these groups has to be taken into the account. The choice of Web 2.0 service is important in order to motivate students for the innovative approach to the learning process. Service that is already most popular should be chosen, because there is a number of users that will actively engage in this new form of learning (first cluster), and there is a number of students that can be easily motivated to become users of this facility (second cluster). While students that are skilled active online users may have high expectations from advanced functionalities of the e-learning service, cautious passive online users may need more reassurance in the security properties of the service, even if the functionality of the web 2.0 service is more modest. These users will also benefit from well developed online help for Web 2.0 services, which is especially important for students "online beginners". Beginners will be more willing to participate if they can see additional benefits in comparison to traditional learning that can justify time invested in training and learning how to use a new tool. The most challenging student group is the group of non-users that lack not only the skill for using the Internet but also motivation. It seems that the most motivating property of the Web 2.0 services seems to be the possibility of collaborative learning. In this way each student within the learning group makes higher impact on other classmates than in traditional learning by publishing and making available all of the works, research and tasks they make for the course. Every student can see other students' work, which can be motivating since each opinion over a particular subject is expressed and made available. Also, the fact that all of the work is available online motivates students to be original and in great measure prevents cheating in terms of copying other students work. The possibilities of influencing their classmates should be seen as the most important factor for the students that are unwilling to participate in online learning activities.

Being able to satisfy all of these needs can pose a great challenge and additional effort for the teachers in the starting phase of the course design. This can greatly discourage teachers in employing Web 2.0 services in the learning process, even though after implementation the role of the teacher is drastically changed and less emphasised than in traditional learning process.

In order to alleviate this workload for teachers the institution should establish policies and practices that can motivate teachers effectively. These may include introduction of LMS or virtual learning environments, staff development possibilities and assessment frameworks.

VI. CONCLUSION

E-learning 2.0 is a new paradigm of distance learning facilitated through the use of Web 2.0 Internet services. The main characteristic of this approach is the shift of the roles of teacher and students within the learning process. Teachers' influence is diminishing while the interaction between students gains more importance on the realization of the learning process though the collaborative learning.

In this paper, we examined what are the requirements of students that can successfully achieve these goals. In order to identify student attitudes and their preferences and habits in using Internet services a survey was conducted among students at the University of Zagreb. A cluster analysis was employed on the gathered data. Four distinctive groups of typical student attitudes were identified. Groups that have most positive attitude to using Web 2.0 services in learning process are the "skilled active online users" and "cautious passive online users". Survey results strongly indicated that adequate course design must take into consideration the theory of diffusion of innovation, and cover particular needs of all distinctive groups of users. Groups of students which are less inclined to using Web 2.0 services as a platform for e-learning are "online beginners" and "sceptical non-users" that require special attention during the course design.

Currently available practices for supporting e-learning at the Faculty of Economics have provided a good starting platform for development of e-learning courses but additional effort is required to customize final courses for the identified students groups according to the survey results. Crucial role in achieving the goal of implementing e-learning 2.0 is the support of the institution towards the teaching staff but also in the implementation and maintenance of the adopted learning management system.

REFERENCES

[1] Z. Pozgaj and N. Vlahovic, The Impact of Web 2.0 on Informal Education, Proceeding of the 33rd international convention on information and communication technology, electronics and microelectronics - MIPRO, May 24. -28, 2010, Opatija, Croatia

[2] V. Bosilj Vukšić and M. Pejić Bach (eds.), Business Informatics, Element, Zagreb, 2009, p. 105 (in Croatian)

[3] Ž. Požgaj, Distance learning – reality or vision, Proceedings of 15th International Convention MIPRO 2002, Opatija, 2002, pp.19-24.

 ISBN: 978-1-61208-110-6

[4] T. Rekkedal and S. Quist-Eriksen, Internet Based E-learning, Pedagogy and Support Systems, http://learning.ericsson.net/socrates/doc/norway .doc [01/21/2006].

[5] Q. B. Chung, Sage on the Stage in the Digital Age: The Role of Online Lecture in Distance learning, The Electronic Journal of e-Learning, Volume 3, Issue 1, p. 1-4, http://www.ejel.org [12/18/2006]

[6] Ž. Požgaj and B. Knežević, E-learning: Survey on Students' Opinions, Proccedings of 29 th International Conference on Information Technology Interfaces ITI2007, Cavtat/Dubrovnik, 2007, pp. 381 – 386.

[7] P. Isias, P. Miranda, and S. Pifano, Dessigning E-leaarning 2.0 courses: recommendations and giudelines, Research, Reflections and Innovations in Integrating ICT in Education, pp. 1081-1085.

[8] H. Ajjan and R. Hartshorne, Investigating faculty decision to adopt Web 2.0 technologies: Theory and empirical tests. Internet and Higher Education 11, 2008, pp. 71-80.

[9] R. Mason and F. Rennie, E-Learning and Social Networking Handbook: Resources for Higher Education, Routledge, 2008.

[10] M. Kirkwood and L. Price, Learners and learning in the twenty-first century: What do we know about students' attitudes towards experiences of information and communication technologies that will help us design courses?, Studies in Higher Education, Vol 30, No, 3, 2005, pp. 257-274.

[11] N. Vlahović, Lj. Milanović, and R. Škrinjar, Turning Points in Business Process Orientation Maturity Model: An East European Survey; In: Mastorakis, N., editor. WSEAS Transactions on Business and Economics; Vol 7, No. 1, WSEAS Press, Athens, Greece. 2010, pp. 22 – 32.

[12]] J. Ireland, H. M. Correia, and T. M. Griffin, Developing quality in e-learning: a framework in three parts, Qualitiy Assurence in Education, Vol. 17, No. 3, Emerald Publishing, 2009, pp. 250-263.

[13] C. S. Chai, H. L. Woo, and Q. Wang, Designing Web 2.0 based constructivist-oriented e-learning uints, Campus-Wide Information Systems, Vol. 27, No. 2, Emerald, 2010, pp. 68-78.

[14] S. Downes, E-Learning 2.0, ACM eLearn Magazine, 2005, http://www. elearnmag.org/subpage.cfm?section=articles&article=29-1 [04/15/2010]

[15] D. Jennings, Virtually effective: The measure of a learning environment; In: G. O'Neill, S. Moore, B. McMullin, editors. Emerging issues in the practice of University Learning and Teaching; All Ireland Society for Higher Education (AISHE), Dublin, 2005, pp. 159-167.

[16] R. Klamma, M. A. Chatti, E. Duval, H. Hummel, E. H. Hvannberg, M. Kravcik, E. Law, A. Naeve, and P. Scott, Social Software for Life-long Learning. Educational Technology & Society, 10 (3), 2007, pp.72-83.

[17] E-Learning Consultant Glossary. Address: http://www.e-learningsite.com/elearning/glossary/glossary.htm , 2003. [04/15/2010]

[18] NRW Medien GmbH (ed.), The Market for e-learning providers in Nordrhein-Westfalen(in german), Der Markt der E-Learning-Produzenten in Nordrhein-Westfalen. Düsseldorf (Germany), 2003.

[19] D. Tavangarian, M. E. Leypold, K. Nölting, M. Röser, and D. Voigt, Is e-Learning the Solution for Individual Learning?, Electronic Journal of e-Learning, Vo. 2, Issue 2, Academic Conferences Limited, 2004, pp. 273-280

[20] A. Rossett and K. Sheldon, Beyond The Podium: Delivering Training and Performance to a Digital World. San Francisco: Jossey-Bass/Pfeiffer, 2001, p. 274.

[21] M. Rosenberg, e-Learning: Strategies for Delivering Knowledge in the Digital Age. New York: McGraw-Hill, 2001, p. 28.

 ISBN: 978-1-61208-110-6

A System-On-Chip Platform for HRTF-Based Realtime Spatial Audio Rendering

Wolfgang Fohl, Jürgen Reichardt, Jan Kuhr
HAW Hamburg
University of Applied Sciences
Hamburg, Germany
Email: fohl@informatik.haw-hamburg.de, juergen.reichardt@haw-hamburg.de, jankuhr@hartschall.de

***Abstract*—A system-on-chip platform for realtime rendering of spatial audio signals is presented. The system is based on a Xilinx Virtex-5 ML507 FPGA platform. On the chip an embedded μBlaze microprocessor core and FIR filters are configured. Filtering is carried out in the FPGA hardware for performance reasons whereas the signal management is performed on the embedded processor. The azimuth and elevation angles of a virtual audio source relative to the listener's head can be modified in real time. The system is equipped with a compass sensor to track the head orientation. This data is used to transform the room related coordinates of the virtual audio source to the head related coordinates of the listener, so that a fixed position of the virtual sound source relative to the room can be attained regardless of the listener's head rotation. Head related transfer functions (HRTF) were sampled in steps of 30° for azimuth and elevation. Interpolation for intermediate angles is done by either interpolating between the coefficients of the measured HRTFs at the four adjacent angles (azimuth and elevation), or by feeding the audio signal through the corresponding four filters, and mixing the outputs together. In the latter case the required four filter processes per output stereo channel do not result in longer computing time because of the true parallel operation of the FPGA system. The system output is identical to the output of a corresponding Matlab prototype.**

***Keywords*-Mixed-reality audio; realtime HRTF interpolation; system-on-chip;**

I. Introduction

Spatial sound rendering is important for audio playback, and for creating realistic virtual environments for simulations and games. For headphone-playback devices, techniques based on head related transfer functions (HRTF) are widely used, not only in virtual reality applications, but also for stereo enhancements in home audio systems and mobile audio players [1]–[4]. There are already consumer products available, that use HRTF-based audio spatialisation with head tracking [5], [6].

For a realistic spatial impression of a virtual sound source, the perceived source location must stay fixed relative to the room when the listener's head is turned, so a headphone-based system will have to perform a coordinate transformation between head-related and room-related coordinates. A quick update of head position data is necessary to prevent a perceivable delay between head movement and HRTF adjustment.

The system described here is aimed at mixed-reality audio applications, which require a mobile device with realtime behaviour. For such applications, systems-on-chip consisting of a field programmable gate array (FPGA) with an embedded microprocessor on it are versatile and flexible platforms. The application of the time-variant HRTFs to the audio signal is done on the FPGA hardware. The filters are configured in the VHDL language (VHDL stands for Very high speed integrated circuit Hardware Description Language [7]). The tasks of signal routing and signal management are performed by the C program running on the embedded μBlaze-Processor.

In the next sections an overview of related work is given, and the fundamentals of audio spatialisation with HRTFs are outlined. Then the design of our system is described with emphasis on the partitioning of the application to hardware and software, and on the interface between the embedded μBlaze processor and the surrounding FPGA chip. Results are presented and the paper finishes with the discussion of results, summary, and outlook to future work.

II. Theoretical Background

A. Related Work

Since its beginnings in the mid-1970's, dummy head stereophony has found continuous research and development interest [8]. With increasing computing power of audio workstations it has become feasible to perform realtime rendering of a virtual audio environment [1]. The problem of proper out-of-head localisation has been addressed by many authors. It turns out, that a HRTF-based solution combined with a room reverberation model yields the best results [9]. HRTF rendering algorithms will always have to interpolate between the stored filter coefficients for measured angles. In a recent investigation the threshold of spatial resolution in a virtual acoustic environment has been investigated [3]. The reported result is, that the auditory localization has a resolution of 4° to 18°, depending of the source direction. Interpolation with minimum phase + allpass [2] In PC-based realtime systems an effective way of designing HRTF filters is to perform a minimum phase plus allpass decomposition, where the minimum phase part models the frequency response of the HRTF, and the allpass part, which is usually replaced by a delay, models the phase response [2]. Crossfading algorithms for HRTF filter interpolations are described in [2] and [3]. FPGA systems turn out to be suitable platforms for mobile audio processing [10], [11].

B. Basic Concepts

1) Spatial Audio Rendering: The human auditory system uses (at least) three binaural properties of a sound signal to determine the direction of the source: Inter-aural intensity differences (IID), inter-aural time differences (ITD), and the angular variation of the spectral properties of the sound. Concerning only the inter-aural time and intensity differences leaves an ambiguity, the "cone of confusion": All source locations on this cone yield the same ITD and IID. This ambiguity is partly removed by the angle-dependent spectral properties of the perceived sound, which result from the transmission properties of the signal path from the source to the eardrums. These three properties are completely represented by

 ISBN: 978-1-61208-110-6

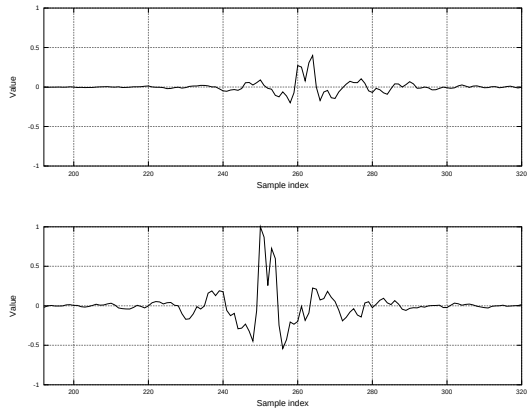

Figure 1. HRTF impulse responses for 0° elevation and 30° azimuth angle. Top: left ear, bottom: right ear

the head related transfer functions (HRTF) for given azimuth and elevation angles.

Figure 1 shows the impulse response of the HRTF at 0° elevation and 30° azimuth angle, where the time and level differences can clearly be seen. As the right ear is closer to the source, the absolute values of the right impulse response samples are larger. The sound reaches the right ear earlier which causes the shift of the maximum of the right channel to shorter delays.

The spatial rendering is done by filtering the source sound with the HRTFs of the corresponding angle and playing the resulting stereo signal back by a set of headphones.

In dynamic listening situations, listeners resolve the ambiguity of the cone of confusion by slightly turning their heads: the resulting changes in ITD, IID and sound spectrum allow a proper localization of the source.

2) FPGA System-On-Chip: The low-level FPGA architecture consists of a pool of logic blocks for combinational and registered logic, RAM-memory and DSP slices. DSP slices consist of a MAC block (*multiply-accumulate*) and registers of appropriate width to perform the multiply-accumulate operations in digital signal processing. The logic cells and DSP slices are interconnected by a user programmable switch matrix. By programming this switch matrix the user defined functionality of the system is obtained.

To handle the complexity of larger systems, the design tools for the FPGA system support a block structured approach by defining *intellectual property blocks* (IP cores), that implement special functions like FIR filters or even microprocessors (in our case the emulation of a μBlaze processor, see section III-C). Once these IP cores have been developed or purchased, the high-level design task is to properly interconnect these cores and to supply the necessary glue logic.

With the availability of a microprocessor core on the FPGA chip, it is possible to design a complete software / hardware system, where the software part is written in C and executed on the processor core, and the hardware part is specified in the VHDL language and is performing time-critical and hardware-related tasks. Systems with this architecture are called *system-on-chip (SoC)*.

C. HRTF coefficients

In preliminary measurements, HRTFs were measured with a dummy head measuring system and an audio spectrum analyser. Attenuation and phase differences were measured for 500 logarithmically spaced sine wave frequencies. The results are in good conjunction with other HRTF measurements [12], [13].

From the frequency response data the FIR filters modelling the angle dependent sound properties were designed using standard frequency-domain design techniques, with the special feature of considering the measured phase by adding it to the linear base phase. This directly introduces the interaural time delays to the FIR coefficients as can be seen in Figure 1, which is actually a plot of the FIR coefficient values over coefficient index.

D. HRTF interpolation techniques

Two approaches of interpolating HRTFs for angles between the sampled positions are considered: the first approach is the *interpolation of the filter coefficients*, the second approach is the *crossfading (mixing) of appropriate filter outputs*. In the stationary case (no variation of source angle) these two approaches are equivalent. In the next two paragraphs the two techniques are explained for the case of constant elevation. If also the elevation angle is to be interpolated, the interpolation of four filters has to be performed (see section III-D5).

1) Coefficient Interpolation: The coefficient interpolation for an angle φ that lies in the interval $[\varphi_k, \varphi_{k+1}]$ is done by linear interpolation of each of the FIR parameters b_i. The implementation of Equation 1 requires $2L$ additions and L multiplications per filter, where L is the FIR filter length.

$$b_i(\varphi) = b_i(\varphi_k) + \frac{\varphi - \varphi_k}{\varphi_{k+1} - \varphi_k} \cdot \Big(b_i(\varphi_{k+1}) - b_i(\varphi_k)\Big) \qquad i \in \{0 \dots L-1\} \tag{1}$$

It should be noted that this approach is not applicable to IIR filters.

2) Crossfade Interpolation: The crossfading approach according to Equation 2 obtains the output signal y_φ by mixing the filter outputs of the two filters corresponding to the interval limits φ_k and φ_{k+1}. The relative contribution of the two outputs is controlled by the parameter m.

$$y_\varphi = (1-m) \cdot y_{\varphi_k} + m \cdot y_{\varphi_{k+1}} \quad \text{with} \quad m = \frac{\varphi - \varphi_k}{\varphi_{k+1} - \varphi_k} \tag{2}$$

This interpolation is also suited for IIR filters. It requires only three multiplications and three additions *per audio sample* at the extra expense of running the audio material through two filters simultaneously.

A key feature of FPGA systems is the ability of *true simultaneous* execution of the filtering: The parallel operation of multiple filters does *not* require more *processing time*, instead it requires more FPGA *resources*, in particular, it requires at least *one DSP slice per filter*. The maximum filter length per DSP slice is given by the ratio of system clock frequency and audio sampling rate [11], [14]. The device presented here works with 125 MHz processor clock frequency and 44.1 kHz audio sampling rate, allowing a maximum filter length of $\frac{125 \cdot 10^6\,\mathrm{Hz}}{44.1 \cdot 10^3\,\mathrm{Hz}} \approx 2800$.

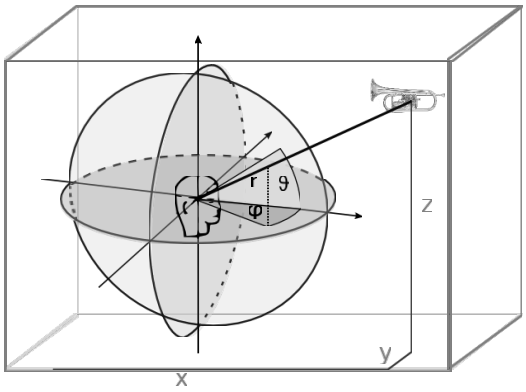

Figure 2. Head Related Coordinates r, ϑ, ϕ and Room Related Coordinates x, y, z

E. Head tracking

For a realistic spatial impression of headphone-based 3D-audio the transformation from head-related to room-related coordinates according to Figure 2 is necessary. For a virtual audio source that is supposed to remain fixed in the room, the orientation of the listener's has to be continuously measured and a corresponding correction of the apparent source direction has to be applied.

III. System Design and Implementation

A. Requirements

The requirements listed here are a consequence of the intended use of the system as a mobile device for spatial audio rendering in mixed-reality environments.

- Low power consumption, small and light system.
- Audio signals in CD quality: 44.1 kHz sampling rate, 16 Bit word length, 2 channels.
- Multiple parallel FIR filters with $\geq$ 512 coefficients.
- Tracking of the head azimuth and pitch (forward) angle. For future developments also the roll (sideways) angle and the acceleration data for three axes must be measured.
- Sufficient memory to hold the filter coefficient sets.
- Audio latency $\leq$ 10 ms to avoid perceivable delay.
- Architecture must be extensible to multiple independently moving virtual audio objects.

B. System Components

Our system is based on a Xilinx Virtex-5 ML507 FPGA evaluation board. In addition to the FPGA chip this board provides a large number of resources and interfaces, the most important ones being an AC97 audio interface, a RS 232 serial interface, a Compact Flash (CF) memory interface, for which a file system driver is provided, and a DDR2 RAM interface. In addition there is a DIP-switch interface for simple user interaction (e.g., switching of operating modes).

The azimuth and elevation angles of the listener's head are provided by a compass sensor (Ocean Server OS5000 [15]), that is mounted on the headphone clip and is connected to the system via the RS 232 link.

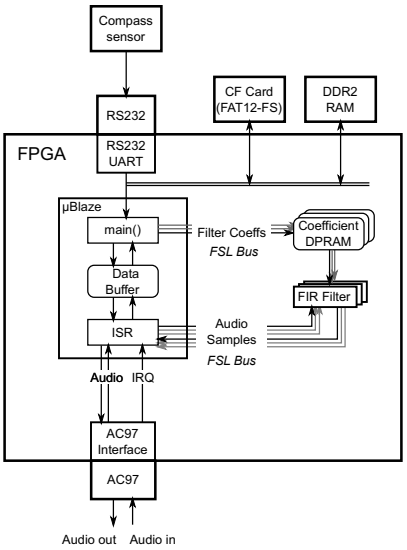

Figure 3. FPGA Components and Interfaces

C. Hardware Architecture

On the FPGA chip is the embedded μBlaze processor IP core for signal and data management, the FIR filter blocks, and the block interconnection logic. The μBlaze is a 32-bit big-endian RISC processor with a library to access the FPGA chip hardware and a runtime environment for a C `main()` routine. The processor was configured without a floating-point coprocessor to save FPGA resources for the HRTF filters. The FPGA chip is configured with 64 kB on-chip RAM for the μBlaze processor, Dual-ported RAM blocks for the FIR filter coefficients, 256 MB of external DDR2 memory and a 512 MB CF card with a FAT12 filesystem, which can be accessed by the standard C file I/O routines. Figure 3 gives an overview of the relevant system components and interfaces.

External devices like the AC97 and the RS 232 can be accessed by the μBlaze program with library functions provided by Xilinx. The interconnection with the on-chip FIR blocks is established via the *Fast Simplex Link* (FSL) bus. The FSL bus is an unidirectional bus which also performs the synchronisation of sender and receiver to the system clock. Three FSL bus instances per filter were implemented for parameter transfer, audio input, and audio output.

Incoming audio samples generate an interrupt which will be served by the interrupt service routine (ISR) on the μBlaze.

The FIR filters for the HRTF filtering are implemented in a direct form I (DF1) structure as a sequential processing pipeline utilising only one DSP slice per filter block [14].

The active FIR filter coefficients are stored in a dual-ported RAM (DPRAM), so the coefficient update and the filtering may be executed asynchronously.

D. Software Architecture

1) Operating Modes: Two basic modes were implemented: a *realtime mode*, and an *offline mode*. In realtime mode, spatial audio rendering is triggered by the interrupts of the AC97 interface. Each incoming sample raises an interrupt, the ISR takes the input sample, transfers it to the filters, receives the filtered audio sample, performs

 ISBN: 978-1-61208-110-6

the mixing if required, and sends it to the AC97 interface for playback.

In *offline* mode audio data is read from wav-files stored on the CF memory card. Audio samples are processed in the same way as in realtime mode, but in offline mode the whole program is executed at maximum speed in the cyclical `main()` program, and the output is stored in a wav-file on the CF card for further evaluation.

In realtime mode the timing and latency of the two interpolation techniques are investigated; in offline mode the correctness of implementation is checked by comparing the output wav-files with the results of corresponding MATLAB computations.

For both modes either of the two interpolation techniques, *coefficient interpolation* or *crossfading* may be selected as interpolation mode. In coefficient interpolation mode, each data update from the compass sensor triggers the calculation of a new interpolated coefficient set for the filters for left and right audio channel according to Equation 1. The coefficient sets are then transferred via the FSL bus to the coefficient DPRAM on the hardware. In crossfading interpolation mode according to Equation 2, each update of the compass sensor data triggers the computation of a new mixing factor, which is written to the global data space where the ISR can access it.

2) Filter Implementation: The FIR filter coefficients of the HRTFs were calculated in MATLAB from measurement data. The normalisation of the filter coefficients was done in an empirical way, starting with $L1-$ normalized coefficients. These coefficients lead to very low output amplitudes and thus a poor S/N ratio. For typical input signals a normalisation factor was determined experimentally, that led to no audible overflow.

Data has been converted to 16-bit Q15 integer format using the MATLAB fixed-point toolbox, and stored in binary format on the CF card. Intermediate results in the filter block are stored in 32 bit wide registers.

Filter coefficients are loaded from CF memory by the `main()` routine of the μBlaze C program, and are transferred to the hardware filters via the FSL bus connections of the filters.

3) Head Tracking: The azimuth and elevation (pitch) angles of the compass sensor are read in the `main()` routine, and are used for calculating the audio source angles in head-related coordinates. The compass sensor provides a third angle, giving the sideways bend of the head (*roll* angle). This angle is neglected because performing the necessary trigonometrical computations in integer arithmetic on the microprocessor would be too time-consuming.

4) Signal Routing: All audio signals are processed by the μBlaze ISR and transferred to the filters via the FSL bus. To minimize overhead, the 16 bit samples for left and right channel are combined to a 32 bit word and transferred as one item to the filters. The filter connection logic then extracts the two samples from the transferred word. Figure 4 shows the interconnection between the μBlaze processor and the hardware filters.

Filter coefficients are transferred from processor to hardware from within the processor's `main()` function using the same technique of transferring two 16 bit coefficients at once.

5) Implementation of Crossfading: Figure 5 shows the principle of crossfading interpolation for azimuth and elevation. The mono input signal is fed to four stereo FIR filter pairs, for the top left, top right, bottom left, and bottom right position of the interpolation interval. From the compass sensor data the azimuth and elevation

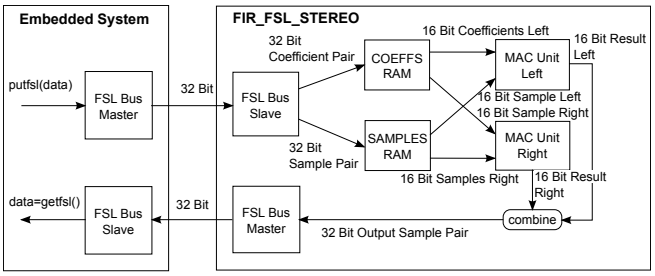

Figure 4. Data Flow for the Filtering Process

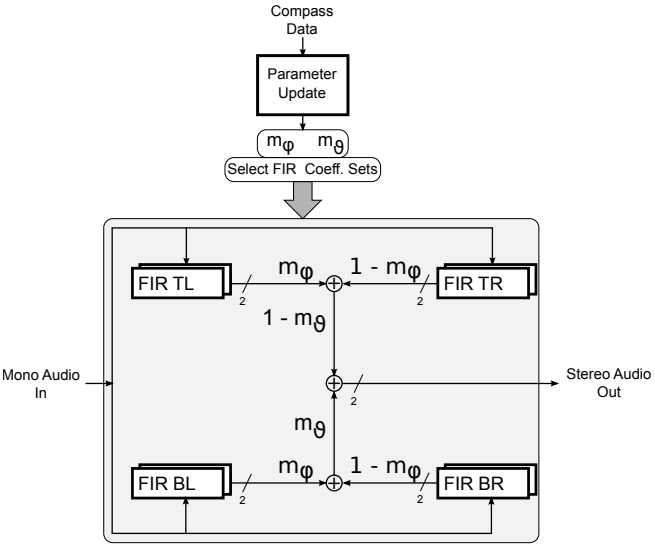

Figure 5. Signal Crossfading Interpolation between Top Left (TL), Top Right (TR), Bottom Left (BL), and Bottom Right (BR) Filter Outputs

mixing parameters m_φ and m_ϑ are computed, and the filter outputs are superposed according to the two mixing parameters. When the azimuth and elevation data from the compass data indicate that the current interpolation interval has been left, the FIR coefficient sets are reloaded according to the new interval.

IV. Results and Discussion

A. Verification of the Static Filtering Algorithms

The filtering algorithms for both interpolation techniques have also been implemented in Matlab using the fixed-point toolbox, and the results have been compared with the wav-files that are produced by the FPGA system in offline mode. Test cases were filtering at the measured HRTF angles and at different constant interpolated azimuth and elevation positions.

In these measurements the outputs of the FPGA system and the Matlab implementation were bit-wise identical.

Both interpolation algorithms produced identical output signals.

B. Signal Processing Latencies

To assess and optimize system performance, and to examine, if the data transfer times will limit the maximum number of audio objects (i.e., independent filter processes), detailed timing measurements have been carried out.

 ISBN: 978-1-61208-110-6

Table of System Latencies:

t_{AC97}: AC 97 audio subsystem (Note 1)	1 ms
t_{FSL}: FSL transfer (round-trip) of one 32 bit word	300 ns
t_{FIR}: FIR processing L=512	4.2 μs
t_{Parm}: Parameter transfer for 512 32-bit parameter pairs	120 μs
t_{gd}: Filter group delay	$\leq$ 7 ms
t_{CS}: Compass sensor sampling time	25 ms
t_{CT}: Compass sensor data transfer (Note 2)	2.4 ... 22 ms
t_{AL}: System audio latency for N audio objects (Note 3)	
$t_{AL} = t_{AC97} + N \cdot t_{FSL} + t_{FIR} + t_{gd}$	8 ms
t_{HLI}: Head tracking latency	
$t_{HLI} = t_{AL} + t_{CS} + t_{CT} + t_{Parm}$	35 ... 55 ms

Notes:

1) This value has been measured in a previous work [11]. It is the time for transferring a stereo audio sample from the AC97 to the FPGA, decode it in hardware, re-code and pass it back to the AC97 output *without* routing the audio signal through the μBlaze processor.
2) The compass data is transferred in ASCII. Small values have less digits and thus need less transfer time then large values.
3) Delay between audio input and audio output.

The table shows that the filtering of one audio sample takes much longer time than the FSL bus transfer, so the ISR can be substantially sped up by replacing the standard *blocking* data transfer routines `putfsl()` and `getfsl()` by their *nonblocking* counterparts `nputfsl()` and `ngetfsl()`. In the case of blocking transfer as shown in Figure 6, the `getfsl()` routine has to wait until the filtering process has finished, whereas in the nonblocking case as shown in Figure 7 the `ngetfsl()` routine returns immediately. In the latter case the ISR gets the result of the *previous* filtering process, adding an extra latency of 1 audio sampling time to the system, which is negligible compared to the group delay of the filter.

The limiting factor for the number of audio objects is the fact, that each audio object will require one FSL bus transfer that is executed *sequentially* in the C program of the μBlaze processor. An upper limit can be estimated by the requirement, that all the FSL bus transfers must be completed within one audio sampling period t_S:

$$N \cdot t_{FSL} < t_S \tag{3}$$

For $t_S = 1/44100\,\mathrm{s}$, the upper limit for N is 75 audio objects.

C. Filtering With Compensation of Head Movement

At the moment of writing, only qualitative listening tests have been performed. The compensation of the head movement drastically increases the spatial impression of the rendered audio material. No front-back ambiguity was noticed. A much better externalization of the sound was perceived, even in the case of source locations directly in front of the listener, where externalization is known to be most difficult to obtain [9].

The latency of approximately 40 ms for the compensation of the head movements by evaluating the compass data is perceivable only at fast and abrupt head movements, where it causes a slight irritation. For head movements at moderate speed no latencies are perceivable. This is due to the limited angle resolution (4° to 18°) of the human auditory system [3].

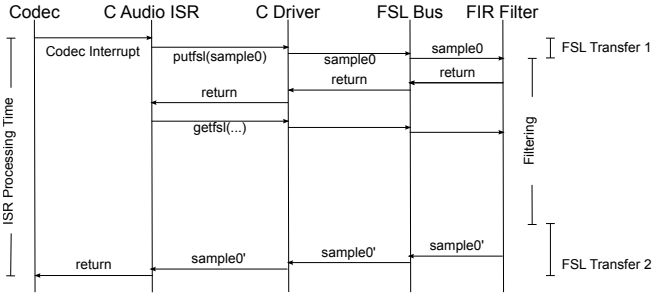

Figure 6. Blocking Filtering Process Sequence Diagram

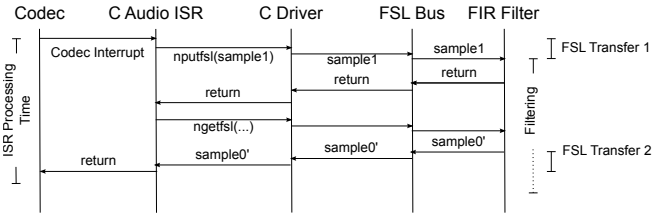

Figure 7. Non-Blocking Filtering Process Sequence Diagram

1) Coefficient Interpolation: The coefficient interpolation algorithm according to Equation 1 leads to artifacts at fast head movements or fast moving sources. This is due to the fact, that the coefficient modification is asynchronous with the filtering, so for fast moving sources the coefficient sets may be inconsistent during the parameter transfer. As shown in section IV-B, this transfer takes $120\,\mu s$, which is approximately 5 audio sampling times, so 5 audio samples will be filtered with inconsistent filter coefficients, which will cause the audible artifacts.

2) Crossfading: With the crossfading interpolation algorithm according to Equation 2 and Figure 5, no artifacts were audible in our listening tests, as long as the source azimuth and elevation angles remain in the same interpolation interval. In the moment, where the interval boundaries are crossed, there is the risk of artifacts which arise from the same reason as in the parameter interpolation case: The parameter update of the four involved filter pairs takes longer than one audio sample, so the filter coefficients are inconsistent during this update.

V. Conclusion

A. Summary

The system-on-chip platform presented in this paper turned out to be well suited as a mobile component of a mixed-reality audio system. The maximum number of virtual audio sources is limited by the 126 DSP slices on the FPGA chip. Two slices are needed for the headphone compensation, so 124 slices remain for the HRTF filters. One audio object requires 8 DSP slices (4 stereo filter pairs at the borders of the interpolation interval), so $\lfloor 124/8 \rfloor = 15$

independent objects could be rendered with the current system design.

Compared to the upper limit of 75 audio objects from the consideration of the bus transfer times in section IV-B, it turns out that the number of available DSP slices is actually the limiting factor for the number of audio objects.

The preferred HRTF interpolation technique is the crossfading of filter outputs. Here the only problem to overcome is the disturbance that occurs when the limits of the interpolation interval are left. In the next section a possible solution to this problem will be outlined.

The coefficient interpolation technique is not suited for this system platform, because one parameter update takes about 5 audio sampling times. During this time the filtering occurs with inconsistent coefficient sets leading to audible artifacts in the output signal.

B. Outlook

Systematic listening tests will have to be conducted to investigate whether the interpolation introduces a perceivable degradation of audio quality. Furthermore, tests will have to be carried out to determine the source localization accuracy with and without head movement compensation.

One issue with the current implementation of the crossfade interpolation is the disturbance caused by reloading the filter coefficients, when the source angles cross the boundaries of the interpolation intervals. This problem can be overcome by introducing additional "standby" filters that provide the output signals of the adjacent angle intervals.

The current implementation uses the audio input of the AC97 subsystem as audio source. To render multiple audio objects, there will be needed multiple source audio streams. These audio streams will have to be transferred to the system via the network or the USB interfaces of the Virtex board.

With the HRTF filtering only the *direction* of a virtual audio source can be rendered. To additonally reproduce the *position* of a virtual source, the influence of the *source distance* on the perceived sound must be modeled and applied to the output signal. Currently we are investigating these effects with the aim of creating a sufficiently simple distance model that can be executed in real-time on the Virtex system-on-chip platform.

REFERENCES

[1] J. W. Scarpaci and H. S. Colburn, "A System for Real-Time Virtual Auditory Space," in *Proceedings of ICAD 05-Eleventh Meeting of the International Conference on Auditory Display, Limerick, Ireland*, vol. 9, 2005, pp. 6 – 9. [Online]. Available: http://www.dell.org

[2] B. Carty and V. Lazzarini, "Binaural HRTF Based Spatialisation: New Approaches and Implementation," in *Proc. Of the 12th Int. Conference on Digital Audio Effects (DAFx-09), Como, Italy*, 2009.

[3] A. Lindau and S. Weinzierl, "On the Spatial Resolution of Virtual Acoustic Environments for Head Movements in Horizontal, Vertical and Lateral Direction," in *Proc. Of the EAA Symposium on Auralization, Espoo, Finland*, vol. 17, 2009, pp. 15 – 17.

[4] A. Lindau, "The Perception of System Latency in Dynamic Binaural Synthesis," in *NAG/DAGA 2009 - Rotterdam*, 2009, pp. 120 – 180.

[5] Beyerdynamic, "Headzone System," Accessed 08/05/2010, available at http://europe.beyerdynamic.com/shop/hah/headphones-and-headsets/at-home/headphones-amps/headzone-home-hz.html.

[6] S. R. LLC, "Realiser A8," Accessed 08/05/2010, available at http://www.smyth-research.com/products.html.

[7] V. Analysis and S. Group, "Behavioural languages–part 1: Vhdl language reference manual," IEC Standard 61691-1-1: 2004, 2004.

[8] J. Blauert, *Spatial Hearing The Psychophysics of Human Sound Localization*. MIT Press, 1983, vol. 9.

[9] T. Liitola, "Headphone Sound Externalization," Master's thesis, Helsinki University of Technology, Department of Electrical and Communications, 2006.

[10] S. Kurotaki, N. Suzuki, K. Nakadai, H. G. Okuno, and H. Amano, "Implementation of Active Direction-Pass Filter on Dynamically Reconfigur able Processor," in *IEEE/RSJ International Conference on Intelligent Robots and Systems*, 2005, pp. 8912 – 8913.

[11] W. Fohl, J. Matthies, and B. Schwarz, "A FPGA-based Adaptive Noise Cancelling System," in *Proc. of the 12th Int. Conference on Digital Audio Effects (DAFx-09), Como, Italy*, 2009.

[12] B. Gardner and K. Martin, "HRTF Measurements of a KEMAR Dummy-Head Microphone," MIT Media Lab, Cambridge, MA 02139, MIT Media Lab Perceptual Computing Technical Report No.280, 1994.

[13] O. Warusfel, "LISTEN HRTF Database," Accessed 08/05/2010, available at http://recherche.ircam.fr/equipes/salles/listen/index.html.

[14] J. Reichardt and B. Schwarz, *VHDL-Synthese Entwurf digitaler Schaltungen und Systeme*, 5th ed. München: Oldenbourg Wissenschaftsverlag, 2009.

[15] O. S. T. Inc., "Digital Compass Users Guide, OS5000 Series," Accessed 08/05/2010, available at http://www.ocean-server.com/download/OS5000_Compass_Manual.pdf.

 ISBN: 978-1-61208-110-6

Classification of Emotional Speech in Anime Films by Using Automatic Temporal Segmentation

Yutaro Hara
Graduate School of Computer and Information Sciences
Hosei University
3-7-2 Kajino-cho, 184-8584 Koganei, Japan
Email: yutaro.hara.h2@gs-cis.hosei.ac.jp

Katunobu Itou
Faculty of Computer and Information Sciences
Hosei University
3-7-2 Kajino-cho, 184-8584 Koganei, Japan
Email: itou@hosei.ac.jp

***Abstract*—This paper describes emotional speech classification in anime films. An emotional speech corpus was constructed by using data collected over 8 h. The corpus consists of emotional speech material of a total of 984 utterances. Five emotions, namely, joy, surprise, anger, sadness, and the neutral case, were labeled and divided into training and test data. In a previous study, Attack and Keep and Decay were adopted as parameters to describe temporal characteristic of the power transition. This paper proposed an improved method of A-K-D unit, and evaluated it. As a result, acoustic features of the proposed method were more effective than the conventional method when we used for GMM.**

***Keywords*—Emotion, animated film, Speech analysis, Pattern classification.**

I. Introduction

A number of “anime” films have been produced in Japan. More than 140 anime films were produced in 2008; these include children’s film as well as other genres such as sci-fi, horror, and sports. They are often produced as a drama series. Anime films involve typical directorial techniques and a voice acting technique called “anime tone” in Japan. In this acting technique, the speaking style and emotional expression involve a characteristic sound and prosody, which have not been thoroughly studied.

In recent years, many studies have been conducted on emotional classification. However, emotional classification is complicated because it depends on clear acoustic features that may not be determined, and hence, depends on the word length and speaker. This study deals with the emotional speech of voice actors/actresses in a cartoon film; as yet, few studies have been conducted on cartoon films. A previous study [1] focused on emotional speech classification of voice actors/actresses in an animated movie that is “The Incredibles” [2]. Automatic classification enables the maintenance of an emotional speech corpus by automatically applying to it the label of emotions of the corpus of some other cartoon film. In addition, a support of the performance exercise of the emotional expression is enabled if it apply to an automatic classification system of the emotional expression such as the cartoon film and will open possibility of some production of the cartoon film work.

Section II introduces emotions and some samples used in this paper. Section III presents an explanation of temporal structure of emotional speech; it also introduces acoustic features and the proposed method of A-K-D unit estimation. Section IV describes 4 classifiers used in this paper. Section V describes Feature Subset Selection (FSS) and Sequential Forward Floating Selection (SFFS). In Section VI, we present a result of the recognition rate in each emotion. In Section VII, we discuss the result and effectiveness of the proposed method. Section VIII describes the general summary and some future issues.

II. Speech Database of Emotions

Emotional speech has been dealt with a spontaneous speech and a conscious speech. The spontaneous speech is difficult to specify the emotion of the moment of a speech, and it is easy to be dependent on a mental condition and environment. Otherwise, the conscious speech such as acting emotional speech can detect the emotion by the power and tone of voice, and it is easy to judge a emotion for some observers subjectively. Furthermore, if the voice acting is mastered like some voice actors/actresses, the emotional expression is easy to accord with a subjective evaluation with only a pitch. We use acting emotional speech of voice actors/actresses because of the superior ability that it is difficult to be dependent on a mental condition and the environment and the ease of a subjective emotional classification by some observers.

In this study, emotional speech is extracted from DVD-Video of “Honey and Clover” a Japanese anime film. This film is based on everyday school life and consists of 24 episodes. This film contains a lot of emotional speech samples because 4 male and 3 female important characters appeared in the film.

Emotional speech samples of 4 male voice actors and 2 female voice actresses were extracted manually with a sampling frequency of 48 kHz and a bit rate of 16 bits from 24 episodes (total film length was approximately 8 h) of this anime drama. The samples were classified into seven categories such as “Joy”, “Surprise”, “Anger”, “Sadness”, “Neutral”, “Others” and “No emotion”. We selected the previous 5 emotions because they’re used most [1][3][4][5]. We discarded speech samples that have any overlap of plural speech and were annotated with different emotional labels by two annotators. A details of the emotional categories of the sample are listed in Table 1.

“Other” in Table 1 shows that the samples which did not include 5 emotions meet some condition, and which did not

 ISBN: 978-1-61208-110-6

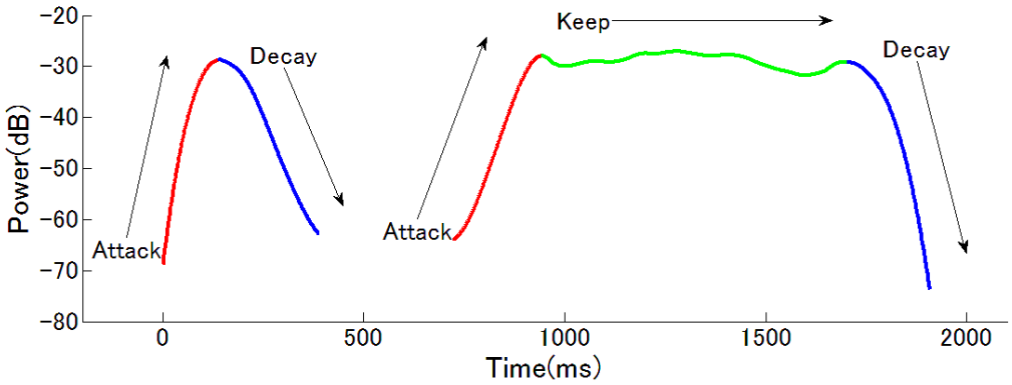

Figure 1. A-K-D unit of emotional speech

TABLE 1. Number of samples of emotional speech

	Joy	Surprise	Anger	Sadness	Neutral	Other
Training	167	75	128	196	195	/
Test	35	42	27	57	62	/
Total	202	117	155	253	257	1671

accord in the subjectivity emotional classification. In addition, it is assumed that 5 emotions do not include plural emotions.

III. Acoustic Feature Analysis of Emotional Speech

A. Temporal structure of emotional speech

Emotional speech has characteristic temporal structures in power transition and pitch transition such as three-layerd models that were modeled by F0 contour,power envelope, spectrum [6], and Support Vector Machine (SVM) and Gaussian Mixture Model (GMM) and K-nearest neighbor method (k-NN) that were modeled by energy mean of fall-time and Energy mean of rise-time and so on [4]. Some temporal characteristics of a power transition and a F0 transition were considered as derivatives (delta features) of whole segments [4][6]. Mitsuyoshi [5] proposed a temporal structure model of utterance based on power transition. In this model, an utterance is divided into three parts; “Attack”, which lasts from the beginnings of the utterance to the peak in power domain, “Keep”, which lasts during keeping the power level, and “Decay”, which begins decreasing the power. Emotional speech, especially, has characteristics in Attack and Decay. In Japanese anime, the beginnings of “Joy” utterance and “Anger” utterance have high pitch, so that the mean or the maximum of F0 of Attack unit is higher than other emotional types. For “Sadness” utterances, duration of Keep tend to be shorter, and their power and pitch do not vary so much. In Decay unit of sadness utterances, therefore, magnitude of derivation of F0 and/or power tend to be small. Mitsuyoshi classified a speech into A-K-D for a unit and it was called “ A-K-D unit ”. This study performed A-K-D unit detection, and modeled acoustic features in each unit (Attack-Keep-Decay). Figure 1 shows examples of A-K-D unit of emotional speech.

In the study [5], in Attack unit and Decay unit, inclination and maximum value and continues length in power were used. In Keep unit, continues length and power average and Δpower average and, power variance were used. With these acoustic features, emotional classification was performed for the spontaneous speeches and the conscious speeches. The structural modeling based on the A-K-D unit is promising, but, there are some problems to be solved. In this paper, we improved the detection algorithm of A-K-D unit. We also examined more acoustic features, such as ΔF0 and minimum value of power and F0 based on the A-K-D unit. Further, when the A-K-D unit is estimated, we do not consider F0 only the power transition.

B. Acoustic features

In emotional speech analysis, F0 and power have been widely used as acoustic features [1][4][5][6][9]. In this study, both F0 and power are used as acoustic features. In addition, 77 dimension acoustic features shown in Table 2 were used in total. F0 represents a pitch of a speech, and it was extracted using STRAIGHT [7] in this study. Power is calculated as ratio of a total of power spectra in 70 msec segment to a silent section (the power level of background noise). MFCC(Mel-Frequency Cepstrum Coefficient) represents a frequency response of a human vocal tract, and have been used in emotional speech recognition and speech recognition [8]. In addition, we normalized all acoustic features by using the average of acoustic features of Neutral data as a standard of each speaker.

C. A-K-D unit estimation

1) Conventional method of A-K-D unit estimation: When the Δp in equation (1) crossed the threshold, Attack begins.

$$\Delta p = p_n - p_{n-1} (n = 1, 2, 3, 4, ...) \quad (1)$$

Attack is defined to last during the segment where $\Delta p > 0$. Keep is defined to last while an absolute value of Δp keeps under the threshold. When it crosses the threshold, Keep finishes and Decay begins. The Decay slope was defined by $\Delta p \leq 0$. Decay finished when $\Delta p \geq 0$ was detected, which means one A-K-D unit ended, and next A-K-D unit estimation begins. The conventional method of estimating A-K-D unit consists of the above-mentioned items.

 ISBN: 978-1-61208-110-6

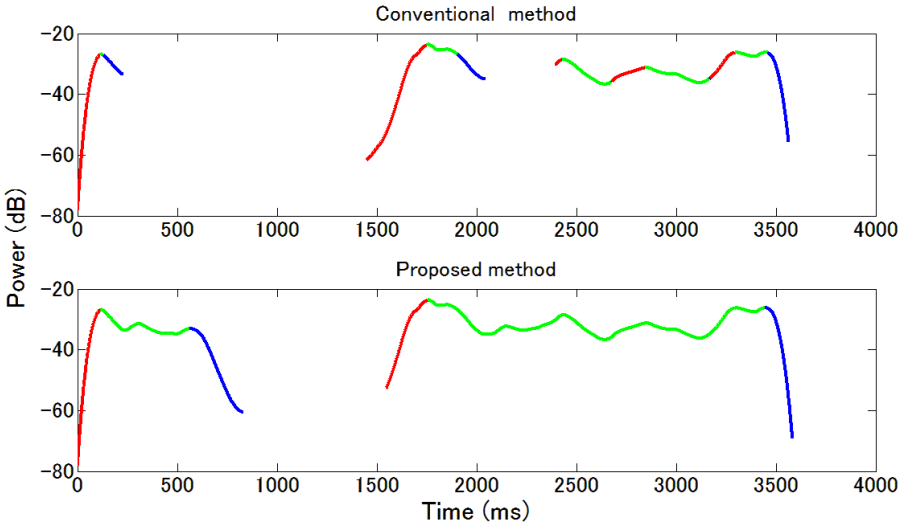

Figure 2. A-K-D unit of conventional method and proposed method (red line is Attack, green line is Keep, blue line is Decay)

TABLE 2. Acoustic features

Extraction segment	Acoustic feature
A-K-D Unit	F0 maximum
	F0 minimum
	F0 range
	F0 mean
	F0 median
	F0 Standard deviation
	F0 inclination
	Power maximum
	Power minimum
	Power range
	Power mean
	Power median
	Power Standard deviation
	Power inclination
	Continuous length
Whole Speech	F0 maximum
	F0 minimum
	F0 range
	F0 mean
	F0 median
	F0 Standard deviation
	ΔF0 mean
	ΔF0 maximum
	ΔF0 minimum
	ΔF0 mean of positive incline
	ΔF0 mean of negative incline
	Power maximum
	Power minimum
	Power range
	Power mean
	Power median
	Power Standard deviation
	ΔPower mean of positive incline
	ΔPower mean of negative incline
	ΔPower median of positive incline
	ΔPower median of negative incline
	12 dimensions MFCC

The conventional method, however, the power transition have a single peak as left plot of Figure1, a unit that should be recognized as Decay may be misrecognized as Keep.

The other way around, the case of the power transition does not have a single peak, When $|\Delta p|$ crossed the threshold, the unit that should be recognized as Keep may be misrecognized as Attack or Decay. In this study, we proposed a method to improve the above-mentioned point.

2) Proposed Method of A-K-D unit estimation: As preprocessing of A-K-D unit estimation, as a first, we performed the Voice Activity Detection (VAD). The V/UV decision of STRAIGHT was used for VAD. If there is an interval between a voiced segment and the next voiced segment within 100msec, it was treated as one voiced segment. After the VAD, we performed A-K-D unit estimation based on the power transition in the voiced segment. In addition, the power was smoothed by using Savitzky-Golay Filter with 3 degree and a window width of 201 msec.

When voiced segment begins, Attack begins. When the equation (2), Attack finishes and Keep begins.

$$|R^n_{Attack}| > 0, R^{n+1}_{Attack} < 0 \tag{2}$$

Here, R is an inclination of the power that is calculated every 10 msec.

Next, we decide a rough shape of the power transition. Here, R_{Attack} is the inclination of the power from Attack began to Attack finished, and R_{KD} is the inclination of the power from Keep began to voiced segment finished.

If $|R_{KD}| < R_{Attack}$

- There is not a single peak like left plot of Figure 1, that is to say, there is not Keep unit.
- Attack began equals Decay begins.
- When the voiced segment finishes, Decay finishes.

Else

- There is a single peak like right plot of Figure 1, that is to say, there is Keep unit.

$$R^n_{Keep} > 0, R^{n+1}_{Keep} < 0 \tag{3}$$

$$\left|R^{n+1}_{Keep}\right| > max(|R_{Keep}|) \tag{4}$$

- When the equation (3) and (4), Keep finishes and Decay begins.

 ISBN: 978-1-61208-110-6

- When the voiced segment finishes, Decay finishes.

When one A-K-D unit ended, next A-K-D unit estimation begins This study decided one A-K-D unit by the above-mentioned algorithm. Figure 2 shows a result of A-K-D unit estimation by conventional method and proposal method. As the threshold in the conventional method, we used the average of Δp of the whole speech.

In Figure 2, the proposal method can estimated A-K-D unit more precisely than the conventional method. The region that should be recognized as Keep was misrecognized to be Attack and Decay, and there was the region where Decay unit is not found in after Keep unit in the conventional method of Figure 2. But, these points were improved by the proposal method.

IV. EMOTIONAL SPEECH MODEL

Emotional speech classification has been performed with various classifiers in some previous studies [3][4][9][10]. In this study, acoustic features of emotional speech of some voice actors/actresses were modeled by K-nearest neighbor method (K-NN), Probabilistic Neural Network (PNN), Support Vector Machine (SVM),Gaussian Mixture Model (GMM). We performed Feature Subset Selection (FSS) in each model, and performed Gaussian Mixture Size Selection (GMSS) in GMM.

A. K-nearest neighbor method (K-NN)

One of the method that is a standard in pattern recognition is neighbor method. When new data are given, nearest neighbor method calculate distance with the other data and classify it in a category same as data in the neighborhood most. On the other hand, K-NN refers to not only the nearest data but also the K unit data of the neighborhood, and it classifies the class where most learning patterns belong to.

B. Probabilistic Neural Network (PNN)

K-NN refers to the K unit data of the neighborhood, but PNN refers to data in the distance to decide a category. PNN has a high recognition precision. Because PNN approximated precisely to the relation of true probability density distribution between each category, by putting a kernel function formed from a sample pattern on top of one another. Therefore, if the number of sample patterns increases, the recognition rate nears an ideal value according to Bayesian statistics, and the classifier with high recognition rate can be realized.

C. Support Vector Machine (SVM)

SVM is one of the classifier with supervised learning. SVM is a method to constitute a classifier of two classes with a linear threshold element, and there are three main characteristics.

1) It can be expect a high generalization ability, because it decide an identification plane by the margin maximization.
2) The learning is resulted in quadratic programming problem by Lagrange multiplier method, and a local optimal solution become a global optimal solution by all means.
3) It performs linear identification on the feature space by defining the feature space that reflected prior knowledge for the space of the identification object. And, It is not necessary to show a conversion to the feature space explicitly by defining the kernel function that expressed dot product on the feature space.

Because of these characteristics, SVM shows high recognition rate for the unlearning data. In addition, SVM shows such characteristics because a optimization is performed for both the recognition error and the generalization in learning. In this study, we implemented SVM by using LIBSVM[11].

D. Gaussian Mixture Model (GMM)

A mixture Gaussian distribution is expressed in probabilistic model called the mixture distribution by piling of some single Gaussian distribution. GMM expresses arbitrary consecutive density functions by coordinating a weight coefficient, and a mean of each distribution, and covariance. There is the case that it is not caught a distribution even if maximum likelihood estimation is used in single Gaussian distribution, but when maximum likelihood estimation is used by linear combination of some Gaussian distribution, GMM can catch the distribution. GMM can express in the next equation:

$$p(x) = \sum_{k=1}^{k} \pi_k N(x \mid \mu_k, \Sigma_k) \tag{5}$$

Here, π_k is the mixture coefficient, and $N(x \mid \mu_k, \Sigma_k)$ is the mixture factor. Each Gaussian distribution have a peculiar mean μ_k and a covariance Σ_k. In addition, the mixture Gaussian distribution is determined by parameters such as the weight coefficient, a mean, and a covariance. These parameters are determined by using maximum likelihood estimation. A function that a likelihood function becomes maximum can demand by using the maximum likelihood estimation.

1) Gaussian Mixture Size Selection (GMSS): The number of Gaussian distributions is an important element in GMM. In a previous study [12], a BIC-based method called Gaussian Mixture Size Selection (GMSS) has been proposed; this method can be used to control the complexity of the speaker model and to determine the number of Gaussian distributions in GMM. In a previous study [9], GMSS was used with the Akaike Information Criterion (AIC) [13]. BIC and AIC are the most commonly used an evaluation standard in an information standard. In this study, we are used for GMSS with AIC.

Let $X = \left\{x_j \in R^d : j = 1, \ldots, N\right\}$ be the training data set, $\lambda = \{\lambda_i : i = 1, \ldots, K\}$ be the candidates for the parameter of models. AIC of GMM is given by the following equation:

$$AIC_i = logP(X \mid \lambda_i) - (2d + 1) \tag{6}$$

Here, $logP(X \mid \lambda_i)$ is the logarithm of the likelihood of training data X by GMM, d is the number of acoustic features.

The mixture size of GMM is determined by evaluating the following:

$$\Delta AIC = AIC_M - AIC_{2M} \tag{7}$$

The mixture size is doubled if ΔAIC is negative. Otherwise, it is represented as M. Thus, the number of Gaussian

 ISBN: 978-1-61208-110-6

distributions can be set according to the training data; when the training data is sparse, the mixture size is expected to be small. In this study, we find the number of the mixture distributions that is most suitable for every class, and make a high performance GMM.

V. Feature Subset Selection (FSS)

The database used for pattern recognition is represented by some examples of acoustic features and generally consists of a high-dimensional vector. Therefore, the calculation cost may increase due to the many features and classes. Furthermore, the best classification cannot be achieved due to noise. However, the recognition rate can be improved by using FSS [14].

By FSS, a candidate with the most effective acoustic feature set for classification is selected from a given the acoustic feature set; then, a subset s consisting of d features containing the information required for classification is identified from a feature set S with D features. Note that since the output value of the evaluation function is high, it is a good feature set.

A. Sequential Forward Floating Selection (SFFS)

There is various ways in Feature Subset Selection. A method that till a certain standard is satisfied by increasing features or reducing features is generally used. A representative method involves the use of a search algorithm called forward type that increases the number of features from 0 to higher values. Sequential forward selection (SFS) proposed by Whitney is a representative of the forward type [15]. Another method involves the use of a search algorithm called backward type that determines the best feature set by reducing the number of features from the total number of features. Sequential backward selection (SBS) proposed by Marill and Green is a representative of the backward type [16]. These algorithms can be easily used in various applications. However, it may not be demanded the best feature combination because they are one-direction search algorithm and cannot carry it out about the all possible feature combination. Therefore, there is SFFS (Sequential Floating Forward Search) suggested as the algorithm that improved SFS and SBS by Pubdil [17]. SFFS is the Floating type algorithm which put Forward type algorithm and Backward type algorithm together. And, SFFS is the higher performance than SFS and SBS. SFFS algorithm is shown in Figure 3.

In many emotional speech classification, SFFS is used as feature subset selection [9][18]. In this study, we use SFFS to demand the high performance feature combination. In addition, SFFS is performed by PNN which is used a error rate as evaluation function.

VI. Result

We examined with open test. The 3 feature combinations were compared by four classifiers which are K-NN and PNN and SVM and GMM. A result is shown in Table 3. In Table 3, a set A is Whole Speech features, a set B is Whole Speech and A-K-D unit (conventional method) features, a set C is Whole Speech and A-K-D unit (proposal method) features.

Initialize:
$Y = \{ y_j \mid j = 1,\cdots,D\}$ *//Input and available measurements//*
$X = \{ x_j \mid j = 1,\cdots, k, x_j \in Y\}$, $k = 0,1,\cdots,D$ *// ouput//*
$X_0 = 0$, $k = 0$
Execute:
Repeat
Step1(Inclusion)
$x^+ = \text{argmax } J(X_k + x)$
$X_{k+1} = X_k + x^+$
$k = k+1$
Step2(Conditional Exclusion)
$x^- = \text{argmax } J(X_{k-1} - x)$
if $J(X_{k-1} - x^-) > J(X_{k-1})$
$X_{k-1} = X_k - x^-$
$k = k - 1$
goto Step2
else
goto Step1
until k is the number of features required

Figure 3. SFFS algorithm

Each set were performed Feature Subset Selection. The feature combinations of each set is shown in Table 5.

TABLE 3. Recognition rate in each classifier

	kNN	PNN	GMM	SVM
set A	63%	64%	64%	60%
set B	64%	64%	65%	62%
set C	64%	64%	67%	62%

TABLE 4. the highest recognition rate in GMM with set C

	Joy	Surprise	Anger	Sadness	Neutral	Total
GMM	60%	45%	44%	86%	77%	67%

In K-NN and SVM and GMM, the set B and the set C which are used A-K-D features are a little higher recognition rate than set A which is not used A-K-D features. In addition, a best model is GMM with set C. The Table 4 shows the recognition rate for each emotion in GMM with set C.

VII. Discussion

The set B and the set C which are used A-K-D features showed the higher recognition rate than set A which is not used A-K-D features. Furthermore, it may be said that the acoustic features in A-K-D unit are effective for emotional speech classification, because the acoustic features of each region of Attack and Keep and Decay were selected by Feature Subset Selection in Table 3. In addition, the highest recognition model involved A-K-D features which are estimated by the proposed method. But, because there is not a difference in recognition rate with the classifiers except GMM, it cannot be said that the proposed method is better unconditionally. However, it may be said that the acoustic features of the proposed method is more effective than the conventional method when we use a classifier which can express a feature distribution in detail like GMM.

In this study, an error rate was used as an evaluation function when Feature Subset Selection was performed. But, using the error rate as the evaluation function may be dependent on the classifier and the number of samples of test data or

 ISBN: 978-1-61208-110-6

TABLE 5. Feature combinations for optimal solutions

set A (Whole Speech)
F0 (max,min,range,mean,median), ΔF0 (mean,maxi,mean of pos.incline), Power (min,mean,median), MFCC (7th)
set B (A-K-D unit (Conventional method) + Whole Speech)
Attack F0 (max,mean,median), Keep Power (median), Keep F0 (mean), Decay Power (min), Decay F0 (median), MFCC (1st,2nd), F0 (max,min,mean,median), ΔF0 (mean), Power (range)
set C (A-K-D unit (Proposal method) + Whole Speech)
Attack Power (max,median), Keep F0 (median), Decay Power (max,min,mean,median), Decay length, F0 (max,mean,median), ΔF0 (mean of neg.incline), Power (max,mean), MFCC (1st)

training data. Because the number of samples was uneven with every emotion, there is a possibility that the best Feature Subset Selection was not possible. Therefore, in future, it is necessary to perform the best Feature Subset Selection by using the evaluation function which is difficult to depend on the classifier and the number of samples or equalizing the number of samples.

In addition, Anger is misrecognized a lot by surprise, and surprise is misrecognized a lot by anger. Sadness is got high recognition rate, but it was easy to be misrecognized neutral, and neutral is easy to be misrecognized by sadness. The pair of those emotions resembled in non-language information, and the subjective emotional classification depended on language information are considered as one of the reasons. In future, it is necessary to solve this problem by performing the subjective emotional classification does not depended on language information or speaker oneself.

VIII. Conclusion

This paper focused on A-K-D unit, and we proposed the improved method of A-K-D unit estimation. As a result, acoustic features of the proposed method were more effective than the conventional method when we used for GMM. However, the acoustic features in this study cannot be used to classify all emotions precisely. Therefore, the acoustic features effective for emotional speech classification have to be examined. In addition, because we did not inspect a precision of automatic detection / estimate of the A-K-D unit, it will be necessary to inspect the precision in future. In a related study [8], the Teager energy operator (TEO) is used as an acoustic feature that does not depend on the word length and speaker. By using this operator, the recognition rate may be improved.

And, if this study is used as an application system, it is necessary for a accorded rate in the subjective emotional classification with a speaker or a third person to be inspected. In addition, this paper is not considered a influence of language information in the subjective emotional classification. Therefore, we think that it is necessary to consider how much influence language information has on human subjective emotional classification, by performing the subjective emotional classification for non-language information.

References

[1] N.Amir and R.Cohen, "Characterizing emotion in the soundtrack of an animated film: Credible or incredible?", Affective Computing and Intelligent Interaction, vol.4738/2007, pp.148-158, 2007.

[2] B.Bird (Director), "The Incredibles [motion picture]", United States: Walt Disney Pictures, 2004.

[3] W.J.Yoon and K.S.Park, "A Study of Emotion Recognition and Its Applications", Modeling Decisions for Artificial Intelligence, vol.4617/2007, pp.455-462, 2007.

[4] B.Schuller, G.Rigoll, and M.lang, "Speech emotion recognition combining acoustic features and linguistic information in a hybrid support vector machine - belief network architecture", ICASSP '04, vol.1, pp.I-577-80, 2004.

[5] S.Mitsuyoshi, F.Ren, Y.Takana, and S.Kuroiwa, "NON-VERBAL VOIVE EMOTION ANALYSIS SYSTEM", IJICIC, vol2, pp.819-830, 2006.

[6] C.F.Huang and M.Akagi, "A three-layered model for expressive speech perception", Speech Communication, vol50(10), pp.810-828, 2008.

[7] K.Hideki, Z.Parham, A.D.Cheveign, and R.D.Patterson, "FIXED POINT ANALYSIS OF FREQUENCY TO INSTANTANEOUS FREQUENCY MAPPING FOR ACCURATE ESTMATION OF F0 AND PERIODICITY ", Proc.EUROSPEECH'99, vol.6, pp.2781-2784, 1999.

[8] N.Katsuya et al."Speech Emotion Recognition with the Teager Energy Operator [in Japanese]", IEICE technical report. Speech 105(572), pp.1-6, 2006.

[9] D.Ververidis and C.Kotropoulos, "Emotional Speech Classification Using Gaussian Mixture Models and the Sequential Floating Forward Selection Algorithm", ICME, pp.1500-1503, 2005.

[10] W.Ser, L.Cen, and Z.L.Yu, "A Hybrid PNN-GMM Classification Scheme for Speech Emotion Recognition ", ICPR, pp.1-4, 2008.

[11] C.W. Hsu, C.C.Chang, and C.J.Lin, "A Practical Guide to Support Vector Classification ", 2010.

[12] M.Nishida and T.Kawahara, "Speaker Model Selection Selection Based on the Bayesian Information Criterion Applied to Unsupervised Spervised Speaker Indexing", IEEE TRANSAP, vol.13, no.4, pp.583-592, 2005.

[13] H.Akaike."A new look at the statistical model identification", IEEE Trans. Automatic Control, vol.19, no.6, pp.716-723, 1974.

[14] I.Guyon."Gene Selection for Cancer Classification using Support Vector Machines", Machine Learning, vol.46, pp.389-422, 2002.

[15] A.W.Whitney."A Direct Method of Nonparametric Measurement Selection", IEEE Transactions on Computers, vol.20, pp.1100-1103, 1971.

[16] T.MARILL and D.M.GREEN, "On the effectiveness of receptors in recognition systems", Trans. on IT, vol.9, pp.1-17, 1963.

[17] P.Pudil, J.Novovicova, J. Kittler, "Floating search methods in feature selection", Pattern Recognition Letters, vol.15, pp.1119-1125, 1994.

[18] D.Ververids and C.Kotropoulos, "Fast and accurate sequential floating forward feature selection with the Bayes classifier applied to speech emotion recognition", Signal processing, vol188(12), pp.2956-2970, 2008.

 ISBN: 978-1-61208-110-6