IARIA

CONTENT 2018

The Tenth International Conference on Creative Content Technologies

ISBN: 978-1-61208-611-8

February 18 - 22, 2018

Barcelona, Spain

CONTENT 2018 Editors

Hans-Werner Sehring, Namics, Germany

Pascal Lorenz, University of Haute-Alsace, France

CONTENT 2018

Forward

The Tenth International Conference on Creative Content Technologies (CONTENT 2018), held between February 18 - 22, 2018 - Barcelona, Spain, continued a series of events targeting 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, 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.

The conference had the following tracks:

- Image and graphics
- Web content
- Domains and approaches

Similar to the previous edition, this event attracted excellent contributions and active participation from all over the world. We were very pleased to receive top quality contributions.

We take here the opportunity to warmly thank all the members of the CONTENT 2018 technical program committee, as well as the numerous reviewers. The creation of such a high quality conference program would not have been possible without their involvement. We also kindly thank all the authors that dedicated much of their time and effort to contribute to CONTENT 2018. 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 also gratefully thank the members of the CONTENT 2018 organizing committee for their help in handling the logistics and for their work that made this professional meeting a success.

We hope CONTENT 2018 was a successful international forum for the exchange of ideas and results between academia and industry and to promote further progress in the area of creative content technologies. We also hope that Barcelona provided a pleasant environment during the conference and everyone saved some time for exploring this beautiful city.

CONTENT 2018 Chairs

CONTENT Steering Committee

Raouf Hamzaoui, De Montfort University - Leicester, UK
Dan Tamir, Texas State University, USA
Mu-Chun Su, National Central University, Taiwan
Nadia Magnenat-Thalmann, University of Geneva, Switzerland
Paulo Urbano, Universidade de Lisboa, Portugal
José Fornari, UNICAMP, Brazil

CONTENT 2018 Industry/Research Advisory Committee

Hans-Werner Sehring, Namics, Germany
René Berndt, Fraunhofer Austria Research GmbH, Austria
Daniel Thalmann, Institute for Media Innovation (IMI) - Nanyang Technological University, Singapore
Klaus Drechsler, Fraunhofer-Institute for Computer Graphics Research IGD, Germany
Stefanos Vrochidis, Information Technologies Institute - Centre for Research and Technology Hellas, Greece

CONTENT 2018

Committee

CONTENT Steering Committee

Raouf Hamzaoui, De Montfort University - Leicester, UK
Dan Tamir, Texas State University, USA
Mu-Chun Su, National Central University, Taiwan
Nadia Magnenat-Thalmann, University of Geneva, Switzerland
Paulo Urbano, Universidade de Lisboa, Portugal
José Fornari, UNICAMP, Brazil

CONTENT 2018 Industry/Research Advisory Committee

Hans-Werner Sehring, Namics, Germany
René Berndt, Fraunhofer Austria Research GmbH, Austria
Daniel Thalmann, Institute for Media Innovation (IMI) - Nanyang Technological University, Singapore
Klaus Drechsler, Fraunhofer-Institute for Computer Graphics Research IGD, Germany
Stefanos Vrochidis, Information Technologies Institute - Centre for Research and Technology Hellas, Greece

CONTENT 2018 Technical Program Committee

Jose Alfredo F. Costa, Federal University - UFRN, Brazil
Mostafa Alli, Tsinghua University, China
Leonidas Anthopoulos, University of Applied Science (TEI) of Thessaly, Greece
Konstantinos Avgerinakis, CERTH-ITI, Greece
Kambiz Badie, Research Institute for ICT & University of Tehran, Iran
René Berndt, Fraunhofer Austria Research GmbH
Christos Bouras, University of Patras | Computer Technology Institute & Press <Diophantus> Greece
Marcelo Caetano, INESC TEC, Porto, Portugal
Juan Manuel Corchado Rodríguez, Universidad de Salamanca, Spain
João Correia, University of Coimbra, Portugal
Raffaele de Amicis, Oregon State University, USA
Rafael del Vado Vírseda, Universidad Complutense de Madrid, Spain
Myriam Desainte-Catherine, LaBRI - Université de Bordeaux, France
Klaus Drechsler, Fraunhofer-Institute for Computer Graphics Research IGD, Germany
Joël Dumoulin, HumanTech Institute | University of Applied Sciences of Western Switzerland, Switzerland
Miao Fan, Tsinghua University, China
José Fornari, UNICAMP, Brazil
Alexander Gelbukh, Instituto Politécnico Nacional, Mexico
Afzal Godil, National Institute of Standards and Technology, USA

Seiichi Gohshi, Kogakuin University, Japan
Raouf Hamzaoui, De Montfort University, Leicester, UK
Tzung-Pei Hong, National University of Kaohsiung, Taiwan
Gang Hu, University of Electronic Science and Technology of China, China
Chih-Cheng Hung, Kennesaw State University, USA
Wilawan Inchamnan, Dhurakij Pundit University, Thailand
Pavel Izhutov, Stanford University, USA
Kimmo Kettunen, National Library of Finland | University of Helsinki
Wen-Hsing Lai, National Kaohsiung First University of Science and Technology, Taiwan
Yuhua Lin, Microsoft, USA
Alain Lioret, Paris 8 University, France
Nadia Magnenat-Thalmann, University of Geneva, Switzerland
Prabhat Mahanti, University of New Brunswick, Canada
Maryam Tayefeh Mahmoudi, ICT Research Institute, Iran
Manfred Meyer, Westphalian University of Applied Sciences, Bocholt, Germany
Vasileios Mezaris, Information Technologies Institute (ITI) - Centre for Research and Technology Hellas (CERTH), Greece
Boris Mirkin, National Research University "Higher School of Economics", Russia / University of London, UK
María Navarro Cáceres, University of Salamanca, Spain
Somnuk Phon-Amnuaisuk, Universiti Teknologi Brunei, Brunei
P.Krishna Reddy, International Institute of Information Technology Hyderabad (IIITH) Gachibowli, India
Himangshu Sarma, NIT Sikkim, India
Marco Scirea, IT University of Copenhagen, Denmark
Hans-Werner Sehring, Namics, Germany
Anna Shvets, Maria Curie-Sklodowska University in Lublin, Poland
Mu-Chun Su, National Central University, Taiwan
Atsuhiro Takasu, National Institute of Informatics, Japan
Dan Tamir, Texas State University, USA
Daniel Thalmann, Institute for Media Innovation (IMI) - Nanyang Technological University, Singapore
Božo Tomas, University of Mostar, Bosnia and Herzegovina
Nikita Spirin, University of Illinois at Urbana-Champaign, USA
Paulo Urbano, Universidade de Lisboa, Portugal
Stefanos Vrochidis, Information Technologies Institute - Centre for Research and Technology Hellas, Greece
Krzysztof Walczak, Poznan University of Economics, Poland
Toyohide Watanabe, Nagoya Industrial Science Research Institute, Japan
Yubao Wu, Georgia State University, USA

Copyright Information

Table of Contents

Schemas for Context-aware Content Storage

Hans-Werner Sehring
Namics
Hamburg, Germany
e-mail: hans-werner.sehring@namics.com

Abstract—**Data, to an increasing degree, is not used directly as content represented in documents, but it serves as a foundation for content tailored for and delivered to users working in different and varying contexts. To this end, the actual content is dynamically assembled from base data with respect to a certain context. This is particularly true for content management applications, e.g., for websites that are targeted at a user's context. The notion of context comprises various dimensions of parameters like language, location, time, user, and user's device. Most data modeling languages, including programming languages, are not well prepared to cope with variants of content, though. They are designed to manage universal, consistent, and complete sets of data. The Minimalistic Meta Modeling Language (M3L) as a language for content representation has proven particularly useful for modeling content in context. Towards an operational M3L execution environment, we are researching data schemas to efficiently store and utilize M3L models. Such schemas serve as a testbed to discuss context-aware data representation and retrieval in this paper. This is done by expressing context-aware models, in particular M3L statements, by means of traditional persistence technology.**

Keywords-data modeling; content modeling; context-aware data modeling; content; content management; context.

I. INTRODUCTION

In the digital society [1], data is required to represent all kinds of *content*, ranging from structured content of text documents to unstructured, typically binary representations of video and audio content. It is used for many purposes, the most obvious ones being information and commerce. Content is published by means of documents, often multimedia documents incorporating different media that are interrelated to form hypermedia networks. So-called publication channels offer the medium for one kind of publication, e.g., a website, a document file, or a mobile app. Content is typically represented in a channel-agnostic way in order to support multi- or even omni-channel publishing.

It is quite common to deliver content to users in a way that addresses the *context* in which they are when requesting the content. This may include the channel they are using, the working mode they are in, the history of previous usage scenarios, etc. Targeting content to users' contexts can range from simply arranging content in a specific way, over specifically assembled documents, to content that is synthesized for the current requests. Examples are a prominent display of teasers for content that is assumed to be of interest to the user, the production of documents matching a user's native language, adjustment of document quality based on the current network bandwidth and the receiving device, and creating content that represents some base data in knowledgeable form.

For such content targeting scenarios, data needs to be stored in a way that allows generating different views on the content, mainly by selecting content relevant in a certain context. Data representing all forms of content in such a system, therefore, needs to be attributed with the contexts in which it is applicable or preferred. Obviously, some notion of context is required for such representations [2].

Data modeling and programming languages typically do not exhibit features to represent context and to include it in evaluations. Database management systems, being the backbone of practically every information system, are particularly optimized for one connected set of data that is supposed to be consistent and complete. This means that they are not well equipped for dynamic content production, neither regarding content representation nor efficient context-dependent retrieval.

Data retrieval needs particular attention when content is dynamically assembled depending on some context in which it is requested. For the tasks of context-aware content management, complex collections of data to be used as content are requested frequently. A context-aware schema has to efficiently support the underlying queries that are employed to identify relevant content.

For the discussion of data models, we consider content in contexts as it is expressible using the Minimalistic Meta Modeling Language (M3L). This language allows expressing content in a straightforward way. Being a modeling language, there is no obvious mapping to established data structures, though.

The rest of this paper is organized as follows. Section II reviews related work in the area of context-aware data and content models. Section III gives a brief overview over the M3L and describes those parts of the language that are required for the discussion in this paper. Section IV presents a first conceptual model of an internal representation of M3L concepts. Section V makes this model more concrete by means of logical representations, comparable to the logical view on databases. Aspects of alternative implementations are touched in Section VI. The conclusion and acknowledgment close the paper.

II. Related Work

Context is important in the area of content management, but also other modeling domains. This section names some existing modeling approaches for contextual information.

A. *Content Management Products*

Most commercial content management products have introduced some notion of context in their models and processes. They utilize context information to *target* content to users. Some use the term *personalization*, which is similar to, but different from contextualization [3].

In most cases, there are publication *rules* associated with content, similar as discussed in [4]. These rules are based on so-called *segments*. Every user is assigned one or more segments. When requesting content, the rules are evaluated for the actual segment(s) in order to select suitable content.

Content authors and editors maintain the content rules. Segments are assigned to users automatically by the systems based on the users' behavior (user interactions), the user journey (e.g., previously visited sites and search terms used for finding the current website), and context information (e.g., device used and location of the user).

Segments offer a rather universal notion of context, though there is no explicit context model.

B. *Context-aware Data Models*

Parallel to the notion of context used for content, there exists some work on the influence of environments on running applications. In mobile usage scenarios, context refers mainly to such environmental considerations, e.g., network availability, network bandwidth, device, or location.

Context changes are incorporated dynamically into evaluations in these scenarios [5].

Context-awareness is not limited to data models. It is also used for adaptable or adaptive software systems, e.g., to map software configurations to execution environments [6], or to control the behavior of a generic solution [7].

C. *Concept-oriented Content Management*

Concept-oriented Content Management (CCM) [8] is an approach to manage content reflecting knowledge. Such content does not represent simple facts, but instead is subject to interpretation. Furthermore, the history of things is described by content, not just their latest state.

CCM is not directly concerned about modeling context. Instead, it aims to introduce a form of pragmatics into content modeling that allows users on the one hand to express differing views by means of individual content models, and on the other hand to still communicate by exchanging content between individualized models.

CCM uses a notion of personalization that goes far beyond the one of content management systems (see above).

It is similar to contextualized content usage, although the system does not know about the context of a user. Instead, users carry out personalization (in CCM terms) manually.

A CCM system reacts to model changes and relates model variants to each other. The basis for this is systems generation: based on the definitions of users, schemas, APIs, and software modules are generated.

Some aspects of the considerations presented in Section VI were gained from the research on the generation of CCM modules for persistence.

III. The Minimalistic Meta Modeling Language

The *Minimalistic Meta Modeling Language* (*M3L*, pronounced "mel") is a modeling language that is applicable to a range of modeling tasks. It proved particularly useful for context-aware content modeling [9].

For the purpose of this paper, we only introduce the static aspects of the M3L in this section. Dynamic evaluations that are defined by means of different rules are not presented here because – at least in the current state of investigation – they lay outside the scope of content models.

The descriptive power of M3L lies in the fact that the formal semantics is rather abstract. There is no fixed domain semantics connected to M3L definitions. There is also no formal distinction between typical conceptual relationships (specialization, instantiation, entity-attribute, aggregation, materialization, contextualization, etc.).

A. *Concept Definitions and References*

A M3L definition consists of a series of definitions or references. Each definition starts with a previously unused identifier that is introduced by the definition and may end with a semicolon, e.g.:

```
Person;
```

A reference has the same syntax, but it names an identifier that has already been introduced.

We call the entity named by such an identifier a *concept*.

The keyword `is` introduces an optional reference to a *base concept*, making the newly defined concept a *refinement* of it.

A specialization relationship as known from object-oriented modeling is established between the base concept and the newly defined derived concept. This relationship leads to the concepts defined in the context (see below) of the base concept to be visible in the derived concept.

The keyword `is` always has to be followed by either `a`, `an`, or `the`. The keywords `a` and `an` are synonyms for indicating that a classification allows multiple sub concepts of the base concept:

```
Peter is a Person; John is a Person;
```

There may be more than one base concept. Base concepts can be enumerated in a comma-separated list:

```
PeterTheEmployee is a Person, an Employee;
```

The keyword `the` indicates a closed refinement: there may be only one refinement of the base concept (the currently defined one), e.g.:

```
Peter is the FatherOfJohn;
```

Any further refinement of the base concept(s) leads to the redefinition ("unbinding") of the existing refinements.

Statements about already existing concepts lead to their redefinition. For example, the following expressions define the concept `Peter` in a way equivalent to the above variant:

```
Peter is a Person; Peter is an Employee;
```

B. *Content and Context Definitions*

Concept definitions as introduced in the preceding section are valid in a context. Definitions like the ones seen so far add concepts the topmost of a tree of contexts. Curly brackets open a new context, e.g.:

```
Person { Name is a String; }
Peter is a Person{"Peter Smith" is the Name;}
Employee { Salary is a Number; }
Programmer is an Employee;
PeterTheEmployee is a Peter, a Programmer {
 30000 is the Salary; }
```

We call the outer concepts the *context* of the inner, and we call the set of inner concepts the *content* of the outer.

In this example, we assume that concepts `String` and `Number` are already defined. The subconcepts created in context are unique specializations in that context only.

As indicated above, concepts from the context of a concept are inherited by refinements. For example, `Peter` inherits the concept `Name` from `Person`.

M3L has visibility rules that correlate to both contexts and refinements. Each context defines a scope in which defined identifiers are valid. Concepts from outer contexts are visible in inner scopes. For example, in the above example the concept `String` is visible in `Person` because it is defined in the topmost scope. `Salary` is visible in `PeterTheEmployee` because it is defined in `Employee` and the context is inherited. `Salary` is not valid in the topmost context and in `Peter`.

C. *Contextual Amendments*

Concepts can be redefined in contexts. Implicitly, this happens by definitions as those shown above. For example, in the context of `Peter`, the concept `Name` receives a new refinement.

Concepts can be redefined in a context explicitly:

```
AlternateWorld {
 Peter is an Ape {
  "Peter Miller" is the Name; } }
```

We call a redefinition performed in a context different from that of the original definition a *conceptual amendment*.

In the above example, `Peter` is both a `Person` (inherited) and an `Ape` (additionally defined), while the name has been changed.

A redefinition is valid in the context it is defined in, in sub contexts, and in the context of refinements of the context (since the redefinition as part of the content is inherited).

IV. A CONCEPTUAL MODEL FOR CONTENT REPRESENTATIONS

A conceptual model, as known from database modeling, serves as a first step towards data models for context-aware content. The notion of "concept" is ambiguous here: The aim is a model of (M3L) concepts. A conceptual model for this allows us to abstract from the M3L as a language. The model is not supposed to address practical properties such as operational complexity.

A set of M3L concept definitions can be viewed as a graph with each node representing a concept, labeled with the name of the concept. There are two kinds of edges to represent specialization and contextualization. In fact, such a graph forms a hypergraph to account for contextualization. Every node can contain a graph reflecting definitions as the concept's content.

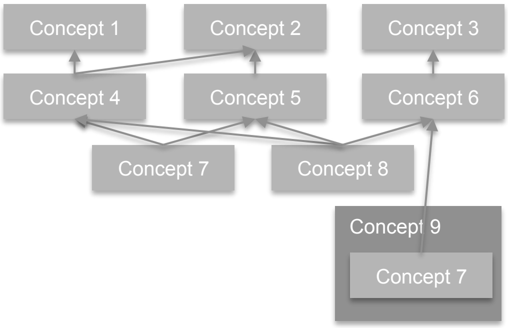


Figure 1. M3L concept refinements and contexts.

The following subsections detail specialization and contextualization relationships, as well as contextual redefinitions.

A. *Representing Specialization*

Conceptually, a specialization/generalization relationship can straightforward be seen as a many-to-many relationship between concepts. Fig. 1 shows an example.

Arrows with filled heads, directed from a concept to its base concepts, represent specialization relationships in the figure. For example, *Concept 4* is a refinement of *Concept 1* and *Concept 2*.

Fig. 1 furthermore indicates an amendment in a context, namely *Concept 9*. While *Concept 7* is a refinement of *Concept 4* and *Concept 5* in the default context, it is additionally a refinement von Concept 6 in the context of Concept 9 (if it is an `is a`/`is an` definition; otherwise, *Concept 7* would only be a refinement of *Concept 6* in the context of *Concept 9*).

B. *Representing Context*

Since contexts form a hierarchy, contextualization can be represented by a one-to-many relationship between concepts in the roles of context and content.

Fig. 2 represents such a hierarchy by nesting boxes shown for concepts. The contextualization relationship is thus visually represented by containment. For example, *Concept 2* is part of the content of *Concept 1*, or *Concept 2* is defined in the context of *Concept 1*.

The outermost context is the default context. There is no corresponding concept for this context.

C. *Representing Contextual Information*

Specialization and contextualization act together. Refinements of a concept inherit its content; concepts from that content are valid also in the context of the refinement. Each context allows concept amendments. These are a second way to add variations of concepts.

In order to represent contextualized redefinitions, we introduce two kinds of context definitions: *Initial Concept Definition* and *Contextual Concept Amendment*. Both can be placed in any context.

An initial context definition is placed in the topmost context in which a concept is defined. Redefinitions of concepts are represented by concept amendments inside the concept in whose context the redefinition is performed.

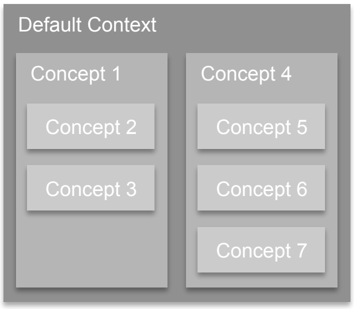


Figure 2. M3L concept definitions in contexts.

Fig. 3 illustrates this. As before, contexts are depicted as nested boxes. There is one *Context* and a *Sub Context*. Both show a *Concept* that has originally been defined as a refinement of *Base Concept* and is itself refined to *Refinement*. In the context on *Sub Context*, the concept gets the additional base concept *Base Concept 2*, and there is another refinement *Refinement 2*. These additions are recorded in the amendment in *Sub Context*.

Amendments have a reference to the next higher definition. This reference is called *Original*. In Fig. 3, it is shown by the dotted line.

Traversal of the original references allows collecting all definitions in order to determine the effective definition.

V. Logical Content Representation

This section refines contextual content representation models to a level similar to that of a logical data model. This way it discusses properties of data representations without taking implementation details into account.

The complexity of lookups is of major importance for the schema design. During the evaluation of M3L statements, many graph traversals are required to find all valid contexts, all base concepts (to determine content sets) and all refinements (to narrow down concepts before applying rules; this evaluation process is not laid out in this paper).

The most important design decision is the degree of (de)normalization of the schema. The basic assumption is that content is mainly queried, so that creation and update cost is less important than lookup cost.

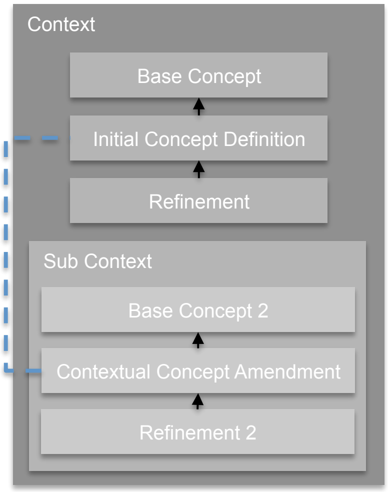


Figure 3. M3L concept amendments in contexts.

We consider two designs of denormalized schemas: materialization of reference sets and storage of relationships in way that allows efficient queries. Efficient storage is based on the usage of numeric IDs to reference concepts and computing relationships based on ID sets. An example of such an approach is the BIRD numbering scheme for trees [10] that allows range queries to determine sub trees.

A. Storing Refinements

Compared to the straightforward conceptual model, the logical schema is denormalized in order to avoid repeated navigation of specialization relationships when collecting the set of (transitive) base concepts or refinements of a concept.

Two approaches are investigated: aggregated concepts and transitive refinement relationships.

Aggregated data collects necessary information to avoid nested queries for refinements. All base concepts and all refinements are stored in an object representing the concept definition. Context-dependent content is added in contextual concept amendments (s.a.) that are stored as part of the context hierarchy.

The description objects additionally reference each other via original references.

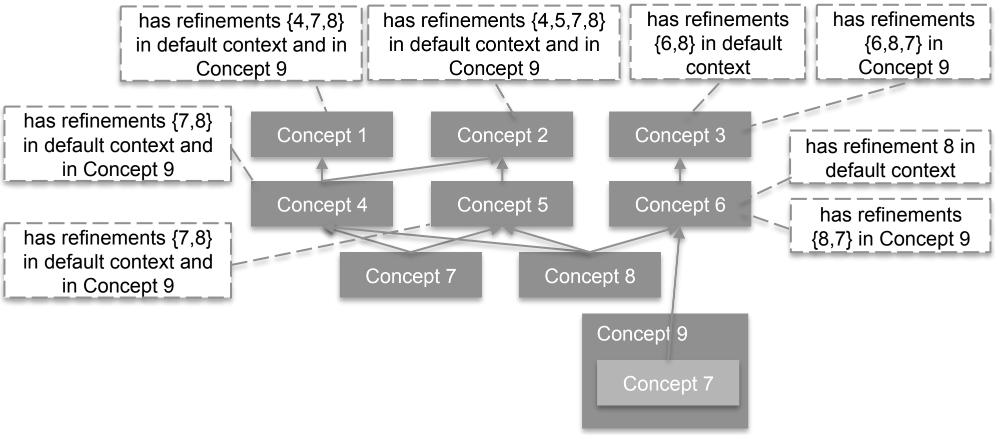

has refinements {4,7,8} in default context and in Concept 9
has refinements {4,5,7,8} in default context and in Concept 9
has refinements {6,8} in default context
has refinements {6,8,7} in Concept 9
has refinements {7,8} in default context and in Concept 9
Concept 1
Concept 2
Concept 3
has refinement 8 in default context
Concept 4
Concept 5
Concept 6
has refinements {8,7} in Concept 9
has refinements {7,8} in default context and in Concept 9
Concept 7
Concept 8
Concept 9
Concept 7

Figure 4. Representation of refinements using materialized transitive refinement relationships.

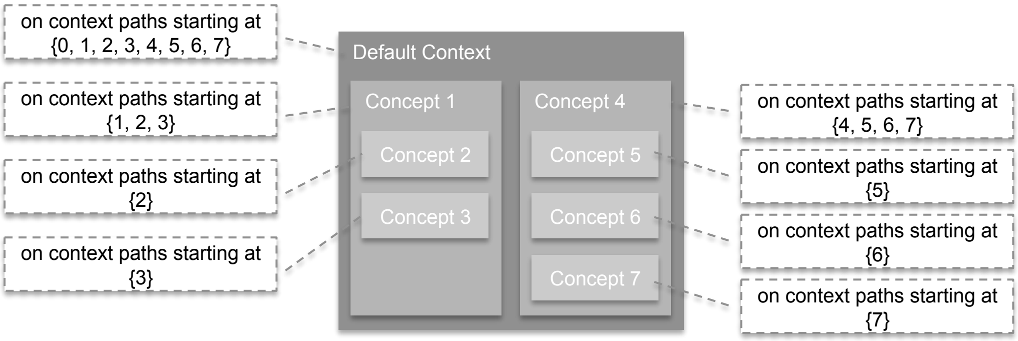

on context paths starting at {0, 1, 2, 3, 4, 5, 6, 7}
on context paths starting at {1, 2, 3}
on context paths starting at {2}
on context paths starting at {3}
Default Context
Concept 1
Concept 2
Concept 3
Concept 4
Concept 5
Concept 6
Concept 7
on context paths starting at {4, 5, 6, 7}
on context paths starting at {5}
on context paths starting at {6}
on context paths starting at {7}

Figure 5. Representation of context hierarchies my materializing paths.

Alternatively, just transitive refinement relationships are materialized for every concept in every context. This way, transitive refinements are directly available, and base concepts can be collected using a simple query.

Fig. 4 shows an example for the sample from Fig. 1. The dashed boxes show the transitive refinements per relevant context. Base concepts can be determined by queries.

For example, the (transitive) base concepts of *Concept 4* are those concepts that have this concept as a refinement. Specifically, these are *Concept 1* and *Concept 2* (in both the default context and in the context of *Concept 9*).

Storing the context together with the refinement relationships is vital for handling singleton (`is the`) relationships, in particular the unbinding of concepts.

B. Storing Context Hierarchies

Performance is particularly important for the retrieval of the hierarchy of contexts a concept is defined or amended in. The effective definition of a concept (including aggregated base concepts and content) relies on this concept hierarchy.

By blending in the context information into the transitive refinements, as shown in the previous subsection, the situation is leveraged to a large degree. Still, the content that a concept has in a certain context is also relevant to concept evaluations.

As for the specialization/generalization relationships, two approaches are discussed here: materialized content collections in all contexts and information about paths in the context hierarchy.

The materialization of contextual definitions works the same way as that of refinements: with every concept definition amendment, we store the effective content in the respective context. This has to be computed on definition.

For the second approach, Fig. 5 illustrates the attribution of paths to the schematic example of Fig. 2. For each concept, we note down the concepts lying on the path in the context hierarchy from the default context to a specific context. For example, *Concept 1* lies on the paths from the default context to itself, to *Concept 2*, and to *Concept 3*.

We used numeric IDs to reference the concept (with the ID 0 given to the pseudo-concept for the default concept). IDs have to be ordered from the default context to sub contexts. By querying for all concepts on the path of a concept, ordered by ID, we retrieve the path to that concept.

VI. Physical Content Storage Models

This section briefly discusses some implementation approaches of context-aware content models. Specifically, we present the basics of a mapping to relational databases and one to a document-oriented database.

A. Mapping M3L to a Relational Database

There is a range of approaches for storing trees and graphs in relational databases [11]. On the basis of these, we add materialized transitive relationships as described above.

Relational tables for the transitive context hierarchy can be defined by statements like (with numeric type INT):

```
CREATE TABLE concept (id INT PRIMARY KEY);
CREATE TABLE paths (
 concept_id       INT REFERENCES concept(id),
 terminal_concept INT REFERENCES concept(id),
 PRIMARY KEY (concept_id, terminal_concept)
);
```

The table *concept* holds concepts (both initial definitions and amendments) with artificial, numeric IDs (other data is omitted here). The second table holds the path information as indicated in Fig. 5. *concept_id* refers to the concept, *terminal_concept* refers to the concept on whose path the concept lies.

Data stored this way can be queried by, e.g.,

```
SELECT c.* FROM concept c, paths p
 WHERE c.id = p.concept_id
 AND p.terminal_concept = i
 ORDER BY p.concept_id DESC;
```

to retrieve the path to concept *i*.

Transitive refinements can be stored in a table:

```
CREATE TABLE transitive_refinements (
 base_concept_id INT REFERENCES concept(id),
 refinement_id   INT REFERENCES concept(id),
 context_id      INT REFERENCES concept(id),
 PRIMARY KEY (base_concept_id, refinement_id,
             context_id));
```

The base concepts of, e.g., *Concept 4* can be queried by:

```
SELECT base_concept_id
 FROM transitive_refinements
 WHERE refinement_id = 4 AND context_id = 0;
```

in the default context (with ID 0), or by:

```
SELECT base_concept_id
 FROM transitive_refinements
 WHERE refinement_id = 4 AND context_id = 9;
```

for the context of *Concept 9*.

```
db.concept.insert({ name: "Default Context", content: [
 { name: "Concept 1", baseConcepts: null,                                    content: null },
 { name: "Concept 2", baseConcepts: null,                                    content: null },
 { name: "Concept 3", baseConcepts: null,                                    content: null },
 { name: "Concept 4", baseConcepts: ["Concept 1", "Concept 2"],              content: null },
 { name: "Concept 5", baseConcepts: ["Concept 2"],                           content: null },
 { name: "Concept 6", baseConcepts: ["Concept 3"],                           content: null },
 { name: "Concept 7", baseConcepts: ["Concept 4", "Concept 5"],              content: null },
 { name: "Concept 8", baseConcepts: ["Concept 4", "Concept 5", "Concept 6"], content: null },
 { name: "Concept 9", baseConcepts: null,                                    content: [
  {name: "Concept 7", baseConcepts: ["Concept 4", "Concept 5", "Concept 6"], content: null,
   original: "Concept 7" } ] } ] })
db.concept.aggregate([
{$unwind:"$content"},{$replaceRoot:{newRoot:"$content"}},{$match:{name:"Concept 9"}},
{$unwind:"$content"},{$replaceRoot:{newRoot:"$content"}},{$match:{baseConcepts:"Concept 6"}}])
```

Figure 6. Document definitions to map M3L to MongoDB and a sample query.

B. Mapping M3L to a Document Database

As an example of so-called NoSQL approaches, we conduct ongoing experiments with MongoDB, a widely used document-oriented database management system.

The definition of concept relationships is done a similar way as in relational databases: records have IDs, and records store IDs for references. There are no distinct relation structures, though. References are stored as document fields.

In contrast to a purely relational structure, documents allow representing nested contexts in a natural manner by embedded documents.

As an example of a schema, the *insert* statement shown in Fig. 6 stores the whole graph of Fig. 1.

This structure can be queried as required. For example, to find concepts with base concept *Concept 6* in the context of *Concept 9*, the *aggregate* statement in Fig. 6 can be applied.

VII. Conclusion

This section sums up the paper and gives an outlook on future work.

A. Summary

In this paper, we laid out approaches to context-aware content management, in particular using the Minimalistic Meta Modeling Language (M3L).

Though it is easily possible to map context representations to existing data management approaches, care has to be taken to achieve efficient implementations.

A logical schema for the representation of contextual content is presented, and first implementations are conducted. These demonstrate the feasibility of the schemas.

B. Outlook

The work on the data model mappings for M3L concept definitions is ongoing work; there is ample room for further optimizations of the relational database schema. The mapping to document-oriented database needs much more elaboration before comparisons can be made.

The utilization of databases to support M3L concept evaluation is an open issue. Practical rule sets will guide the investigations in the future.

Experiments with different implementations are ongoing. Data models have yet to be rated based on practical results.

Acknowledgment

Though the ideas presented in this paper are in no way related to Namics, the author is thankful to his employer for letting him follow his research ambitions based on experience made in customer projects. The discussions with colleagues, partners, and customers are highly appreciated.

References

[1] M. Gutmann, “Information Technology and Society,” Swiss Federal Institute of Technology Zurich / Ecole Centrale de Paris, 2001.

[2] C. Bolchini, C. A. Curino, E. Quintarelli, F. A. Schreiber, and L. Tanca, “A Data-oriented Survey of Context Models,” ACM SIGMOD Record, vol. 36, pp. 19-26, December 2007.

[3] A. Zimmermann, M. Specht, and A. Lorenz, “Personalization and Context Management,” User Modeling and User-Adapted Interaction, vol. 15, pp. 275-302, Aug. 2005.

[4] S. Trullemans, L. Van Holsbeeke, and B. Signer, “The Context Modelling Toolkit: A Unified Multi-layered Context Modelling Approach,” Proc. ACM Human-Computer Interaction (PACMHCI), vol. 1, June 2017, pp. 7:1-7:16.

[5] G. Orsi and L. Tanca, “Context Modelling and Context-Aware Querying (Can Datalog Be of Help?),” Proc. First International Conference on Datalog Reloaded (Datalog '10), Mar. 2010, pp. 225-244.

[6] D. Ayed, C. Taconet, and G. Bernard, “A Data Model for Context-aware Deployment of Component-based Applications onto Distributed Systems,” GET/INT, 2004.

[7] S. Vaupel, D. Wlochowitz, and G. Taentzer, “A Generic Architecture Supporting Context-Aware Data and Transaction Management for Mobile Applications”, Proc. International Conference on Mobile Software Engineering and Systems (MOBILESoft '16), May 2016, pp. 111-122.

[8] J. W. Schmidt and H.-W. Sehring, “Conceptual Content Modeling and Management,” Perspectives of System Informatics, vol. 2890, M. Broy and A.V. Zamulin, Eds. Springer-Verlag, pp. 469-493, 2003.

[9] H.-W. Sehring, “Content Modeling Based on Concepts in Contexts,” Proc. Third Int. Conference on Creative Content Technologies (CONTENT 2011), pp. 18-23, Sep. 2011.

[10] F. Weigel, K. U. Schulz, and H. Meuss, “The BIRD Numbering Scheme for XML and Tree Databases – Deciding and Reconstructing Tree Relations using Efficient Arithmetic Operations,” Proc. Third international conference on Database and XML Technologies (XSym'05), Aug. 2005, pp. 49-67.

[11] V. Tropashko, SQL Design Patterns: The Expert Guide to SQL Programming. Rampant Techpress, 2006.

Social Media Analytics in Support of Documentary Production

Giorgos Mitsis, Nikos Kalatzis, Ioanna Roussaki,
Eirini Eleni Tsiropoulou, Symeon Papavassiliou
Institute of Communications and Computer Systems
Athens, Greece
e-mails: {gmitsis@netmode, nikosk@cn,
ioanna.roussaki@cn, etsirop@netmode,
papavass@mail}.ntua.gr

Simona Tonoli
Mediaset
Milan, Italy
e-mail: Simona.Tonoli@mediaset.it

Abstract—**Recent market research has revealed a globally growing interest on documentaries that have now become one of the most populated content-wise genre in the movie titles catalog, surpassing traditionally popular genres such as comedy or adventure films. At the same time, modern audiences appear willing to immerse into more interactive and personalized viewing experiences. Documentaries, even in their linear version, involve high costs in all phases (pre-production, production, post-production) due to various inefficiencies, partly attributed to the lack of scientifically-proven cost-effective Information and Communications Technology (ICT) tools. To fill this gap, a set of innovative ICT tools is delivered that focus on supporting all stages of the documentary creation process, ranging from the documentary topic selection to its final delivery to the viewers. This paper provides an overview of the respective tools, elaborating on two specific tools that primarily focus on the interests and satisfaction of the targeted audience: the Integrated Trends Discovery tool and the Social Recommendation & Personalization tool, elaborating on their design, functionality and performance, and concludes with exposing the future plans and potential regarding these tools.**

Keywords-documentary production; social-media analytics; Integrated Trends Discovery tool; Social Recommendation & Personalization tool.

I. INTRODUCTION

From the earliest days of cinema, documentaries have provided a powerful way of engaging audiences with the world. They always had social and market impact, as they adapted to the available means of production and distribution. More than any other type of films, documentarians were avid adapters of new technologies, which periodically revitalized the classical documentary form. The documentary is a genre which lends itself straightforwardly to interaction. People have different knowledge backgrounds, different interests and points of view, different aesthetic tastes and different constraints while viewing a programme. Therefore, it becomes evident that some form of personalized interactive documentary creation will enhance the quality of experience for the viewers, facilitating them to choose different paths primarily with respect to the documentary format and playout system. The convergence between the documentary production field and of digital media enables the realization of this vision.

As the range of ICT platforms broadens, documentary makers need to understand and adopt emerging technologies in order to ensure audience engagement and creative satisfaction, via the use of personalization and interactive media. One of the major challenges for stakeholders in the arena of documentary creation is the development of processes and business models to exploit the advantages of those technical achievements, in order to reduce the overall cost of documentary end-to-end production, to save time and to deliver enhanced personalized interactive and thus more attractive documentaries to the viewers.

PRODUCER [1] is an H2020 EU project that aims to pave the path towards supporting the transformation of the well-established and successful traditional models of linear documentaries to interactive documentaries, by responding to the recent challenges of the convergence of interactive media and documentaries. This is achieved via the creation of a set of enhanced ICT tools that focus on supporting all documentary creation phases, ranging from the user engagement and audience building, to the final documentary delivery. In addition to directly reducing the overall production cost and time, PRODUCER aims to enhance viewers' experience and satisfaction by generating multi-layered documentaries and delivering more personalized services, e.g., regarding the documentary format and playout.

In order to provide the aforementioned functionality, the PRODUCER platform implements 9 tools, each focusing on a specific documentary production phase. These tools are: Integrated trends discovery tool, Audience building tool and Open content discovery tool (that support the documentary pre-production phase), Multimedia content storage, search & retrieval tool and Automatic annotation tool (that support the core production phase), Interactive-enriched video creation tool, 360° video playout tool, Second screen interaction tool and Social recommendation & personalization tool (all four focusing on the documentary post-production phase). The architecture of the PRODUCER platform is presented in more detail in [2]. This paper elaborates on two of the PRODUCER tools: the Integrated Trends Discovery tool and the Social Recommendation & Personalization tool.

In the rest of the paper, Section 2 elaborates on the design & functionality of the Integrated Trends Discovery tool, presenting initial evaluation results for one of its mechanisms. Section 3 focuses on the description of the Social Recommendation & Personalization tool, while it elaborates on specific performance evaluation/benchmarking results related to its functionality. Finally, in Section 4, conclusions are drawn and future plans are presented.

II. INTEGRATED TRENDS DISCOVERY TOOL

This section elaborates on the ITD tool, i.e., its innovations, architecture, user demographics inference mechanism and respective evaluation.

A. Rationale and Innovations

In recent years, there is an increasing trend on utilizing social media analytics and Internet search engines analytics for studying and predicting behavior of people with regards various societal activities. The proper analysis of Web 2.0 services utilization, goes beyond the standard surveys or focus groups and has the potential to be a valuable source of information leveraging internet users as the largest panel of users in the world. Analysts from a wide area of research fields have the ability to reveal current and historic interests of individuals and to extract additional information about their demographics, behavior, preferences, etc. One of the c aspects of this approach is that the user base consists of people that the researchers have never considered.

Some of the research fields that demonstrate significant results through the utilization of such analytics include epidemiology (e.g., detect influenza [3][4] and malaria [5]) epidemics), economy (e.g., stock market analysis [6], private consumption prediction [7], financial market analysis and prediction [8], unemployment rate estimation [9]) politics (e.g., predicting elections outcomes [10]).

On the other hand, there are limitations on relying only on these information sources as certain groups of users might be over- or under-represented among internet search data. There is a significant variability of online access and internet search usage across different demographic, socioeconomic, and geographic subpopulations.

With regards content creation and marketing, the existing methodologies are under a major and rapid transformation given the proliferation of Social Media and search engines. The utilization of such services generates voluminous data that allows the extraction of new insights with regards the audiences' behavioral dynamics. In [11], authors propose a mechanism for predicting the popularity of online content by analyzing activity of self-organized groups of users in social networks. Authors in [12] attempt to predict IMDB (http://www.imdb.com/) movie ratings using Google search frequencies for movie related information. In a similar manner, authors in [13] are inferring, based on social media analytics, the potential box office revenues with regards Internet content generated about Bollywood movies.

The existing research approaches are mainly focusing in post-production phase of released content. Identifying the topics that are most likely to engage the audience is critical for content creation in the pre-production phase. The ultimate goal of content production houses is to deliver content that matches exactly what people are looking for. Deciding wisely on the main documentary topic, as well as the additional elements that will be elaborated upon, prior to engaging any resources in the documentary production process, has the potential to reduce the overall cost and duration of the production lifecycle, as well as to increase the population of the audiences interested, thus boosting the respective revenues. In addition, the existence of hard evidence with regards potential audience's volume and characteristics (e.g., geographical regions, gender, age) is an important parameter in order to decide the amount of effort and budget to be invested during production.

There are various social media analytics tools that are focusing on generic marketing analysis e.g., monitoring for a long time specific keyword(s) and websites for promoting a specific brand and engaging potential customers. These web marketing tools rely on user tracking, consideration of user journeys, detection of conversion blockers, user segmentation, etc. This kind of analysis requires access to specific websites analytics and connections with social media accounts (e.g., friends, followers) which is not the case when the aim is to extract the generic population trends. In addition, these services are available under subscription fee that typically ranges from 100 Euros/month to several thousand Euros/month, a cost that might be difficult to be handled by small documentary houses.

The ITD Tool aims to support the formulation, validation and (re)orientation of documentary production ideas and estimate how appealing these ideas will be to potential audiences based on data coming from global communication media with massive user numbers. The ITD tool integrates existing popular publicly available services for: monitoring search trends (e.g., Google Trends), researching keywords (e.g., Google Adwords Keyword Planner), analyzing social media trends (e.g., Twitter trending hashtags). In more details, the ITD tool innovations include the following:

- Identification and evaluation of audience's generic interest for specific topics and analysis/inference of audience's characteristics (e.g., demographics, location)
- Extraction of additional aspects of a topic though keyword analysis, quantitate correlation of keywords, and association with high level knowledge (e.g., audience sentiment analysis)
- Discovery and identification of specific real life events related with the investigated topic (e.g., various breakthroughs of google/twitter trending terms are associated with specific incidents)
- Utilisation of data sources that are mainly openly accessible through public APIs which minimises the cost and increases the user base.

B. Architecture

A functional view of ITD tool's architecture is provided in Figure 1. Its core modules are described hereafter.

RestAPI: This component exposes the backend's functionality through a REST endpoint. The API specifies a set of trend discovery queries where the service consumer provides as input various criteria such as keywords, topics, geographical regions, time periods, etc.

Trends Query Management: This component orchestrates the overall execution of the queries and the processing of the replies. It produces several queries formulated properly that are forwarded to the respective connectors/wrappers to dispatch the requests to several existing TD tools/services available online. Given that each external service will reply in different time frames (e.g., a call to Google Trends discovery replies within a few seconds while Twitter stream analysis might take longer time periods) the overall process is performed in an asynchronous manner, coordinated by the Message Broker. The Query Management enforces querying policies tailored to each service in order to optimize the utilization of the services and to avoid potential bans. To this end, results from calls are also stored in ITD tool's local database in order to avoid unnecessary calls to the external APIs that have recently performed.

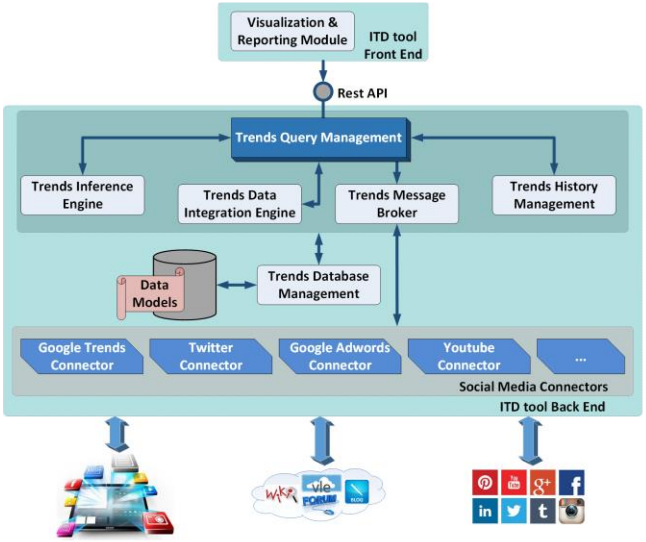


Figure 1. Architecture of the Integrated Trends Discovery Tool.

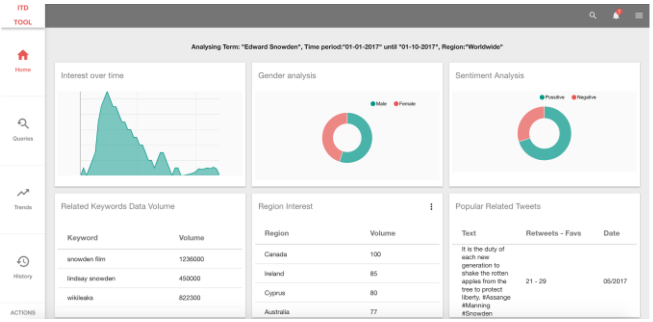
Trends Message Broker: This component realizes the asynchronous handling of requests. It is essentially a messaging server that forwards requests to the appropriate recipients via a job queue based on distributed message passing system.

Social Media Connectors: A set of software modules that support the connection and the execution of queries to external services through the provided available APIs. Connectors are embedding all the necessary security related credentials to the calls and automate the initiation of a session with the external services. Thus, the connectors automate and ease the actual formulation and execution of the queries issued by the Query Management component. Some example APIs that are utilized by the connectors are: Google Adword API, Twitter API, YouTube Data API v3.

Trends Data Integration Engine: This module collects the intermediate and final results from all modules, homogenize their different formats, and extracts the final report with regards the trends discovery process. The results are also modelled and stored in the local data base in order to be available for future utilization.

Trends Database Management & Data model: The ITD tool maintains a local database where the results of various calls to external services are stored. The Database Management module supports the creation, retrieval, update and deletion of data objects. This functionality is supported for both contemporary data but also for historic results (Trends History Management). Hence, it is feasible for the user to compare trend discovery reports performed in the past with more recent ones and have an intuitive view of the evolution of trend reports in time.

Front End: The Front-End visualizes the results providing the following output: (i) a graph of terms (each term is escorted by an audience popularity metric and is correlated with other terms, where a metric defines the correlation level), (ii) audience interest per location (country/city), (iii) interest per date(s) (significant dates, identification of seasonal habits), (iv) audiences sentiment analysis, (v) audiences gender analysis (vi) related questions with the topic. An ITD GUI snapshot is depicted in Figure 2.



Analysing Term: "Edward Snowden", Time period:"01-01-2017" until "01-10-2017", Region:"Worldwide"

Related Keywords Data Volume

Keyword	Volume
snowden film	1236000
lindsay snowden	450000
wikileaks	822300

Region Interest

Region	Volume
Canada	100
Ireland	85
Cyprus	80
Australia	77

Popular Related Tweets

Text	Retweets - Favs	Date
It is the duty of each new generation to shake the rotten apples from the tree to protect liberty. #Assange #Manning #Snowden	21 - 29	05/2017

Figure 2. Snapshot of the Integrated Trends Discovery Tool GUI.

Trends Inference Engine: In some cases, the external services are not directly providing all information aspects of the required discovery process. To this end, by applying the appropriate inference mechanisms on the available data allows the extraction of additional information escorted by a confidence level with regards the accuracy of the estimation. Details of this module are presented in the following section.

C. *Inference of User Demographics*

During the preproduction phase of a documentary, producers are highly interested in estimating trends in correlation with potential audiences' gender and age classification. This kind of information is not freely available from social media services due to user privacy protection data policies. There are various state of the art attempts that focus on inferring user demographics though probabilistic approaches based on user related data freely available on social media (e.g., tweets content, linguistic features, followers' profile) [14][15][16][17].

With regards to the documentary preproduction phase, the task of age and gender estimation is tackled by the ITD tool via the utilization of classification algorithms trained with ground-truth data sets of a number of tweeter users. Twitter service proved to be the most proper for extracting user profile information as Twitter account data and content are openly available. The trained network is then utilized in order to generalize the training process and estimate missing information from wider networks of twitter users.

The inference process is coordinated by the Trends Inference Engine. The engine uses the TwitterAPI to retrieve tweets where the keywords connected with certain topics are mentioned. Based on the respective Twitter Account ids, profile information is collected for each account. Based on profile attributes (e.g., "name", "screen_name", "profile photo", "short description", "profile_color") each user is classified to age & gender category and each classification is escorted by a confidence level.

The actual classification process is based on a statistical model where recurring patterns of users' profile attributes are accompanying a certain age and/or gender class. Learning is

ISBN: 978-1-61208-611-8

performed based on a ground truth dataset containing records of real Twitter profile information and the respective gender/ age. The ITD tool is capable to utilise various classification algorithms but as a first proof of concept the Naive Bayes is evaluated. Naive Bayes (NB) is an algorithm that fulfills the requirements set by similar problems and has performed well in many complex real world situations [18]. NB follows a supervised learning approach for estimating parameters of the classifier, such as means and variances of the variables. The algorithm provides quantifiable probability distributions for each possible class and requires a small amount of training data. In addition, NB can handle both categorical and numerical attributes. Compared with Bayesian Networks, there is no need for domain expert interference in designing dependencies between input attributes. On the other hand, it assumes that attributes are independent from each other with respect to the classification outcome, something that it is not always the case, while the computing resource consumption can get significantly high.

A user's profile is modelled as s $= \{c_1, c_2, \dots, c_n\}$, where c_i is the value of user profile information of type $i, (i = 1,2,\dots,n)$. Gender classes are modelled as g_j (j =1,2,3) corresponding to: “Female”, “Male” and “Unknown”. Age classes are modelled as a_i (i=1,…,7) corresponding to the following 7 age states: 18-24, 25-34, 35-44, 45-54, 55-64, 65 or more, and Unknown.

Based on the ground truth dataset age and gender classes can be associated with specific user profiles in the form of tuples such as (gender, profile) => (g_j, s) and (age, profile) => (a_i, s). Bayes rule for calculating prediction probabilities according to the defined problem becomes:

$$P\big[g_j|s\big] = P\big[g_j\big] \times \frac{\prod_{j=1}^{n} P[c_j|g_j]}{P(s)}$$

where g_j is the expected gender classification outcome and $s = \{c_j\},\ j = 1, \dots, n$ is the current evidence input.

Similarly, Bayes rule for estimating the user's age is:

$$P\big[a_j|s\big] = P\big[a_j\big] \times \frac{\prod_{j=1}^{n} P[c_j|a_j]}{P(s)}$$

Based on these rules the actual estimation is realised through the maximisation of these probabilities: $a = \arg\max\{P\big[a_j|s\big]\}$ and $g = \arg\max\{P\big[g_j|s\big]\}$.

D. Evaluation

The presented architecture is under implementation by the authors of this article and a first release is already available at: http://itd.lab.netmode.ece.ntua.gr/. The ITD backend tool is developed in Django-Python framework, the front-end is based on Angular-Material while the following services have been integrated through the respective APIs: Google Trends, Google Adwords, Twitter, Youtube. The first evaluation processes with regards the overall utilization of the tool are encouraging and allow to discover early in the development phase potential shortcomings.

Such an issue is related with the volume of calls to external services. For example, Twitter API limits the allowed calls to 15 every 15 minutes per service consumer. As this issue was expected, a caching mechanism is utilized where results from each call to the Twitter API are also stored in the local database. Hence the ITD builds each own information store in order to avoid unnecessary calls. To this end, as the tool is utilized from various user the local information store is getting more complete.

With regards the ITD inference engine a first evaluation attempt realized for the gender estimation mechanism. The evaluation has been based on a public data set (https://www.kaggle.com/crowdflower/twitter-user-gender-classification) of ground truth data containing information of 10021 twitter users' profiles. The dataset contains the gender of distinct twitter users escorted by profile information. As a first step on the evaluation process and given that stylistic factors are often associated with user gender, the Twitter profile colour has been initially utilized.

Each colour's RGB value (red, green, blue) is fed to the Bayesian classifier as a distinct numerical feature. Thus, each class (male, female, unknown) is associated with three numerical features. The aim is to handle colour features not as independent enumerated attributes, but as continuous numerical values, as shades of the same colour are expressed via close RGB values. The Bayesian classifier has been developed using the “scikit-learn” library (http://scikit-learn.org/), and given the fact that the colour attributes are expressed as continuous values, the Gaussian Naive Bayes algorithm has been adapted to the needs of the described problem.

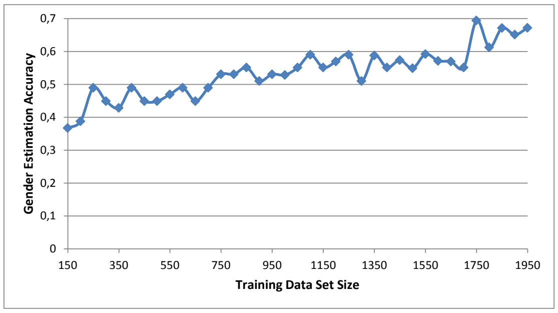


Figure 3. Evaluation of the gender estimation mechanism performance.

In order to evaluate the gender inference algorithm, the initial dataset (~10000 records) has been divided into 40 parts each containing about 250 records. Each dataset part was gradually incorporated to the classifier, while the last 250 records were used for evaluation. The initial evaluation attempts didn't provide high performance results. A data cleansing process was subsequently performed removing records that had the default predefined Twitter profile colors that resulted in a dataset of ~2000 records. The same evaluation process was then conducted where each of the 40 parts contained 50 records. The respective evaluation results are presented in Figure 3 and are rather encouraging, demonstrating about 70% of accurate classification when the entire training data set is incorporated.

The evaluation process is planned to proceed with further testing of the proposed approach based on more datasets, originating from additional social media (not only Twitter), to compare with similar existing approaches and to incorporate additional user profile attributes, including text analysis of provided profile description and Tweets text.

III. Social Recommendation & Personalization Tool

This section elaborates on the SRP tool, i.e., its functionality, architecture, recommendation extraction algorithm and respective evaluation/benchmarking.

A. *Functionality & Design*

Personalization & Social Recommendation are dominant mechanisms in today's social networks, online retails and multimedia content applications due to the increase in profit to the platforms as well as the improvement of the Quality of Experience (QoE) for its users and almost every online company has invested in creating personalized recommendation systems. Major examples include YouTube that recommends relevant videos and advertisements, Amazon that recommends products, Facebook that recommends advertisements and stories, Google Scholar that recommends scientific papers, while other online services provide APIs such as Facebook Open Graph API and Google's Social Graph API for companies to consume and provide their own recommendations [19].

The Social Recommendation & Personalization (SRP) tool of PRODUCER holistically addresses personalization, relevance feedback and recommendation, offering enriched multimedia content tailored to users' preferences. The tool's functionalities can be used in any type of content that can be represented in a meaningful way, as explained later. The application is thus not restricted to documentaries.

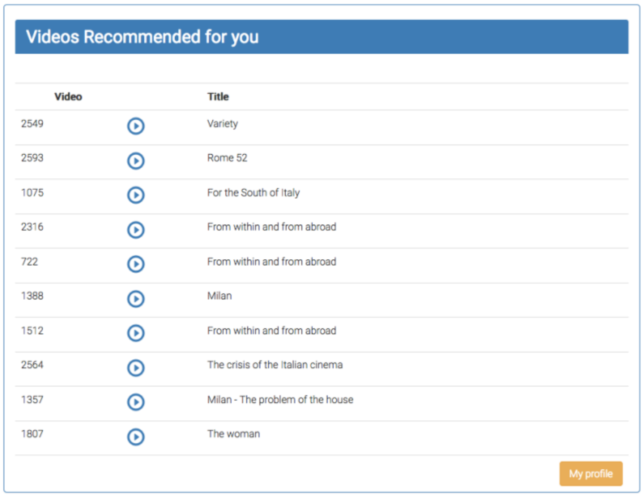
The recommendation system we built is not restricted to the video itself, but applies also to the set of enrichments accompanying the video. Interaction with both video and enrichments is taken into consideration into updating the user's profile, thus holistically quantifying the user's behaviour. Its goal is to facilitate the creation of the documentary and allow the reach of the documentary to a wider audience. To do so, the SRP tool is responsible for proposing appropriate content for specific target groups to the producer of the film via a personalization mechanism.

The first process the SRP tool has to perform is to index the content in a meaningful way, an important step as also indicated in [20][21]. Each video/enrichment is mapped to a vector, the elements of which are the scores appointed to the video/enrichment expressing the relevance it has to each category we have defined. The categories used come from the upper layer of DMOZ (http://dmoztools.net/), an attempt to create a hierarchical ontology scheme for organizing sites, that fits the generic nature of the PRODUCER videos.

Each multimedia content item is therefore described as follows: $X_P = [X_{P_1}, X_{P_2}, \dots, X_{P_N}]$, where P_i are the specified categories and X_{P_i} are appointed using the Doc2Vec algorithm [22]; the metadata of each item are passed through a neural network which represents the item with a multidimensional vector. The same procedure is done with the defined categories, and the vector X_P is constructed by finding the similarity of the multi-dimensional vectors of the item with each of the categories.

In order to be able to identify content relevant to target audiences, the tool needs to collect information and preferences of viewers since user profiles constitute another integral part of a recommendation system. Within the platform the SRP tool operates, the viewer registers and provides some important demographics (i.e., gender, age, country, occupation and education), as well as some of his/her preferences on specified topics, that will be used to identify the audience group that the viewer is part of. Alternatively, instead of providing this information explicitly, the viewer can choose to login with his/her social network account (e.g., Facebook, Twitter) and this information could be extracted automatically.

The user profile created via this process is static and is not effective for accurate recommendation of content since: a) the user is not able to accurately express his/her interests and b) his/her interests change dynamically. Thus, in addition to the above process the SRP tool implicitly collects information for the user's behavior and content choices. Using information about the video he/she watched or the enrichments that caught his/her attention, the SRP tool updates the viewer's profile to reflect more accurately his/her current preferences.



Videos Recommended for you

Video		Title
2549	▶	Variety
2593	▶	Rome 52
1075	▶	For the South of Italy
2316	▶	From within and from abroad
722	▶	From within and from abroad
1388	▶	Milan
1512	▶	From within and from abroad
2564	▶	The crisis of the Italian cinema
1357	▶	Milan - The problem of the house
1807	▶	The woman

Figure 4. Recommendation of content to PRODUCER viewers based on their user profile.

The created user profile, allows the tool to suggest content to the viewer to consume (Figure 4), as well as a personalized experience when viewing the content by showing only the most relevant enrichments for his/her taste. Through a content-based approach, the user's profile is matched with the content's vector by applying the cosine similarity measure as:

$$sim_{up}^{cf}(i,j) = \frac{U_i \cdot X_P^j}{\|U_i\| \left\|X_P^j\right\|} \tag{1}$$

where U_i is the user's profile vector and X_P^j is the content's vector.

The collaborative approach is complementary with the content-based recommendation using information from other viewers with similar taste, to increase diversity. The idea is to use already obtained knowledge from other users in order make meaningful predictions for the user in question. To do so, the similarity between users is computed as follows:

$$sim_{uu}(i,j) = \frac{U_i \cdot U_j}{\|U_i\|\|U_j\|} \tag{2}$$

where the H more similar users are denoted as neighbors. We then compute the similarity of the neighbors to the item:

$$sim_{up}^{cbf}(i,j) = \sum_{s=1}^{H} sim_{up}^{cf}(i,s) \cdot sim_{uu}(s,j) \tag{3}$$

and the final similarity between the user and the item is calculated via a hybrid scheme by using the convex combination of the above similarities:

$$sim_{up}^{h}(i,j) = (1-\theta) sim_{up}^{cbf}(i,j) + \theta sim_{up}^{cf}(i,j) \tag{4}$$

where $\theta : 0 \leq \theta \leq 1$ is a tunable parameter denoting the importance of the content-based and the collaborative approach on the hybrid scheme. A value of $\theta = 0.5$ has been shown to produce better results than both approaches used individually [23].

Based on the collected data above and constructed viewers' profiles, the producer of the documentary can filter the available content based on the preferences of the targeted audience. For this purpose, the k-means algorithm [24] is used to create social clusters of users. Based on the generated clusters, a representative user profile is extracted and is used to perform the similarity matching of the group with the content in question. The SRP tool assigns a score to each item and ranks the items based on that score.

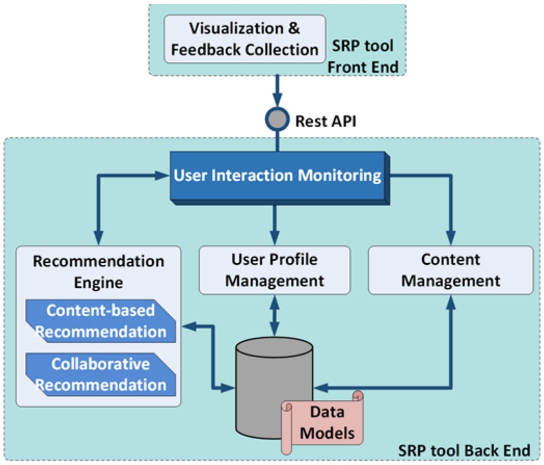

Visualization & Feedback Collection
SRP tool Front End
Rest API
User Interaction Monitoring
Recommendation Engine
Content-based Recommendation
Collaborative Recommendation
User Profile Management
Content Management
Data Models
SRP tool Back End

Figure 5. Architecture of the Social Recommendation & Personalization Tool.

After the creation of the documentary, the SRP tool can provide a filtering on the enrichments that are paired with the video, so that they do not overwhelm the viewer, filtering out less interesting enrichments. After specifying the target audience, the SRP tool can provide the list of suggested enrichments that the producer can either accept automatically or select manually based on his/her preferences, enabling the delivery of personalized documentary versions, tailored to audience interests.

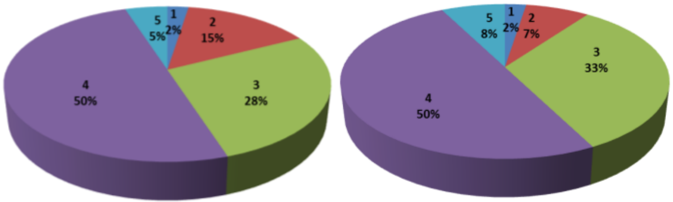
SRP tool's architecture is presented in Figure 5.

B. *Evaluation & Benchmarking*

In order to perform an initial evaluation of the SRP tool, actual user studies were performed during the MECANEX project [25]. The targeted group of users requested to participate in the study where approximately 150 students from the National Technical University of Athens, because their technical, informatics and/or marketing background would be useful in evaluating the tool. Eventually, 40 subjects participated and successfully completed the provided questionnaire, mainly students at the Techno-economics Master's program, an interdisciplinary graduate program designed for professionals.

During the study, each user had to register to the system by providing a username and a password, as well as some demographic information (e.g., name, age, education). He/she could then explicitly choose some initial topics of interest, resulting in a diversified set of preferences that were used by the algorithm to perform some initial recommendations. Based on this initial profile, ten videos from a set of available 2500 videos were shown to the user, who could then choose which one to watch and interact with. Using the information regarding the user interactions with the content, the SRP tool updated the respective user's profile, and a new set of videos was provided to the user. The users were asked to stop using the system as soon as they believed they were ready to rate its quality. The overall results of the study are presented in Figure 6.

More specifically, in Figure 6.a we can see that for the majority of the users, the SRP tool succeeded in predicting their expected profile after the use of the system, with 55% rating the matching of their profile with 4 or 5 stars. The above results come as verification to the simulations of the effectiveness of the algorithm performed in [26]. The overall experience of the tool was also rated highly by the subjects (Figure 6.b) with more than 50% giving 4 or 5 stars rating once more, which indicates that the proposed SRP tool is a well performing recommendation system.



(a) (b)

Figure 6. (a) Matching of final users' profile with their likes/preferences (1: Not really, 5: Matched exactly), (b) Rating of overall experience of the tool (1: Very Bad, 5: Very good).

More results concerning the impact and the effectiveness of the SRP tool in the MECANEX platform can be found in the public deliverable of the project [26]. Further, more thorough evaluation of the tool will be performed within the PRODUCER project's timeline.

IV. CONCLUSIONS

This paper introduced the PRODUCER platform for personalized documentary creation based on trend discovery. It briefly presented the set of tools offered by this platform, as well as its high level architecture. It then elaborated on two of its tools focusing on the targeted audience interests, identification and satisfaction. On the one hand, the ITD tool allows the identification of the most engaging topics to specified target audiences in order to facilitate professional users in the documentary preproduction phase. On the other hand, the SRP tool significantly improves the viewers' perceived experience via the provision of tailored enriched documentaries that address their personal interests, requirements and preferences. Initial prototype implementations of these tools are already available, while final prototypes will be delivered by spring 2018.

Both tools will be demonstrated and evaluated for a period of 3 months (March–May 2018) in an operational environment from an Italian broadcaster and a Belgium documentary production SME. This evaluation process will provide valuable feedback for further improving the overall functionality of the tools. Future plans also include the tools' integration with proprietary documentary production support services/infrastructures, as well as their extended evaluation and benchmarking against the various user requirements identified and against the Key Performance Indicators targeted (such as: cost reduction, time saving, increase of revenue in the entire documentary creation process).

Acknowledgment

This work has been partially supported by the European Commission, Horizon 2020 Framework Programme for research and innovation under grant agreement no 731893.

References

[1] The PRODUCER project. http://www.producer-project.eu, 2017. [*Retrieved January 2018*]

[2] G. Mitsis et. al, "Emerging ICT tools in Support of Documentary Production", 14th European Conference on Visual Media Production, 2017.

[3] J. Ginsberg, et. al, "Detecting influenza epidemics using search engine query data", Nature 457, pp. 1012-1014, 2009.

[4] A. J. Ocampo, R. Chunara, and J. S. Brownstein, "Using search queries for malaria surveillance, Thailand", Malaria Journal, Vol. 12, pp. 390-396, 2013.

[5] S. Yang, et. al, "Using electronic health records and Internet search information for accurate influenza forecasting", BMC Infectious Diseases BMC series, inclusive and trusted, Vol. 17, pp. 332-341, 2017.

[6] F. Ahmed, R. Asif, S. Hina, and M. Muzammil, "Financial Market Prediction using Google Trends", International Journal of Advanced Computer Science and Applications,Vol. 8, No.7, pp. 388-391, 2017.

[7] N. Askitas and K. F. Zimmermann, "Google econometrics and unemployment forecasting", Applied Economics Quarterly, Vol. 55, pp. 107-120, 2009.

[8] S. Vosen and T. Schmidt, "Forecasting private consumption: survey-based indicators vs. Google trends", Journal of Forecasting, Vol. 30, No. 6, pp. 565–578, 2011.

[9] S. Goel, J. M. Hofman, S.Lahaie, D. M. Pennock, and D. J. Watts, "Predicting consumer behavior with Web search", Natl Acad Sci USA, Vol. 107, No. 41, pp. 17486–17490, 2010.

[10] B. O'Connor, R. Balasubramanyan, B. R. Routledge, and N. A. Smith, "From Tweets to Polls: Linking Text Sentiment to Public Opinion Time Series", International AAAI Conference on Weblogs and Social Media, pp. 122–129, 2010.

[11] M. X. Hoang , X. Dang , X. Wu , Z. Yan , and A. K. Singh, "GPOP: Scalable Group-level Popularity Prediction for Online Content in Social Networks", 26th International Conference on World Wide Web, pp. 725-733, 2017.

[12] A. Oghina, M. Breuss, M. Tsagkias, and M. de Rijke, "Predicting IMDB movie ratings using social media", 34th European conference on Advances in Information Retrieval Springer-Verlag, pp. 503-507, 2012.

[13] B. Bhattacharjee, A. Sridhar, and A. Dutta, "Identifying the causal relationship between social media content of a Bollywood movie and its box-office success-a text mining approach", International Journal of Business Information Systems, Vol. 24, No. 3, pp. 344-368, 2017.

[14] J.D. Burger, J. Henderson, G. Kim, and G. Zarrella. "Discriminating gender on Twitter", Conference on Empirical Methods in Natural Language Processing. Association for Computational Linguistics, pp. 1301–1309, 2011.

[15] A. Culotta, N. R. Kumar, and J. Cutler, "Predicting the Demographics of Twitter Users from Website Traffic Data", AAAI, pp. 72–78, 2015.

[16] Q. Fang, J. Sang, C. Xu, and M. S. Hossain, "Relational user attribute inference in social media", IEEE Transactions on Multimedia, Vol. 17, No. 7, pp. 1031–1044, 2015.

[17] Y. Fu, G. Guo, and T. S. Huang, "Age synthesis and estimation via faces: A survey", IEEE transactions on pattern analysis and machine intelligence, Vol. 32, No. 11, pp. 1955–1976, 2010.

[18] I. H. Witten, E. Frank, and M. A. Hall, "Data Mining: Practical Machine Learning Tools and Techniques" book (3rd Edition), Morgan Kaufmann Series in Data Management Systems, Burlington, MA, USA, 2011.

[19] J. Osofsky. After f8: Personalized Social Plugins Now on 100, 000+ Sites. https://developers.facebook.com/blog/post/382, 2010. [*Retrieved January 2018*]

[20] A. Micarelli and F. Sciarrone, "Anatomy and empirical evaluation of an adaptive web-based information filtering system", User Modeling and User-Adapted Interaction, Vol. 14, No. 2-3 (2004), 159–200, 2004.

[21] G. Gentili, A. Micarelli, and F. Sciarrone. Infoweb: An adaptive information filtering system for the cultural heritage domain. Applied Artificial Intelligence, Vol. 17, No. 8-9, pp. 715–744, 2003.

[22] Q. Le and T. Mikolov, "Distributed representations of sentences and documents", 31st International Conference on Machine Learning, pp. 1188–1196, 2014.

[23] E. Stai, S. Kafetzoglou, E. E. Tsiropoulou, and S. Papavassiliou, "A holistic approach for personalization, relevance feedback & recommendation in enriched multimedia content", Multimedia Tools and Applications, 1–44. 2016.

[24] J. MacQueen, "Some methods for classification and analysis of multivariate observations", 5th Berkeley symposium on mathematical statistics and probability, Vol. 1. Oakland, CA, USA., pp. 281–297, 1967.

[25] The MECANEX project. http://mecanex.eu/, 2016. [*Retrieved January 2018*]

[26] MECANEX Deliverable D2.2: Multimedia Content Annotations for Rapid Exploitation in Multi-Screen Environments, 2016.

GR-Media: Spatial Reference Model for Geo-Referenced Media (Work-in-progress paper)

Taehoon Kim, Joon-Seok Kim, Azamat Bolat, Dongmin Kim, Ki-Joune Li

Department of Electrical and Computer Engineering
Pusan National University
Busan, Korea
Email: {taehoon.kim, joonseok, azamat.bolat, dongmin.kim, lik}@pnu.edu

***Abstract*—Recently, the size of digital media contents is growing. Most of the media content is made to reflect the real world. In particular, media created by reflecting the location of the real world is called Geo-Referenced media. However, the media is divided into various types, such as photographs, videos, games and novels. So, it is complicated to make this media as a Geo-Referenced media. The purpose of this study is to define a model that provides location references of various media, and to provide a way to connect real-world and various Geo-Referenced media types. In this paper, media types include photography, painting, video, animation, comics, and novels.**

Keywords–Geo-Referenced Media; Spatial Reference Model.

I. Introduction

A variety of contents, such as images, multimedia, and novels directly or indirectly express and relate objects, places, and locations existing in the real world. For example, the village of Hobbiton, which appeared in The Lord of the Rings movie is shot in places of the Waikato town of Matamata in the real world. This place has become one of the places, which was visited with curiosity by people who received indirect experience through the movies. Such Geo-Referenced media is related to real-world through various type of media, and this association affects other fields like tourism industry and the content industry.

The model connecting the media and the real world is an important requirement for the utilization of contents and the diffusion of new industries. The problem of connecting the media and the real world is classified into three stages as shown in Figure 1:

1) How do you create a model that connects the media and the real world?
2) How do you author data by using authored reference models?
3) What services will you use for your data?

Therefore, in this paper, we propose a spatial reference model for constructing Geo-Referenced media for contents that are closely related to the real world for the first step. The proposed model needs to reflect the existing standard model that expresses indoor and outdoor space and place. In addition, we propose a model considering the reference method depending on the type of Geo-Referenced media.

This paper is organized as follows. In section II, we introduce related research and derive the specific requirements of the GR-Media model. Section III explains the types and characteristics of media. Section IV describes the media contents of the proposed model. Section V summarizes the research contents and discusses future research.

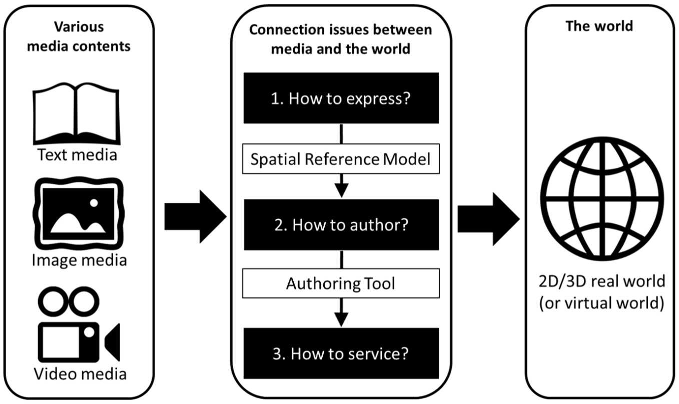


Figure 1. Problem definition for Geo-Reference media

II. Related Work

A. Geo-Reference Media

This section introduces the research about combining multimedia and spatial reference information. First, MediaQ[1][2] is a service that makes easy to find media taken from outdoor space based on location. MediaQ manages the media based on the field-of-view (FoV) information shown in Figure 2.

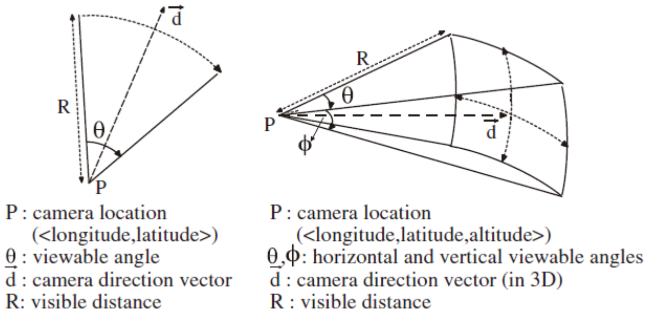

d
R
θ
P
R
θ
P
ϕ
d
P : camera location (<longitude,latitude>)
θ : viewable angle
d : camera direction vector
R: visible distance
P : camera location (<longitude,latitude,altitude>)
θ,ϕ: horizontal and vertical viewable angles
d : camera direction vector (in 3D)
R : visible distance

Figure 2. Illustration of camera field-of-view (FOV)
(a) in 2D (b) in 3D

In addition, there have been studies to add spatial information to media, created in indoor space instead of outdoor space[3][4]. These studies use the same FoV as MediaQ's FoV, but present FoV reflecting the characteristics of the room. In order to manage the indoor location information of the media, the indoor network of the building and the location information of the media are connected.

However, all these studies manage information based on the information of the camera that generates the media. In other words, there is no consideration for the object taken by the camera has taken. However, many objects in the media are closely related to the real world. In order to match the media with the real world, a spatial reference model for objects in the media is needed.

B. Spatial Data and Spatial Reference Model

For the link between media and the real world, it is necessary to link the spatial reference model for media with the existing spatial data model and the spatial reference model. This section introduces representative indoor and outdoor spatial data model and coordinate reference system.

IndoorGML[5] is an international standard for the representation and exchange of indoor spatial information in the Open Geospatial Consortium (OGC), which includes a data model for representing indoor space. In IndoorGML, a cell space is defined as a unit space constituting an indoor space, and an indoor space is represented as a set of cell space. IndoorGML also has an indoor network graph. The node of this graph correspond to the cell space and is represented by a state. And the edge of this graph means connectivity between cell spaces and is expressed as a transition. In case of media generated indoors, connection with the real world can be expressed through matching with the indoor network. In this case, the type of indoor network is the connectivity graph as shown in Figure 3.

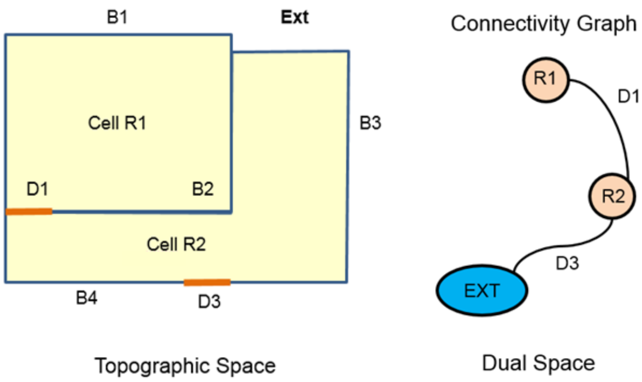


Figure 3. Example of connectivity graph in IndoorGML

CityGML[6] is an international standard established by OGC to represent urban models and includes data models for expressing road, building, tunnel, etc. There are various modules in the document, but ***_BoundarySurface*** most commonly used to express geometry. Therefore, in the case of media generated outdoors, the connection with the real world can be expressed by matching with the ***_BoundarySurface***.

ISO 19111[7] is a standard document describing a coordinate reference system, which has a model for the coordinate reference system as shown in Figure 4. The coordinate reference system consists of datum and coordinate system. The datum describes the relationship between the object and the coordinate system. The **SC_SingleCRS** has only one coordinate system. If more than one coordinate system is required, the **SC_CompoundCRS** can be used. Since there is a CRS already defined as shown in Figure 4, we define the necessary coordinate system to refer to the objects inside the media using the predefined CRS.

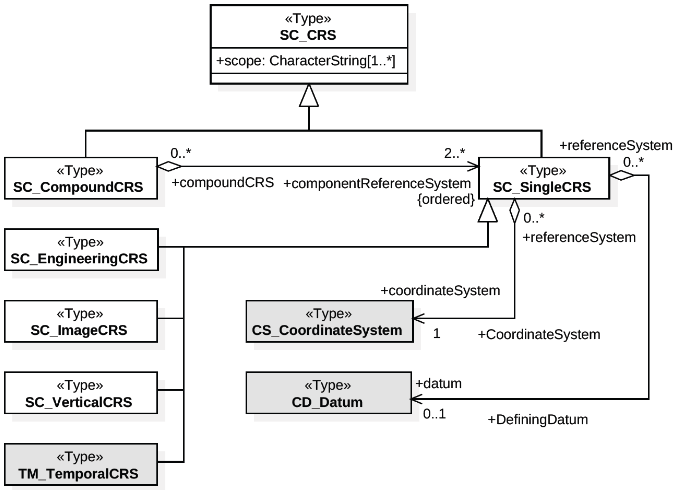


Figure 4. SC_CoordinateReferenceSystem package in ISO19111

As a result, the spatial reference model for Geo-Referenced meida needs to link IndoorGML for indoor media and CityGML for outdoor media. In addition, it is necessary to define the coordinate reference system for expressing the objects inside the media using ISO19111. Finally, we can create a spatial reference model for Geo-Referenced meida using classification according to the characteristics of the media.

III. Geo-Referenced Media Classification

In this paper, we design a spatial reference model for contents such as novel, painting, photography, video, movie, cartoon, and animation. The target contents are all digital media. Media is divided into text media, image media, and video media due to it's characteristics. Table I shows the contents classified according to the classified media types. Novel is classified as text media. Contents that are targeted to one image, such as painting or photographs, are classified as an image media. Cartoons are classified as multi-image media because they are having multiple images classified in cut units. Finally, content whose image changes over time, such as video, movie, and animation, is classified as a video media.

In media, objects that require spatial information are divided into two types. One is the camera object that create the media, and camera objects have FoV information and their presence is determined by the type of media. The other is an object inside the media. The objects inside the media have the following two kinds of information according to the types of the media.

TABLE I. CLASSIFICATION OF CONTENT ACCORDING TO MEDIA TYPE

	Text media	*Image media*	*Multi-image media*	*Video media*
Content	Novel . . .	Painting Photography . . .	Cartoon . . .	Video Movie Animation . . .

- For text media:
  - The location of the object in the text
  - The spatial information of objects in the corresponding real world
- For image(or video) media:
  - Boundary geometry information for objects in images
  - The spatial information of objects in the corresponding real world

Finally, objects inside the media have two pieces of spatial information together: reference information about the media and spatial information of the real world. Therefore, a connection between two pieces of information is required.

IV. GEO-REFERENCED MEDIA MODEL

Based on the spatial information standard introduced in Section II and the classification according to the media characteristics defined in Section III, the GR-Media model as shown in Figure 5 was constructed.

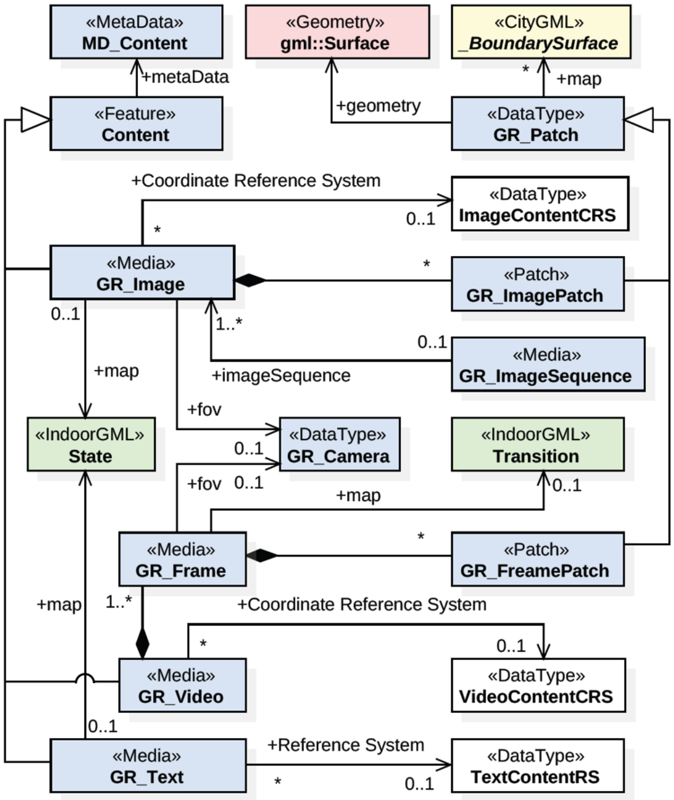


FIGURE 5. GEO-REFERENCED MEDIA MODEL

The characteristics of the GR-Media model are as follows.

All media inherit **Content** class. **Content** class has **MD_Content** class indicating metadata including reference space information. Media is divided into **GR_Image**, **GR_ImageSequence**, **GR_Video** and **GR_Text** class depending on the type.

GR_Image class may have **GR_ImagePatch** class for the representation of objects in an image and uses **ImageContentCRS** class as the coordinate reference system. **ImageContentCRS** basically use **SC_ImageCRS** class.

In the case of **GR_ImageSequence** class, it consists of a sequence of **GR_Image**. **GR_ImageSequence** is required for content with multiple images. For example, in the case of cartoons, it belongs to **GR_ImageSequence** because it is composed of several cut images.

In the case of **GR_Video** class, it is composed of several frame images(**GR_Frame** class). **GR_Frame** can have **GR_FramePatch** class for the representation of the objects in the frame image, and **VideoContentCRS** class is used as the coordinate reference system. **VideoContentCRS** uses **SC_ImageCRS** class to refer to an object in the frame image, and uses **TM_TemporalCRS** class to represent the time. That is, **SC_CompoundCRS** class with two coordinate reference systems is used.

In the case of **GR_Text** class, various types of media exist depending on contents, and there are various reference methods. Therefore, the reference model is designed to have only **TextContentRS** class to be extensible, and is not described in this model. In addition, **GR_Text** provided in this model assumes a text-only media.

GR_Camera class stores FoV information when media is created. Only media with one image are designed to have FoV.

Patches for representing objects inside each media inherit **GR_Patch** class. **GR_Patch** has the geometry of the object as Surface and is mapped to CityGMLś ***_BoundarySurface*** class.

In the case of media representing indoor space, media such as **GR_Image**, which do not have a temporal relationship between media, correspond to the **State** class of the IndoorGML. On the other hand, media such as **GR_Frame**, which have temporally continuous relationship with media, correspond to the **Transition** class of IndoorGML.

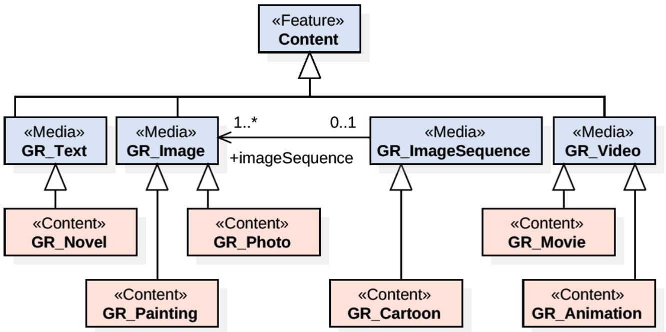


FIGURE 6. EXTENDED CONTENT MODEL

Figure 6 shows the extended model using the GR-media model for the content defined in Section III. First, text-based media, such as fiction inherits the **GR_text** class. Second, media consisting of a single image, such as paintings and photographs, inherits the **GR_Image** class. Third, media composed of multiple images, such as cartoons, inherits the **GR_ImageSequence** class. Finally, for video media, such as movies, animations, the **GR_Video** class is inherited.

An example using the GR-Meida model is shown in Figure 7. **GR_Video** class has basic information about video (title, length, etc. of video). For example, Figure 7 depicts PSYś Gangnam Style music video. In this case, when the Trade tower appears in 32 seconds, information about the frame image is expressed by **GR_Frame** class. Then, make a polygon for the area corresponding to the Trade tower in the frame image, and express it using **GR_FramePatch** class. Finally, the 3D model of CityGML is mapped with the corresponding ***_BoundarySurface*** class to express the connection between the media and the real world.

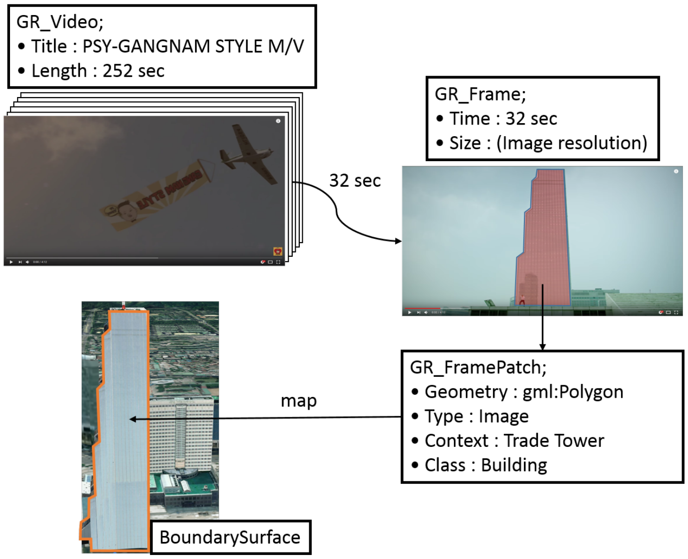


FIGURE 7. EXAMPLE OF GR_VIDEO

V. CONCLUSION

In this paper, we design a spatial reference model for connecting various types of digital media contents and real world.

However, there is a lack of consideration for content with complex situations. For example, it is difficult to use the proposed model in the case where a movie have information about the real world and the virtual world. Also, there is a limitation in covering media content like games, music, etc. In the future, we plan to design a model that considers more diverse content and matches with virtual worlds or multiple worlds.

ACKNOWLEDGMENT

This work was supported by BK21PLUS, Creative Human Resource Development Program for IT Convergence and Ministry of Culture, Sports and Tourism(MCST) and Korea Creative Content Agency(KOCCA) in the Culture Technology(CT) Research & Development Program 2017.

REFERENCES

[1] S. Kim, Y. Lu, G. Constantinou, C. Shahabi, G. Wang, and R. Zimmermann, “MediaQ: Mobile media management framework,” in Proceedings of ACM International Conference on Multimedia System. ACM, 2014.

[2] S. H. Kim, S. A. Ay, and R. Zimmermann, “Design and implementation of geo-tagged video search framework,” Journal of Visual Communication and Image Representation, vol. 21, no. 8, 2010, pp. 773–786.

[3] K.-J. Li, S.-J. Yoo, and Y. Han, “Geo-coding scheme for multimedia in indoor space,” in Proceedings of the 21st ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems. ACM, 2013, pp. 424–427.

[4] J.-S. Kim, S. H. Kim, and K.-J. Li, “Automatic geotagging and querying of indoor videos,” in Proceedings of the Fifth ACM SIGSPATIAL International Workshop on Indoor Spatial Awareness. ACM, 2013, pp. 50–53.

[5] J. Lee et al., “OGC IndoorGML,” Open Geospatial Consortium standard, 2014.

[6] G. Gröger, T. H. Kolbe, A. Czerwinski, and C. Nagel, “OpenGIS city geography markup language (citygml) encoding standard,” Open Geospatial Consortium Inc, 2008, pp. 1–234.

[7] E. ISO, “19111–iso 19111 spatial referencing by coordinates,” International Organization for Standardization, vol. 200, 2003.

An Efficient HDR Video Compression Scheme based on a Modified Lab Color Space

Maryam Azimi, Panos Nasiopoulos
Electrical and Computer Engineering Department
University of British Columbia
Vancouver, Canada
e-mail:{maryama,panos}@ece.ubc.ca

Mahsa T. Pourazad
TELUS Communications Inc.,
Vancouver, Canada
e-mail: mahsa.pourazad@gmail.com

Abstract— **With recent developments in both High Dynamic Range (HDR) capturing and display technologies, consumer distribution of HDR videos is now possible. The current distribution pipeline, however, is based on Standard Dynamic Range (SDR) signal characteristics. To accommodate HDR in the current pipeline infrastructure, some adjustments are necessary. One element of the pipeline that needs to be addressed is compression scheme. Current compression standards, such as HEVC, rely on YC_bC_r as their color encoding. However, this color encoding cannot represent the HDR signal without introducing visual artifacts. Two approaches exist for compensating for these visual errors: re-adjustment of the HDR signal in YC_bC_r as a pre-processing step to compression, or employing a color encoding that better meets the HDR signal requirements. In this paper, we propose to use the perceptually uniform CIELAB color space for HDR video color encoding following the latter approach. To make CIELAB suitable for HDR video color encoding, we change the transfer function and the scaling of the a* and b* channels. The performance of this approach is compared to YC_bC_r color encoding in our study via compressing these signals based on High Efficiency Video Coding (HEVC) standard and comparing the bitrate of these signals at the same quality level. An average bit-rate saving of 12.9% in terms of Overall Signal to Noise Ratio (OSNR), and 41.3% in terms of DE100 are reported for our proposed color encoding scheme compared to the conventional YC_bC_r. A negligible average loss of 0.6% is reported in terms of perceptually transformed Peak Signal to Noise Ratio (tPSNR).**

Keywords- High Dynamic Range (HDR); color encoding; perceptual quantizer; color difference, CIELAB.

I. INTRODUCTION

Recent advances in capturing and displaying technologies have made consumer distribution of HDR content possible. However, the current video distribution pipeline elements, including compression standards, are designed based on Standard Dynamic Range (SDR) videos characteristics. Given the distinct characteristics of HDR [1], special considerations in terms of both processing and compression need to be taken into account for efficient and true-to-original quality HDR distribution.

High Efficiency Video Coding (HEVC) is the latest compression standard used in video distribution pipelines [2]. HEVC and its predecessors rely on YC_bC_r as their color encoding, where Y represents luminance as a weighted combination of relative light Red (R), Green (G), and Blue (B). This relative light information is obtained by applying a perceptual transfer function known as gamma encoding (standardized as BT.1886 [3]). C_b and C_r are the blue and red differences from the luminance channel, respectively. For the limited brightness range (0.01 to 100 cd/m^2) and color gamut (BT.709 [4]) of SDR, 8-bit YC_bC_r represents SDR signal without visible artifacts or color differences.

On the other hand, HDR content is characterized by a higher range of brightness (usually 0.005 to 10000 cd/m^2) and mainly wider color gamut of BT.2020 [5]. For signals with this wide range of brightness and color, gamma encoding and 8-bit quantization cannot perceptually representative. Thus, a new perceptual transfer function, known as Perceptual Quantizer (PQ) was designed specifically for HDR and later was standardized as SMPTE ST 2084, [6]. Although by changing the transfer function and increasing the bit-depth the quality of the HDR signal is improved by precluding most of the quantization artifacts, still a 10-bit PQ YC_bC_r HDR signal has color differences with the original HDR signal [7]. These color errors are even present before compressing the signal and are intensified even more after applying the chroma 4:2:0 subsampling process as showed in [8]. The causes of these errors are the perceptual non-uniformity of the YC_bC_r signal and the existing correlation between luma (Y) and chroma (C_b and C_r) channels in YC_bC_r [9].

Two general approaches try to address the issues related to 10-bit PQ YC_bC_r 4:2:0 HDR color encoding: re-adjusting the YC_bC_r signal, or using a different color encoding which does not carry the known problems of YC_bC_r. The former approach requires additional pre-processing steps before video coding. On the other hand, the latter approach may need some adjustments during video coding to optimize the process for the new color encoding.

In this work, we study the performance of a modified CIELAB color space for HDR compression [10]. To make CIELAB suitable for HDR compression, we propose some changes to its original form. The modified CIELAB signals are then quantized to 10 bits followed by chroma down-sampling and then compressed using the HEVC codec. The performance of the proposed CIELAB color space for compressing HDR videos is compared with existing color encodings including YC_bC_r, in terms of the common HDR objective metrics.

The rest of this paper is organized as follows. Section II provides overview of existing color encodings. Section III provides details on the suggested modifications of CIELAB. Section IV presents and discusses the results and conclusions are drawn in Section V.

II. Overview of Existing Color Encoding Schemes for HDR Video Compresson

The existing HDR video coding system is based on 10-bit $Y'C_bC_r$. This signal is derived from the relative light R'G'B' signal. The prime on the representations of the signals denotes relative light, i.e., perceptually quantized instead of the linear light. It has been shown in [8] that two original linear RGB colors with similar luminance values transformed to 10-bit 4:2:0 $Y'C_bC_r$ using the conventional way followed by chroma subsampling, may result in two very different luminance (Y) values. That is why deriving Y' from relative light R'G'B' is referred to as Non-Constant Luminance (NCL) approach. Since even small changes in luminance values are quite apparent to human eyes, visible artifacts appear on 10-bit 4:2:0 NCL $Y'C_bC_r$ HDR signal.

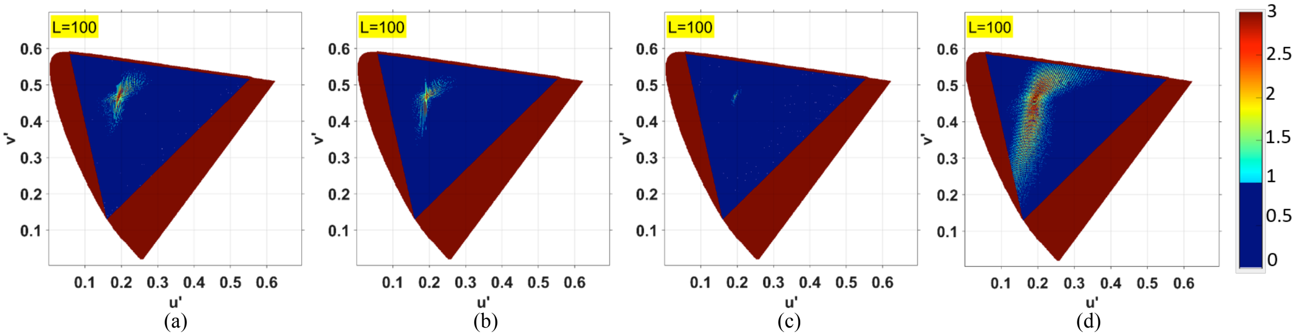
To overcome the non-constant luminance issue, Constant Luminance (CL) derivation of $Y'C_bC_r$ that is based on linear light RGB content can be utilized. However, to put it in practice, the entire infrastructure of the current video transmission system needs to be updated to support CL approach. To avoid such a costly update, a recursive re-adjustment algorithm of the Y' channel (luma) in the 10-bit NCL Y'CbCr to the Y' channel in the 10-bit CL Y'CbCr was proposed in [11]. It is reported in [11] that the visible artifacts of the 10-bit NCL Y'CbCr disappear in the re-adjusted version of the signal. These pre-processing luma adjustments are part of Supplement 15 to ITU-T H-series Recommendations [12] document which offers a description of processing steps and guidelines for converting from 4:4:4 RGB linear light representation video signals into adjusted 10-bit PQ $Y'C_bC_r$ 4:2:0 signal for HDR video transmission. A faster version of this algorithm is proposed in [13] and an approximation of this algorithm was proposed in [14] without the required pre-processing.

To avoid the necessary pre-processing steps associated with using $Y'C_bC_r$, another approach for transmitting HDR is to replace $Y'C_bC_r$ color encoding with a color encoding approach with more de-correlated color and brightness channels. One of such encoding is IC_tC_p [15]. The brightness (I) and color (C_t and C_p) channels information are highly de-correlated. Hence, IC_tC_p does not bear the chroma errors reported with $Y'C_bC_r$. IC_tC_p is designed based on SMPTE ST 2084 as the transfer function.

Another color encoding for HDR video compression is $Y'D'_zD'_x$ [16] that is based on XYZ colors space and SMPTE ST 2084 as the transfer function. This color space was investigated by Moving Picture Experts Group (MPEG) in the early exploration stages of HDR video compression and its requirements. Due to some artifacts seen on some of the HDR videos compressed with $Y'D'_zD'_x$, which in fact were clipping errors, this color encoding was no longer explored by MPEG.

Another proposed color encoding for HDR video compression is Yu"v" [17] that is based on Yu'v' [18]. A transfer function is proposed for this color space in [17]. For dark areas (with luminance value smaller than 5 cd/m^2), u" and v" represent the attenuated u' and v' by Y. Although this introduces dependency on Y for u" and v" channels, it is shown to reduce the noise in dark areas and hence results in better compression efficiency compared to Yu'v'[17].

Figure 1 (a), (b), (c), and (d) show how NCL $Y'C_bC_r$, CL $Y'C_bC_r$, IC_tC_p, and $Y'D'_zD'_x$, respectively represent the whole color gamut of BT.2020 at luminance level of 100cd/m^2 in terms of CIE DE2000 error [19]. Please note to provide fair comparisons of the color encoding schemes rather than transfer functions, we did not include Yu"v" results as it uses a different transfer function from the SMPTE ST 2084. CIE DE2000 is a color difference metric. Errors smaller than one, i.e., invisible errors are represented with dark blue in Figure 1. The visible color differences (with error value larger than 1) are represented with light



(a) (b) (c) (d)

Figure 1. Color encoding error in terms of CIE DE2000 for (a) NCL $Y'C_bC_r$, (b) CL $Y'C_bC_r$, (c) IC_tC_p, and (d) $Y'D'_zD'_x$ shown using an error bar (right) at luminance level of 100 cd/m^2

ISBN: 978-1-61208-611-8

blue to red. Note that these errors are only the quantization errors and chroma sub-sampling is not yet applied. Figure 1 shows that except for IC_tC_p (c), even before compression the colors of an HDR signal are visibly distorted.

In this paper, yet another perceptually uniform color encoding, CIELAB, and its performance for HDR video compression is investigated. As CIELAB is designed for SDR brightness values up to of 100 cd/m^2, some adjustments are made to its original form to accommodate HDR signal, which are presented in details in what follows.

III. PROPOSED MODIFICATIONS TO CIELAB FOR HDR COMPRESSION

CIELAB consists of one brightness channel (L*) which goes up to 100 cd/m^2 and two color channels a* and b* which cover colors from green to red, and from blue to yellow, respectively. Each of these channels is constructed as follows:

$$L^* = 116 f\left(\frac{Y}{Y_n}\right) - 16 \quad , \qquad (1)$$

$$a^* = 500\left[f\left(\frac{X}{X_n}\right) - f\left(\frac{Y}{Y_n}\right)\right] \quad , \qquad (2)$$

$$b^* = 200\left[f\left(\frac{Y}{Y_n}\right) - f\left(\frac{Z}{Z_n}\right)\right] \quad , \qquad (3)$$

$$f(w) = \begin{cases} w^{1/3}, & w > 0.008856 \\ 7.787\, w + 16/116, & w \le 0.008856 \end{cases} \qquad (4)$$

where X_n, Y_n, and Z_n are the XYZ components of the white point. Since HDR luminance values can go up to 10000 cd/m^2, the current CIELAB cannot efficiently handle an HDR signal. To address this issue, an hdr-CIELAB is proposed in [20]. The only change in hdr-CIELAB is the transfer function to have better performance for shadows and highlights compared to conventional CIELAB; otherwise all the other derivations are the same as in CIELAB. Still, the encoded L* in hdr-CIELAB goes only up to 245 cd/m^2.

In this work, we propose to use SMPTE ST 2084 as the transfer function for HDR luminance values in CIELAB. Therefore, the proposed L*, a* and b* channels will be calculated as follows:

$$L^* = Y' \quad , \qquad (5)$$

$$a^* = \begin{cases} \dfrac{X' - Y'}{0.1441 \times 2}, & -0.1441 \le x \le 0 \\ \dfrac{X' - Y'}{0.1083 \times 2}, & 0 < x \le 0.1083 \end{cases} \qquad (6)$$

$$b^* = \begin{cases} \dfrac{Y' - Z'}{0.2338 \times 2}, & -0.2338 \le x \le 0 \\ \dfrac{Y' - Z'}{0.6208 \times 2}, & 0 < x \le 0.6208 \end{cases} \qquad (7)$$

where X’, Y’, and Z’ are the perceptually quantized X, Y and Z signals using SMPTE ST 2084. In (6) and (7), a* and b* channels are scaled to fall within [-0.5 0.5] so that BT.1361 quantization [21] can be applied to them.

The proposed CIELAB for HDR signals is somewhat similar to $Y'D'_ZD'_x$ [16]. However, in the proposed modified CIELAB, color difference channels are scaled differently for positive and negative differences so that codewords are utilized more efficiently.

IV. EXPERIMENTAL SETUP

To evaluate the proposed modified CIELAB color encoding for HDR video compression, we use four HDR video sequences from MPEG HDR video dataset: FireEater2, Market3, BalloonFestival, and SunRise. All of these videos are 1920x1080p, and are in the BT.2020 container although their actual colors fall inside the BT.709 gamut. Figure 2 shows tone mapped snapshots of the first frame of each video sequence.

Figure 3 shows how the original linear light HDR content is encoded to the modified CIELAB, followed by quantization and chroma down-sampling. It is worth noting that our modified CIELAB-based method uses the original sampling filters designed specifically for $Y'C_bC_r$ and as such they are not optimized for our proposed scheme. For compression, we used the HEVC encoder reference software HM 16.15, Main10 profile. We coded the tested videos at four bit-rate levels using four QPs, as suggested in [22]. To compare them with the original ones in terms of quality, the color encoded and compressed signals are then de-compressed and converted back to the linear light domain as shown in Figure 3.

(a) (b) (c) (d)

Figure 2. Snapshots of the first frames of HDR test video sequences (tone-mapped version): (a) FireEater2, (b) Market3, (c) BalloonFestival, and (d) SunRise

Copyright (c) IARIA, 2018. ISBN: 978-1-61208-611-8

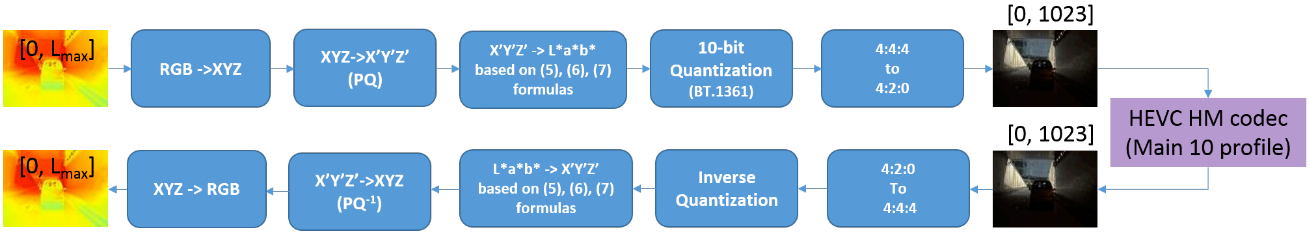

[0, Lmax]
RGB ->XYZ
XYZ->X'Y'Z' (PQ)
X'Y'Z' -> L*a*b* based on (5), (6), (7) formulas
10-bit Quantization (BT.1361)
4:4:4 to 4:2:0
[0, 1023]
HEVC HM codec (Main 10 profile)
[0, 1023]
4:2:0 To 4:4:4
Inverse Quantization
L*a*b* -> X'Y'Z' based on (5), (6), (7) formulas
X'Y'Z'->XYZ (PQ-1)
XYZ -> RGB
[0, Lmax]

Figure 3. Pre/post processing steps of the proposed modified CIELAB for HDR video compression

V. Results and Discussions

Figure 4 (a), (b), (c) and (d) shows the bit-rate versus DE100, Overall Signal to Noise Ratio (OSNR) and perceptually Transformed Peak Signal to Noise Ratio (tPSNR) in terms of (db) for the proposed modified CIELAB, the NCL YC_bC_r, luma-adjusted NCL $Y'C_bC_r$, IC_tC_p and $Y'D'_zD'_x$ for FireEater2, Market3, SunRise and BalloonFestival, respectively. tPSNR is the average of the PSNR X', Y' and Z'. OSNR is the overall SNR of X', Y' and Z' with calculation of the error for each pixel and then averaging the errors. DE100 is the PSNR based value of the average error in terms of CIE DE2000 metric [22]. Table I also shows the bit-rate savings in terms of the same metrics for the proposed color encoding over NCL $Y'C_bC_r$.

As can be seen from Figure 4, the proposed modified CIELAB clearly outperforms the NCL $Y'C_bC_r$, luma-adjusted NCL $Y'C_bC_r$, and $Y'D'_zD'_x$ in terms of DE100. This shows that the proposed method can maintain the original colors better at any given bit-rate. The proposed method performs almost identical to IC_tC_p in terms of DE100.

Moreover, it can be seen form Figure 4 that the proposed method also outperforms NCL $Y'C_bC_r$, luma-adjusted NCL $Y'C_bC_r$, and $Y'D'_zD'_x$ in terms of OSNR, especially at higher bit-rates. All the tested color encoding schemes seem to be performing similarly in terms of tPSNR.

Please note that the chroma down-sampling filter used for the proposed CIELAB is the same as the one in $Y'C_bC_r$. However, a better performance may be achieved in terms of tPSNR and OSNR if a new sampling filter is designed that better matches the a* and b* characteristics. Although this is not in the scope of this paper, it is part of our future work.

TABLE I. Bit-rate savings of the proposed CIELAB compared to NCL Y'CbCr

Metric Video	tPSNR X (%)	tPSNR Y (%)	tPSNR Z (%)	tPSNR XYZ (%)	tOSNR XYZ (%)	DE 100 (%)
FireEater 2	-7.4	8.0	-3.0	-1.0	-24.0	-32.6
Market3	13.1	17.4	2.7	10.5	6.2	-63.6
SunRise	0.2	9.9	-23.0	-6.3	-19.5	0.0
BalloonF estival	5.4	21.5	-17.7	-0.7	-14.3	-69.2
Average	2.9	14.2	-10.2	0.6	-12.9	-41.3

Moreover, the rate-distortion optimization (RDO) setting inside the encoder was maintained the same in all these experiments. Since the current RDO is customized for $Y'C_bC_r$ characteristics, it is expected that further improvements may be obtained by modifying the RDO process according to the proposed modified CIELAB color encoding. This step as well is in the scope of future work.

Another note-worthy observation from Figure 4 is how $Y'D'_zD'_x$ underperforms all the tested color encodings, although its derivation is very similar to what is proposed in this paper. However, as the proposed scaling of a* and b* employs the available codewords more efficiently, it achieves better compression performance compared to $Y'D'_zD'_x$ as observed in Figure 4.

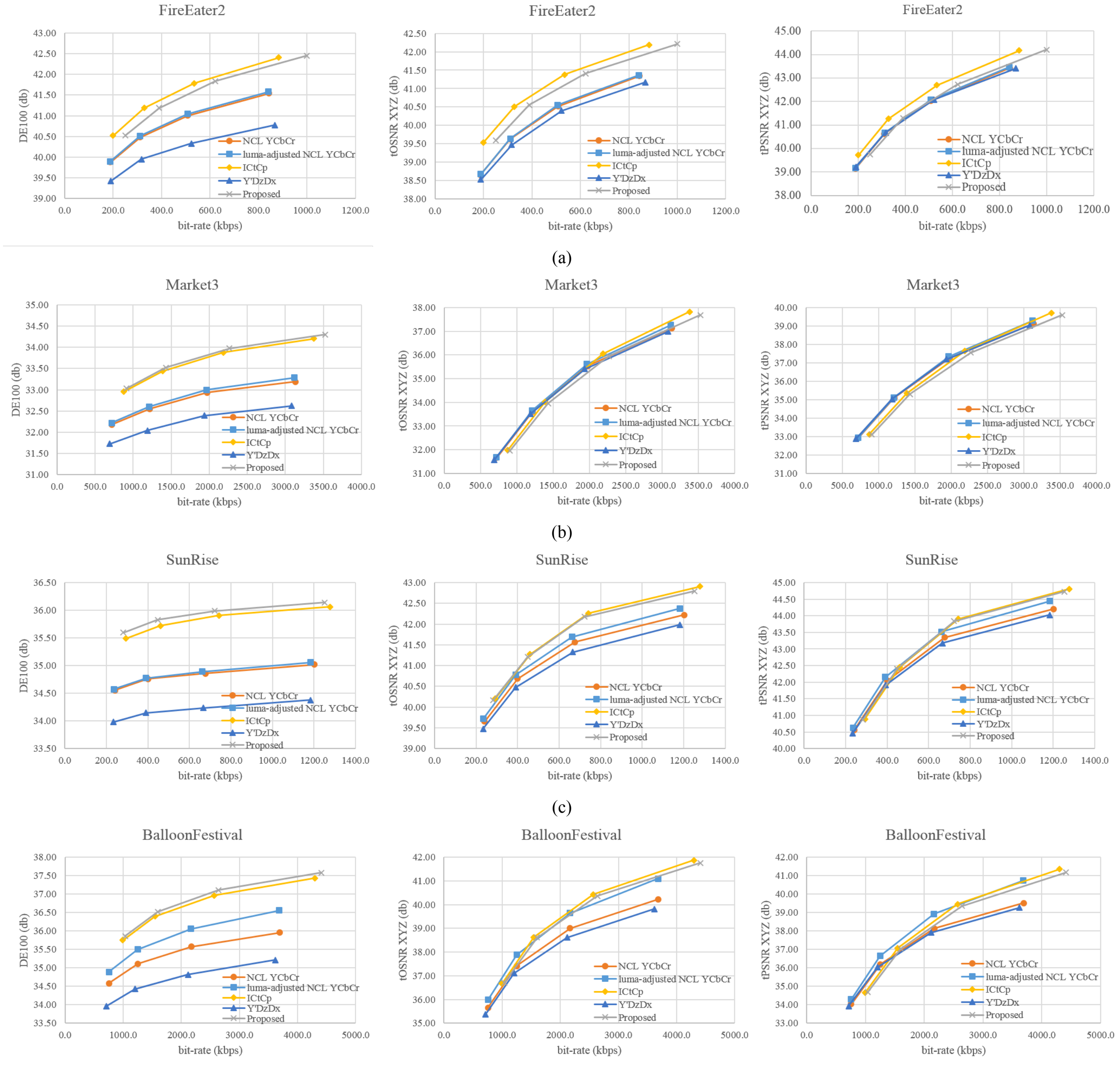
Overall, it is shown that the proposed color encoding results in better performance in terms of DE100 compared to conventional NCL $Y'C_bC_r$, by an average of 41% over the four videos, hence better maintaining the original HDR colors. By using a chroma down-sampling filter that is designed for the proposed space and changing the encoder rate-distortion optimization process, it is expected to improve the performance of the tested color in terms of tPSNR and OSNR.

VI. Conclusions

In this paper, we presented a modified CIELAB color encoding scheme for efficiently compressing HDR content.

Performance evaluations show that the proposed adjusted CIELAB space, even using the chroma down-sampling designed for $Y'C_bC_r$, maintains the original HDR colors better than other existing color spaces and results in an average of 41% bit-rate savings over four videos in terms of DE100 (db). The performance of the proposed modified color space even without changing the chroma sub-sampling filters of $Y'C_bC_r$ is almost similar to that of IC_tC_p.

The slight underperformance of the proposed approach in terms of tPSNR can be improved by changing the chroma down-sampling filter to a more tailored one to the a* and b*



(a)

(b)

(c)

(d)

Figure 4. R-D curves of the proposed color encoding compared to NCL YC_bC_r, luma-adjuste YC_bC_r, $Y'D'_zD'_x$ and IC_tC_p in terms of DE100 (db), OSNR (db) and tPSNR (db) for (a) FireEater2, (b) Market3, (c) SunRise, and (d) BalloonFestival

characteristics. Furthermore, changing the RDO process to be performed in the proposed space instead of $Y'C_bC_r$ may also result in further performance improvement in terms of tPSNR.

REFERENCES

[1] R. Boitard, M. T. Pourazad, P. Nasiopoulos, and J. Slevinsky, “Demystifying High-Dynamic-Range Technology: A new evolution in digital media,” Consumer Electron. Mag., vol. 4, pp. 72 – 86, 2015.

[2] M. T. Pourazad, C. Doutre, M. Azimi, and P. Nasiopoulos, “HEVC: The New Gold Standard for Video Compression: How Does HEVC Compare with H.264/AVC?” Consumer Electron. Mag., vol. 1, pp. 36 – 46, 2012.

[3] Reference electro-optical transfer function for flat panel displays used in HDTV studio production, ITU-R BT.1886, 2011.

[4] Parameter values for the HDTV standards for production and international program exchange, ITU-R BT.709-3, 1998.

[5] Parameter values for ultrahigh definition television systems for production and international programme exchange, ITU-R BT.2020, 2012.

[6] High Dynamic Range Electro-Optical Transfer Function of Mastering Reference Displays, SMPTE Standard ST 2084, 2014.

[7] M. Azimi, R. Boitard, M. T. Pourazad, and P Nasiopoulos, " Visual color difference evaluation of standard color pixel representations for high dynamic range video compression," 25th European Signal Processing Conference (EUSIPCO), pp. 1480 – 1484, Aug. 2017.

[8] E. Francois, "MPEG HDR AhG: about using a BT.2020 container for BT.709 content," 110th MPEG meeting, Strasbourg, France, October 2014.

[9] R. Boitard, R. K. Mantiuk, and T. Pouli. "Evaluation of color encodings for high dynamic range pixels," in Proc. SPIE 9394, Human Vision and Electron. Imaging XX, pp. 93941K1 – 9, Mar. 2015.

[10] A. R. Robertson, "The cie 1976 color-difference formulae," Color Research & Application, vol. 2, no. 1, pp. 7 – 11, 1977.

[11] J. Ström, J. Samuelsson, and K. Dovstam, "Luma Adjustment for High Dynamic Range Video," Proceedings of Data Compression Conference (DCC), pp. 319 – 328, Mar. 2016.

[12] Conversion and coding practices for HDR/WCG Y'CbCr 4:2:0 video with PQ transfer characteristics, ITU-T H.Sup15, 2017.

[13] A. Norkin, "Fast Algorithm for HDR Color Conversion," Proceedings of the IEEE Data Compression Conference (DCC), pp. 486 – 495, Mar. 2016.

[14] F. Xie, R. Boitard, M. T. Pourazad, and P. Nasiopoulos, "Optimizing Non Constant Luminance into Constant Luminance for High Dynamic Range Video Distribution," IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 1487 – 1491, 2017.

[15] T. Lu et al., "ITP Colour Space and Its Compression Performance for High Dynamic Range and Wide Colour Gamut Video Distribution," ZTE Communications, vol.14, no.1, pp. 32 – 38, Feb. 2016.

[16] Y'D'ZD'X Color-difference Computations for High Dynamic Range X'Y'Z' Signals, SMPTE Standard ST 2085, 2015.

[17] C. Poynton, J. Stessen, and R. Nijland, "Deploying Wide Color Gamut and High Dynamic Range In HD and UHD," SMPTE Motion Imaging J. vol. 124, no. 3, pp. 37 – 49, 2015.

[18] G.W. Larson, "LogLuv encoding for full-gamut, high-dynamic range images," J. Graph. Tools, vo. 3, no.1, pp. 15 – 31, 1998.

[19] G. Sharma, W. Wu, and E. N. Dalal, "The ciede2000 color-difference formula: Implementation notes,supplementary test data, and mathematical observations," Color Research & Application, vol. 30, no. 1, pp. 21 – 30, 2005.

[20] M. D. Fairchild, Color Appearance Models, 3rd Edition. Wiley, 2013.

[21] Worldwide unified colorimetry and related characteristics of future television and imaging systems, ITU-R BT.1361, 1998.

[22] A. Luthra, E. Francois, and W. Husak, "Call for Evidence (CfE) for HDR and WCG video coding," ISO/IEC JTC1/SC29/WG11 N15083, Feb. 2015.

Subjective Assessment for Text with Super Resolution on Smartphone Displays

Aya Kubota† Seiichi Gohshi‡

†‡Department of Information Science
Kogakuin University
Tokyo, Japan
e-mail: †em17008@ns.kogakuin.ac.jp, ‡gohshi@cc.kogakuin.ac.jp

***Abstract*— Smartphones appeared on the market only a decade ago. However, the market has since grown rapidly, and people of all ages now use smartphones. In many cases, people read text on their smartphones, but depending on the design of a website, it may be difficult to read its text. By improving the resolution of the text, the readability of text can be improved. One research area for increasing the resolution is super resolution (SR), which includes nonlinear signal processing super- resolution SR (NLSP), a method that can be implemented on smartphones. However, NLSP has never been applied to text in order to improve readability. We applied NLSP for text displayed on liquid crystal display (LCD), and verified its effectiveness. Thus, in this paper, the assessment results for text on LCD are discussed.**

Keywords- Nonlinear signal processing; Super-Resolution; Subjective assessment.

I. INTRODUCTION

Smartphones have become daily necessities in modern society. In addition to processing communication functions, such as telephone and e-mail, it is possible to obtain information in real time via the Internet. When used for the above functions, text must often be read, which could be on operation buttons or explanatory text. Support functions to make text easier to read, such as changing the font size are set in the application that is preinstalled in the operating system (such as mail, smartphone settings, etc.). However, there are websites that do not have a font size larger than a certain size even if the text is enlarged, and sites where the color of the background and the texts is not very different. Problems, such as these can therefore make it difficult to read text.

Improving the resolution of the images can make it easier to read text. Super-resolution (SR) technology is one method to improve resolution. Most 4K TVs are equipped with SR. Nonlinear signal processing SR (NLSP) is a SR technology that can be embedded into smartphones [1]. The algorithm is simple and fast: hence, processing with software is possible, and smartphones with NLSP are already being sold in the market [2]. The effectiveness of NLSP is higher than that of other SR technologies [3][4], and NLSP is effective even in smartphone videos [5].

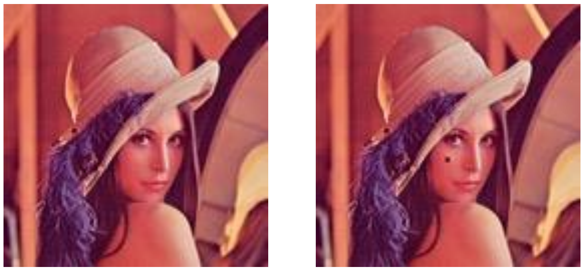
However, the effectiveness of NLSP for text on smartphone display has not been verified. In this study, we verify the effectiveness of using smartphone with NLSP compared one without NLSP.

Images processed with NLSP are introduced only to the display of the smartphone and there is no electric output of the processed image. Therefore, it is impossible to use an objective assessment because the objective assessment requires electric image signal with and without NLSP. Subjective assessment is the only way to assess the difference between the displays. However, subjective assessment is only a reflection of how we feel. It is difficult to ensure the reproducibility of the subjective assessments. The subjective assessments also requires observers and time to assess the image quality.

Although there are issues about the subjective assessment, ITU-R standardized subjective assessment methods. ITU-R BT.710 recommends experimental conditions to obtain reproducible results in subjective assessment experiments [6]. However, BT.710 does not mention practical quantitative scoring assessment which is defined in BT.500. They are the double stimulus continuous quality scale (DSCQS) and the double stimulus impairment scale (DSIS). In our case we need to compare five smartphones and they are different manufactures products, BT.500 and BT.710 do not meet our requirements. One of our authors developed an subjective assessment for multiple displays [6][14]. It applies best-worst method and statistical analysis is introduced to analyze reproducibility. It shows good results if the images/videos are selected appropriately. This paper is organized as follows. In Section II the subjective assessment for multiple displays is explained. In Section III NLSP is explained. In Section IV test images are presented and experiments are explained. In Section V the statistical analysis is adapted to the assessment results and in Section VI the analyzed result is discussed. Section VII is the conclusion of the paper.

II. ASSESSMENT METHOD

Objective assessment and subjective assessment are evaluation methods. Objective assessments analyze the signal and expresses high and low of image quality by a numerical value. However, results of objective assessment do not always match with how we feel. For example, an original image is given in Figure 1(a), and the degraded image is given in Figure 1(b). The peak signal to noise ratio (PSNR) of the degraded image in Figure 1(b) is 40.1112dB. A PSNR 40dB is generally said to be a high image quality [7], but Figure 1(b) contains degradation in the form of a black square in the center of the image. When images



(a) Original image

(b) Degraded image (PSNR: 40.1112dB)

Figure 1. Objective assessment by PSNR

include local degradation, the results of PSNR sometimes deviate from our feeling.

Thus, objective assessments cannot accurately reflect image quality. In addition, objective assessments require comparing the assessment image with the original image. As discussion the previous section, signals processed inside the smartphone cannot be output anywhere outside the display. Therefore, assessment by signal analysis is impossible, and thus the experiment is conducted using subjective assessment.

The best–worst method was adopted as the assessment method using multiple displays. Normalized ranking method and paired comparison method are other assessment methods. Experimental stimuli are ranked at once in the normalized ranking method. The process of the method is simple, but when differences between the stimuli are small, sometimes the differences cannot be detected because of large differences between stimuli influences. In the paired comparison method, stimuli are compared one on one and ranked. Two stimuli are selected, and observers evaluate the stimuli based on the other. Thus, differences between stimuli can be obtained in detail. However, evaluation is performed for all stimulus combinations, which places a heavy burden on the observers. In the best–worst method, observers select the best stimuli and the worst stimuli. After excluding the selected stimuli, observers again select the best and the worst from the remaining stimuli. The best–worst method can detect differences more accurately than the normalized ranking method, and the best–worst method is a smaller burden for observers than the paired comparison method. Therefore, the best–worst method is adopted in this paper.

In this study, an assessment experiment was conducted using five smartphones. The test images are screenshots of a website containing text.

III. NLSP

NLSP is a simple and fast SR technique. The process is similar to enhancer that it increases resolution by emphasizing edges; however, NLSP emphasizes high-frequency components extracted from the input image using a nonlinear function [8]. The nonlinear function can generate high-frequency components that are not included in the original image. These high-frequency components express edges and details of the image. An example nonlinear function is the cubic function ($\mathrm{f(x)} = x^3$). The function can amplify the high-frequency components by as much as three times. Figure 2 shows an example of NLSP processed image. Figure 2(a) is an original image. Figure 2(b) is a NLSP processed image. Figure 2(b) has more details, such as edges of mountain and the surface of it than the original image.

Super-resolution image reconstruction (SRR) and learning-based super resolution (LBSR) are the current mainstream SR technologies. SRR is a technology that generates a high-resolution image from multiple degraded images [9], but the processing requires iteration. When the input image and output image have the same resolution, the technique is not very effective [10]. LBSR is a method that increases resolution using a database [11]. The effectiveness is affected by the database, and the processing requires both an expensive database and iteration. Thus, both the above technologies require complex processing. In addition, their effectiveness is lower than that of NLSP [12][13].

IV. EXPERIMENT

The effect of image processing differs, depending on the images. We adjusted NLSP for text; hence, it was necessary to verify the effect of NLSP for text. A smartphone with NLSP and one without NLSP were compared. The result of the comparison indicates the effects of using NLSP. In addition, the experiment was conducted using smartphones from different manufacturers and verifies the effect of NLSP in comparison with other technologies.

A. Experimental equipment

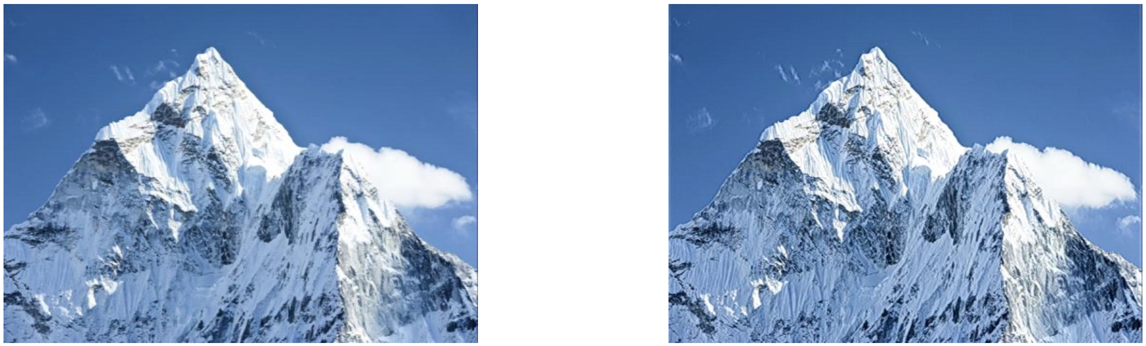
Five smartphones were used in this experiment. To ensure that the results are not caused by display differences, two of the five smartphones featured the same terminal. One was a smartphone with NLSP (smartphone A), and the other was one without NLSP (smartphone B). The remaining three smartphones were smartphones from different manufacturers (smartphone C–E). The display resolution of smartphone A and B was WQHD (2560 × 1440), whereas that of the others was full HD (1920 × 1080). The brightness was adjusted to be close to the same brightness.

B. Test images

Five screenshots of websites containing text were used as experimental images. The images are websites browsed by many people (a site for smartphones, a PC, a map). The site for smartphones are enlarged and viewed when the site has small texts, so an unenlarged site image and two enlarged site images were used. One of the two enlarged images contained text with only small differences in color from the background color. The images are shown in Figure 3. The resolution of all the images is WQHD.

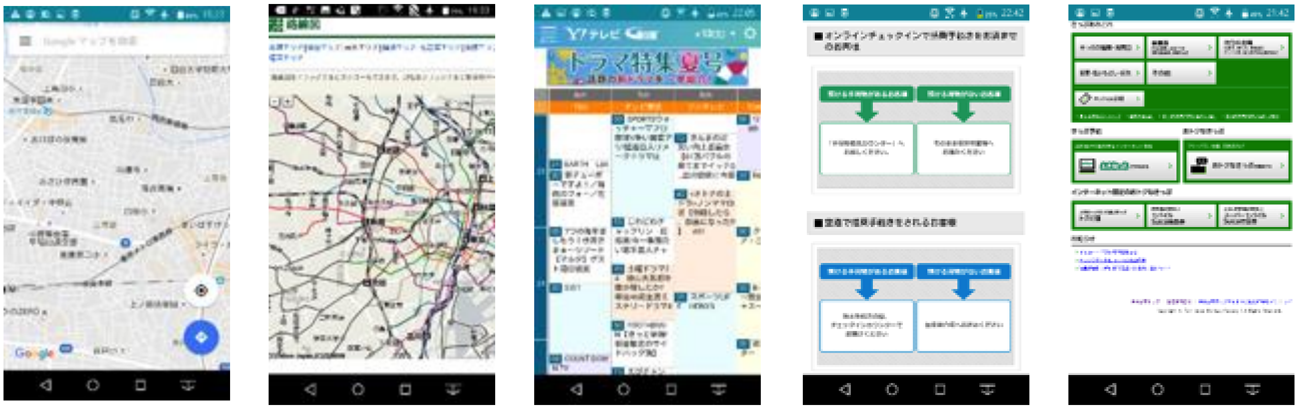
C. Observers

At least 20 observers are required for adequate statistical analysis. In this experiment, 23 observers participating in the experiment had normal visual acuity and color vision. Non-experts who do not work in the image industry cannot



(a) Original image (b) NLSP processed image

Figure 2. Example of NLSP processed image



(a) Map (b) Route (c) TV (d) Airport (e) Ticket

Figure 3. Test images

always distinguish image quality differences, even if experts can distinguish them. If there is a significant difference in the experiment using non-experts, the difference of image quality is large. Therefore, all observers were non-experts.

D. *Experimental method*

Observers evaluated the image quality of the test image and ranked the five smartphones by resolution. The best-worst method was used in the experiment. First, observers select the best (1st rank) and the worst (5th rank) smartphones from the five smartphones. Second, the next best (2nd rank) and the next worst (4th rank) smartphones are selected in the same way from the remaining three smartphones. The remaining smartphone was ranked 3rd.

Observers were instructed on the experimental procedure, the meaning of resolution and the point of evaluation. Explanation of the resolution was conducted using training images to make observers understand correctly. In addition, the observers were instructed not to consider the color, the brightness or noise of the image. When the observers purchase a smartphone, the viewing distance is different for each observer. Thus, the observers could freely adjust the viewing distance. After evaluation, we investigated points where the observers gazed to judge whether observers correctly evaluated differences in resolution.

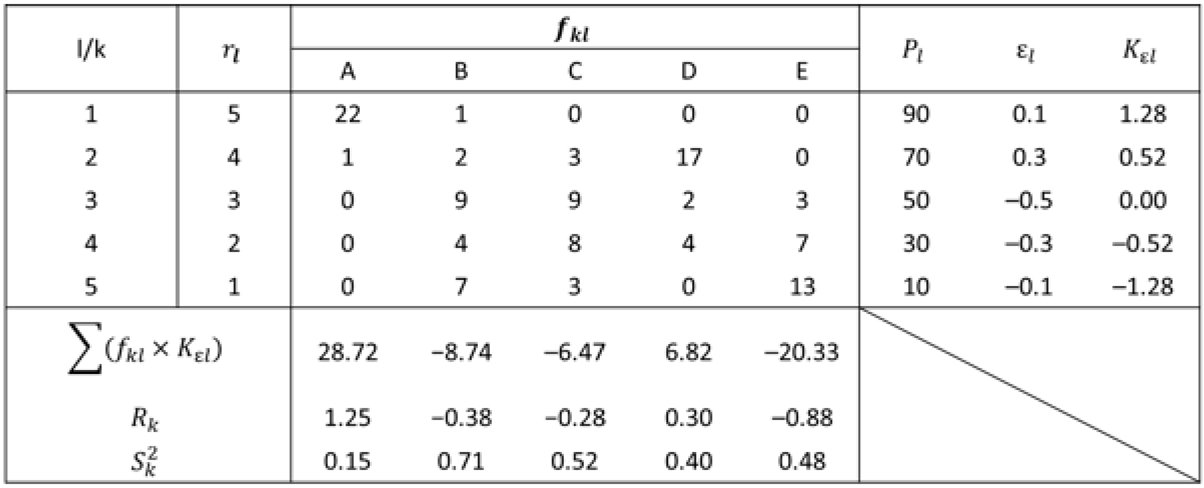
V. RESULTS

The assessment results were analyzed, and the presence or absence of significant differences was identified. The assessment results were quantified, and the average scores representing the image quality of each stimulus were calculated [14]. The calculation requires a normalized score $K_{\varepsilon l}$ which can be calculated using P_l and ε_l. P_l is the average of each segment of the range from 0 to 100 separated into the number of stimuli. In this experiment, the number of stimuli, i.e., the number of smartphones (n), equals 5. The value ε_l is the median of each segment of the standard normal distribution separated into n segments. $K_{\varepsilon l}$ is the percentile of the standard normal distribution. Thus, $K_{\varepsilon l}$ is the distance from the average of the standard normal distribution. The values of $K_{\varepsilon l}$ were given as a normalized score according to rank. The average scores of the total score are the evaluation values for each stimulus.

The aggregate results of “Map” (Figure 3(a)) are shown in Table 1. The rows represent rank, and the columns represent stimuli (smartphones A–E). The values of intersection (f_{kl}) are the number of observers for stimulus k for rank l. Thus, f_{1A} indicates that 22 observers ranked the smartphone with NLSP (smartphone A) 1st.

First, rank is converted to a value. The higher the ranking, the higher the r_l value of the smartphone, where r_l is calculated as follows:

TABLE I. ASSESSMENT RESULTS (Figure 3(a) Map)



l/k	r_l	f_{kl}					P_l	ε_l	$K_{\varepsilon l}$
		A	B	C	D	E			
1	5	22	1	0	0	0	90	0.1	1.28
2	4	1	2	3	17	0	70	0.3	0.52
3	3	0	9	9	2	3	50	−0.5	0.00
4	2	0	4	8	4	7	30	−0.3	−0.52
5	1	0	7	3	0	13	10	−0.1	−1.28
$\sum(f_{kl} \times K_{\varepsilon l})$		28.72	−8.74	−6.47	6.82	−20.33			
R_k		1.25	−0.38	−0.28	0.30	−0.88			
S_k^2		0.15	0.71	0.52	0.40	0.48			

$$r_l = \quad \mathrm{n} - \mathrm{l} + 1 \qquad (1)$$

The percentile values P_l are calculated using r_l as follows:

$$\mathrm{Pl} = \quad \frac{r_l - 0.5}{n} 100 \qquad (2)$$

The calculation results are shown in each row r_l, P_l of Table 1. Next, ε_l is calculated using (3) or (4). If the value of P_l is larger than 50, formula (3) is used. If the value of P_l is 50 or less, formula (4) is used. This is because the values of ε_l are calculated based on the point of the variance 0 of the standard normal distribution.

$$\varepsilon_l = \begin{cases} 1 - \dfrac{P_l}{100} & (P_l > 50) \qquad (3) \\ \dfrac{P_l}{100} & (P_l \leq 50) \qquad (4) \end{cases}$$

The calculation results are shown in row ε_l of Table 1. $K_{\varepsilon l}$ is calculated using ε_l from the normal distribution table. The values of $K_{\varepsilon l}$ shown in Table 1 were given to each stimulus according to the ranking. The average scores (R_l) of the total scores ($\sum(f_{kl} \times K_{\varepsilon l})$) are the evaluation values of the stimulus. For example, the average score R_A is calculated as follows: $R_A = 28.72/23 \fallingdotseq 1.25$. The average scores and total scores are shown in Table 1. The average scores of "Map" (Figure 3(a)) are shown in the yardstick graph in Figure 4. The horizontal axis indicates the average score. The marks on the axis (oval, triangle, square, rhombus, and x) indicate the average scores of each stimulus (smartphone A, smartphone B, smartphone C, smartphone D, and smartphone E, respectively). The higher the average score, the higher the evaluation. In Table 1, the average score of smartphone A is the highest, which indicates that smartphone A has the highest resolution.

A t-test was used to verify the significant difference between the stimuli. The variance of the average score (S_k^2) and the statistical quantity t_0 are calculated as follows:

$$S_k^2 = \quad \frac{\Sigma\{fk_l \times (K\varepsilon_l)^2\}}{\sqrt{\Sigma(fk_l)}} - R_k^2 \qquad (5)$$

$$t_0 = \frac{R_x - R_y}{\sqrt{\sum(f_{kl})\,(S_x^2 + S_y^2)}} \sqrt{\sum(f_{kl}) \sum\{(f_{kl}) - 1\}} \qquad (6)$$

The value $\sum(f_{kl})$ indicates the number of observers. x and y are stimuli. The calculation results are shown in Table 1. The values of t are calculated using the degree of freedom (DoF) from t distribution. In this experiment, the DoF is $\mathrm{DoF} = 2 * \sum(f_{kl}) - 2 = 46 - 2 = 44$. The t value of 1% significant level is $t_{1\%} = 2.414134$ and that corresponding to a 5% significant level is $t_{5\%} = 1.68023$. If the value of t_0 is larger than the value of $t_{5\%}$, there is a significant difference between stimuli.

Here, smartphone A is the highest, and smartphone D is the second highest. The t_0 value between smartphones A and D ($t_0(A, D)$) and the result of the t-test is as follows:

$$t_0(\mathrm{A}, \mathrm{D}) = 10.33 > \ t_{1\%} \qquad (7)$$

In (7), $t_0(A, D)$ is larger than $t_{1\%}$. This result indicates that smartphone A has a higher resolution than smartphone D and has a significance value of 1%. The results of the 3rd rank (smartphone C), 4th rank (smartphone B), and 5th rank (smartphone E) are as follows:

$$t_0(\mathrm{D}, \mathrm{C}) = 4.13 > \ t_{1\%} \qquad (8)$$

$$t_0(\mathrm{C}, \mathrm{B}) = 0.53 > \ t_{1\%} \qquad (9)$$

$$t_0(\mathrm{B}, \mathrm{E}) = 2.77 < \ t_{5\%} \qquad (10)$$

$t_0(D, C)$ and $t_0(C, B)$ are larger than $t_{1\%}$. Therefore, there are significant differences of 1% between smartphones D and C, and smartphones C and B. $t_0(B, E)$ is less than $t_{1\%}$ and $t_{5\%}$, which indicates that there is no significant difference between smartphones B and E. The arrows indicate significant differences in the graph in Figure 4. The asterisks represent the level of significant difference between stimuli. "**" represents a significant difference of 1%, and "*" represents a significant difference of 5%. The analysis results of images [b–e] are shown in Figure 3 (b–e). Smartphone A has the highest resolution and significant differences of 1% between other smartphones in all the images. On the other hand, smartphone E has the worst resolution in all images and significant differences for four out of five images with the other smartphones.

VI. Discussion

Smartphone A (with NLSP) has the highest score and a significant difference of 1% between the other smartphones (which are either without NLSP or from different manufacturers) in all the images. The results indicate that NLSP is valid for text on smartphone displays. The same results were obtained for all the images. Thus, NLSP is valid for images other than the five images used in this paper. There are significant differences between smartphones without NLSP. It is assumed that the results were influenced by the internal processing differences.

In this experiment, a gazing point was not specified for the observers. In addition, there are significant differences in all the images when all the observers are non-experts. From the above, there are clear differences of image quality between images with NLSP and those without NLSP.

VII. Conclusions

Subjective assessments using smartphones with NLSP and those without NLSP were conducted to verify the effectiveness of NLSP for texts. The results of experiments using five smartphones indicated that the image quality of a smartphone with NLSP is the highest, and there are significant differences between the other smartphones.

Statistical analyses indicate that the experimental results are reproducible. The conclusion that a smartphone with NLSP has the highest image quality was obtained for all images, therefore, both the assessment using the best-worst method and the analysis method in this experiment were valid as subjective assessment methods.

References

[1] S. Gohshi, "A new signal processing method for video: reproduce the frequency spectrum exceeding the Nyquist frequency," MMsys '12 Proceedings of the 3rd Multimedia Systems Conference, pp.47-52, Sep. 2012.

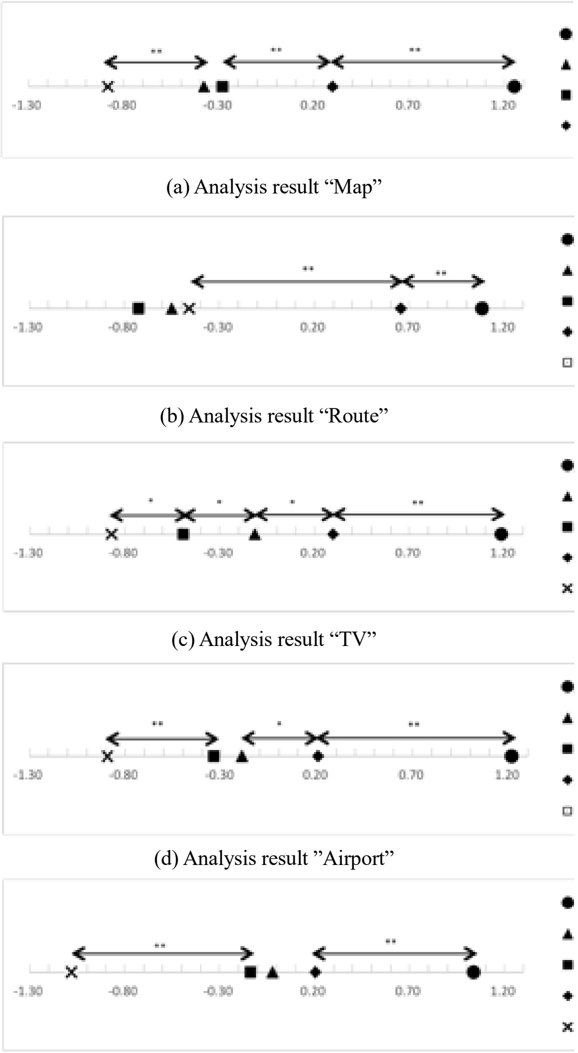
[2] http://www.fmworld.net/product/phone/f-02h/display.html?fmwfrom=f-02h_index. [retrieved: Mar. 2017]

[3] H. Shoji and S. Gohshi, "Subjective Assessment for Resolution Improvement on 4K TVs:Analysis of Learning-Based Super-Resolution and Non-Linear Signal Processing Techniques," The Eleventh International Multi-Conference on Computing in the Global Information Technology (ICCGI2016), pp.10-15, Nov. 2016.

[4] M. Sugie, S. Gohshi, H. Takehisa, and C. Mori, "Subjective Assessment of Super-Resolution 4K Video using Paired Comparison," , 2014 International

[5] Symposium on Intelligent Signal Processing and Communication Systems (ISPACS 2014), pp.42-47, Dec. 2014.

[6] C. Mori and S. Gohshi, "Image Quality of a Smartphone Display with Super-Resolution," Proceedings of the ISCIE International Symposium on Stochastic System Theory and its Applications, Vol.2016, pp.340-344, 2016.



(a) Analysis result "Map"

(b) Analysis result "Route"

(c) Analysis result "TV"

(d) Analysis result "Airport"

(e) Analysis result "Ticket"

Figure 4. Analysis results

[7] Rec. ITU-R BT.710-4, "Subjective Assessment Method for Image Quality in High-Definition Television," Nov, 1998.

[8] Manisha and P. Ahlawat, "Enhanced DNA Based Cryptography," International Journal of Advance Research in Computer Science and Management Studies, vol.3, Issue 6, pp.452-455, Jun. 2015.

[9] S. Gohshi, "A new signal processing method fo video: reproduce the frequency spectrum exceeding the Nyquist frequency," MMSys '12 Proceedings of the 3rd Multimedia Systems Conference, pp.47-52, Sep. 2012.

[10] S. Farsiu and M. Dirk Robinson, "Fast and Robust Multi-Frame Super-Resolution," IEEE Trans Image Process 2004, Vol.13, no.10, pp.1327-1344, Oct. 2004.

[11] S. Gohshi and I. Echizen, "Limitation of Super Resolution image Reconstruction," The Journal of The Institute of Image Information and Television Engineers, Vol.36, No.48, ME2012-125, pp.1-6, Nov. 2012.

[12] W.T. Freeman, T.R. Jones, and E.C. Pasztorm "Example-Based Super-Resolution," IEEE Computer Graphics and Applications, Vol.22, no.2, pp.56-65, Mar. 2002.

[13] M. Sugie, S. Gohshi, H. Takeshita, and C. Mori, “Subjective assessment of super-resolution 4K video using paired comparison,” 2014 International Symposium on Intelligent Signal Processing and communication Systems (ISPACS 2014), pp.17-22, Dec. 2014.

[14] H. Shoji and S. Gohshi, “Subjective Assessment for Learning-Based Super-Resolution and Non-Linear Signal Processing Techniques,” The Eleventh International Multi-Conference on Computing in the Global Information Technology (ICCGI 2016), pp.10-15, Nov. 2016.

[15] T. Fukuda and R. Fukuda, “Ergonomics handbook,” ISBN978-4-86079-036-3, Scientist press co.ltd, pp.41-71, Tokyo, 2009. (in Japanese)

Scalable Video Summarization based on Visual Attention Model

Amr Abozeid
Mathematics department,
Computer Science Division
Faculty of Science,
Al-Azhar University, Cairo, Egypt
email: aabozeid@azhar.edu.eg

Hesham Farouk
Computers and Systems Department,
Electronics Research Institute (ERI),
Cairo, Egypt
email: hesham@eri.sci.eg

Kamal ElDahshan
Mathematics department,
Computer Science Division
Faculty of Science,
Al-Azhar University, Cairo, Egypt
email: dahshan@gmail.com

***Abstract*— Scalable video coding and summarization are becoming important research fields in many adaptive video applications. In this paper, we propose a scalable video summarization framework based on visual attention model. This framework utilizes the scalable video coding to produce scalable summaries. Experiments have been conducted to prove the concept and to measure the time performance of the proposed framework. The results show that the proposed framework is an efficient and promising solution.**

Keywords-Scalable Video Coding; Summarization; Visual Attention Model; Video Processing.

I. Introduction

As a result of the dramatic growth of video creation, research fields, such as video summarization, browsing, adaptation, indexing, and retrieval have been hot topics of recent research. Video Summarization (VS) can be defined as the creation of compact video representation. This new representation can provide the user with brief information about the video content. The advantages of VS include but are not limited to, enhancing browsing, streaming, storage and quick retrieval of videos.

Video summarization is very important nowadays, especially in the context of mobile computing and to ubiquitous accessing needs. Farouk et al. [4] presented a comparative study of mobile video summarization techniques. The comparative study showed that building an adaptive VS approach is required for many applications. VS adaptation can be defined as the ability of automatically producing a summary content that meets the user's preferences and device capabilities.

The main target of Scalable Video Coding (SVC) is to produce one bit-stream that contains multiple layers. Each layer has a specific resolution, quality and frame rate. In SVC, the encoding process is performed once, while many layers (versions) can be extracted from the bit-stream, according to the specific needs (adaptation needs) [5][6].

Although video summarization is an extensively studied topic in the literature [1][3][4][7][8], previous researchers focus mainly on a single scale summary (single output summary). In some cases, producing one scale summary may be insufficient. A scalable video summary has a number of applications. These applications include video summary adaptation, progressive video access, video visualization and interactive video browsing. The SVC concepts can also be applied in the VS context as an additional attributes of the generated summaries. The following Scalable Video Summarization (SVS) modalities are possible:

1. Temporal scalability (keyframes number, duration): adapt keyframes number or duration length to meet the user's request.
2. Spatial scalability (frame size): the video summary is coded at multiple spatial resolutions.
3. Quality scalability: video summary is coded at a single spatial resolution with different qualities.
4. Hybrid scalability: a combination of the three scalability modalities described above.

This paper proposes a scalable video summarization framework based on VAM. This framework is summarized as follows.

1. Extract and partially decoded the base layer of scalable video. The main goal of this step is to reduce the computational cost of the following steps.
2. Feature Extractions: the goal of this step is to extract features (e.g., color, motion, etc.) from the pre-sampled frames. Then build feature based curve for each feature (e.g., color curve).
3. Attention Curve Construction: after the feature based curves are obtained separately, these curves need to be merged in a meaningful way to construct the final attention curve. The attention curve peaks indicate the corresponding video frames or segments which most likely attract user's attention.
4. Scalable keyframe selection: the base layer and enhanced layers video summary are extracted from the scalable video based on the attention curve values.

Initial experiments are conducted to prove the concept of this framework. The results show that the proposed framework is an efficient and promising solution. The rest of this paper is organized as follows. Section II introduces some related work about the SVS. The proposed framework is discussed in Section III. Section IV presents the experiments and the results of the proposed framework. Finally, Section V concludes the paper and suggests future work.

II. Related Work

The concept of SVS was first discussed in [9], where the scalable summary was introduced as a special set of embedded summaries. It presented a framework that consists of two main stages: analysis and generation, similar to SVC.

ISBN: 978-1-61208-611-8

The main objective is to analyze the video sequence once and generate many summaries with different lengths (analyzing once, generate many). During the analysis stage, the input video bitstream is divided into basic units called Group of Pictures (GoPs). A ranked list of these GoPs is built using the hierarchical clustering with average linkage and a ranking algorithm. Finally, the scalable summary is obtained at the generation stage depending upon the length (e.g., keyframes number, skim duration) requested by the user and/or the context. However, the analysis of the input video in [9] is not scaled to the size of the input. As a result, it fails to generate effective summaries for long duration videos due to the substantial increase in the computational cost associated with the analysis stage.

A framework based on sparse dictionary selection was proposed for scalable summarization of consumer (home) videos [10]. It formulates the video summarization problem as a dictionary selection problem. The video frames are considered as an original feature pool. An optimal subset is selected as "dictionary" from this pool under two constraints; sparsity and lower reconstruction error. Sparsity means the extracted dictionary should be as small as possible and selected from the original feature pool in a uniformly scattered way. Low reconstruction error means that the original video can be reconstructed with high accuracy using the selected dictionary (i.e., the selected dictionary is the most representative frame sets). This framework is designed to extract a scalable key frame and/or a video skimming. In contrast to most existing methods, this framework allows users to choose different numbers of keyframes without incurring an additional computational cost.

Etezadifar et al. [11] proposed a new method to improve the performance of the framework proposed in [10]. In this method, VS is performed as a selection and a training sparse dictionary problem simultaneously. Thus, the dictionary selection and learning were iteratively performed. Each iteration is performed independently and the obtained response is replaced by its previous value.

Panda et al. [12] introduced an SVS framework for both the analysis and the generation stages according to the summary length determined by the user. This framework consists of a 3-step analysis stage followed by a 1-step generation stage:

1. The Video Similarity Graph (VSG) is constructed from the input video frames based upon the color feature. Each frame is represented by a 256-dimensional feature vector obtained from the color histogram using the HSV color space. Then, VSG is constructed as a weighted complete graph, where each frame is represented as a vertex. Then the skeleton graph is extracted by choosing the vertices whose degrees are higher than a certain threshold from VSG (i.e., reduce the size of the VSG).
2. A Minimum Spanning Tree (MST) based clustering is used over the skeleton graph to obtain the initial clusters.
3. The initial clusters are propagated using a random walker algorithm [13] to obtain the final clusters.
4. The keyframes (frames that are closest to the centroids of each cluster) are extracted and arranged according to the cluster significance factor.

Perez et al. [14] proposed an SVS approach which provides different levels and views of summary details. This approach is based on the data cube On-Line Analytical Processing (OLAP) operations [15]. The data cube concept has been proposed to facilitate user's navigation through multidimensional space where each move corresponds to a query using some combination of the dimensions. In this work, different audio-visual descriptors are considered. This allows the data cube partitioning in a multimodal audio-visual descriptor space. This approach was designed to process the cultural video document only.

Based on the MPEG-DASH standard [16], a Context-aware Video Summarization and Streaming (CVSS) approach was proposed in [17][18]. The CVSS was proposed to provide an adaptable video streaming for mobile devices, especially, when there are limitations in the available time or mobile energy level. The CVSS consists of 3 phases:

1. The input video is converted into an MPEG-DASH compatible format.
2. A semantic attention value for each segment of the MPEG-DASH video is computed based on VAM. Then, based on the segment attention value, dynamic video summaries are generated with different durations to meet the user’s request.
3. Finally, the video summary is adapted during the streaming session in order to be suitable for the available devices and network contexts.

The main observation is: scalability in video summarization is usually related to the summary length (temporal). However, the SVS concept should be extended to other scalability modalities (e.g., quality, spatial). This may be used to adapt the output summaries to targeted contexts (e.g., device context, network context).

III. Scalable video summarization framework

The proposed framwork consists of an (3-step) analysis stage followed by a (1-step) generation stage. Figure 1 is a block diagram shows all the four steps. The description of these steps are as follows:

A. *Extract and partially decoded*

In this step, the following tasks were implemented.

a. Extract the base layer (layer number 0) from the input scalable video by JVSM BitStreamExtractorStatic function. As shown in Table I, the base layer (layer number 0) has the minimum configurations and this will significantly reduce the computations.
b. Partially decode and extract the base layer frames.

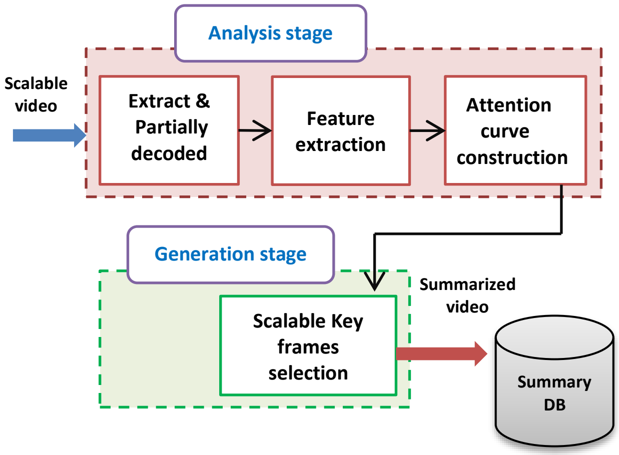


Figure 1. A block diagram of the proposed framework

TABLE I. SCALABLE VIDEO LAYERS

Layer	Resolution	Frame rate	Bitrate	MinBitrate
0	160x 96	1.8750	26.20	26.20
1	160x 96	3.7500	33.00	33.00
2	160x 96	7.5000	40.70	40.70
3	160x 96	15.0000	49.10	49.10
4	160x 96	30.0000	57.20	57.20
5	320x192	1.8750	129.00	129.00
6	320x192	3.7500	165.50	165.50
7	320x192	7.5000	208.10	208.10
8	320x192	15.0000	252.90	252.90
9	320x192	30.0000	288.90	288.90
10	640x368	1.8750	462.20	462.20
11	640x368	3.7500	625.90	625.90
12	640x368	7.5000	843.70	843.70
13	640x368	15.0000	1097.00	1097.00
14	640x368	30.0000	1303.00	1303.00

B. Feature extractions

In this step, color and motion features are extracted. Therefore, this step consists of two sub-steps: Static Attention Curve Extraction and Motion Attention Curve Extraction. These two sub-steps are adapted from [17] and briefly discussed in the next subsections.

1) Static Attention Curve Extraction

In this step, the static attention curve is extracted from the video frames based on the color feature. As shown in Figure 2, the curve describes the video contents by representing the important frames corresponding its peeks. The horizontal parts of the curve mean that the corresponding frames having the same attended areas probability and almost contain the same information. The gradual changes in the curve mean that there is a gradual difference in the content of the corresponding frames. On the other hand, sudden changes in the curve mean that there is a significant difference in the content of the corresponding frames.

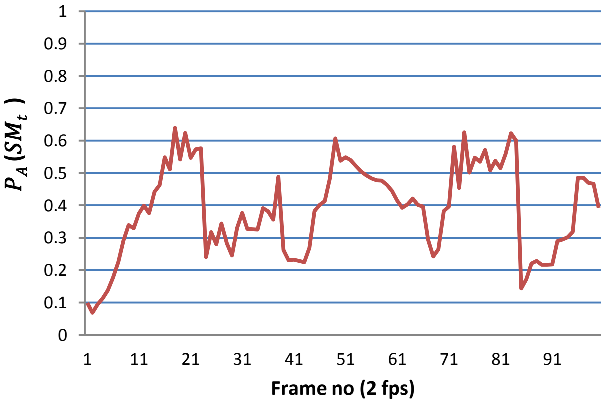


Figure 2. Static attention curve of "Big Buck Bunny" video

The Static Attention Detection Algorithm (see Figure 3) can be summarized as:

a. The Saliency Map (SM) is computed for each frame. A saliency map is a gray image, which contains attended/salient areas (bright areas) and unattended/non-salient areas (dark areas). The attended areas usually attract the user attention.

b. The attended areas of the saliency map are extracted as:

   i. Each SM is divided into non-overlapping Macro-blocks (MB).

   ii. Accordingly, each SM is represented by two sets (A and U). The set A is the set of all non-overlapping attended blocks (areas). Similarly, U is the set of all non-overlapping unattended blocks (areas).

c. After normalizing the value of the $P_A(SM_t)$ for each frame to [0, 1], a static attention curve (SC) is obtained.

```
Input: $F_t$ // the input frame at a time $t$
Output: $P_A(SM_t)$// the probability of attended areas A in $SM_t$
Start
Initialize $A = U = \emptyset$
Compute $SM_t$ for $F_t$
Loop for each $MB_{i,j}$ in the $SM_t$
    If $(C(MB_{i,j}) \geq \varepsilon^{SM})$ then
        Add $MB_{i,j}$ to the attended set A
    Else
        Add $MB_{i,j}$ to the unattended set U
    End loop
$P_A(SM_t) = \frac{|A|}{|A|+|U|}$
End
```

Figure 3. Static Attention Detection Algorithm

2) Motion Attention Curve Extraction

Most of the video summarization approaches are based on motion feature in different ways [19][20][21][22]. In this step, we adopt the Fast Directional Motion Intensity Estimation (FDMIE) algorithm. FDMIE aims to detect the

Motion Intensity (MI) between the consecutive frames. The following two options decrease the complexity of FDMIE.

a. The motion intensity estimation has been applied to the regions in each frame that could potentially attract users' attention due to the motion (i.e., attended areas).

b. The Sum of Absolute Differences (SAD) is used to determine the matching between two blocks. The SAD is more used because it has a higher-quality precision and involves lower computational cost [23][24].

According to the FDMIE algorithm (Figure 4), the motion intensity between the saliency maps is computed.

a. For each block in SM_{t-1} (saliency map extracted from the t-1 frame), FDMIE computes the current minimum (C_{MIN}) distortion between this block and the corresponding block in SM_t by SAD.

b. The FDMIE searches the eight directions around the target block uses the One-at-a-Time Search (OTS) strategy. In OTS, the block-by-block search along a direction is continued if a newly searched block has lower distortion than the previously searched block. Otherwise, the search in that direction stops. The minimum distortion found in each directional search is set as a directional minimum (D_{MIN}) distortion.

c. Then, the Relative Distortion Ratio (RDR) is computed by dividing D_{MIN} by C_{MIN}.

d. If RDR between current D_{MIN} and C_{MIN} is lower than ε^D then other directional searches will be skipped and a final position of the block with C_{MIN} value become the position of the block with D_{MIN} value.

e. Otherwise, other directional searches will be started.

f. After a search round is completed, the lowest distortion among the D_{MIN}s (if found) is set as C_{MIN} and the next search round starts at the block with C_{MIN}.

g. Finally, the motion intensity $MI(P_{i,j})$ of the target block $P_{i,j}$ is computed as the distance between the $P_{i,j}$ position and the position of the block with final C_{MIN} value. Consequently, the motion intensity MI_{t-1} between the saliency SM_{t-1} and SM_t is computed as in (1).

$$MI_{t-1} = \sum_i^w \sum_j^h MI(P_{i,j}), \quad P_{i,j} \in SM_{t-1} \qquad (1)$$

The range of threshold ε^D is [0, 1] and it used to control the FDMIE convergence speed of the algorithm. The higher threshold ε^D will speed up the convergence of the FDMIE, but it will also decrease the prediction quality. For example, if ε^D is set at 0.5 implies that the prediction quality is less than 50% and the number of search blocks is reduced to the half. Initially, the select value of $\varepsilon^D = 0.5$. In future, more experimental studies will be conducted o determine the best value of ε^D .

```
Input: $A_{t-1}, A_t, SM_{t-1}, SM_t$
Output: $MI_{t-1}$
Start
For each $P \in SM_{t-1},\ P \in A_{t-1}$
    1. Initialize flag=false
    2. Compute $C_{MIN} = SAD(P_{i,j}, Q_{i,j})$
    3. For each 8 directions  around the point with $C_{MIN}$
        a. Compute $D_{MIN} = SAD(P_{i,j}, Q_{i+di,j+dj})$
        b. If  $D_{MIN} < C_{MIN}$
            If  $RDR(D_{MIN},\ C_{MIN}) < \varepsilon^D$
                Then $C_{MIN} = D_{MIN}$ and go to step 5.
            Else flag = true
            End for
    4. If flag = true then $D_{MIN}$s are compared. The lowest
       one is set as $C_{MIN}$ and update the corresponding
       position, go to step 1.
    5. Compute $MI(P_{i,j})$ and it to $MI_{t-1}$
End For
Return $MI_{t-1}$
End
```

Figure 4. FDMIE Algorithm

The FDMIE output is a numeric value that represents the motion intensity of the frame F_{t-1} . After normalizing the motion intensity value for each frame to [0, 1] a motion attention curve (MC) is obtained. Figure 5 shows motion attention curve of "Big Buck Bunny" video.

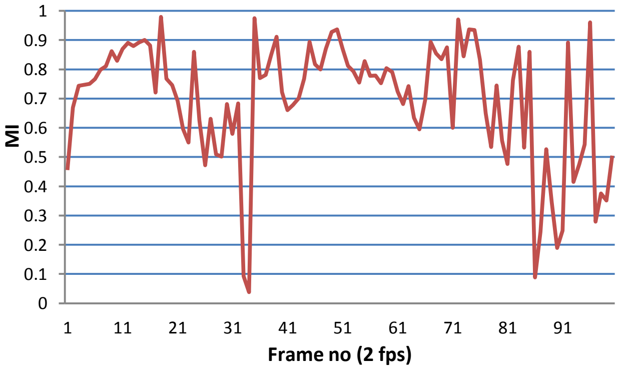


Figure 5. Motion attention curve of "Big Buck Bunny" video

C. Attention Curve Construction

After the static and the motion curves are obtained separately, these curves are merged to construct the final Attention Curve (AV). Figure 6 shows an example of the final attention curve that has been created by the proposed framework. The AV peaks indicate the corresponding video frames most important and usually attract user’s attention. The final attention curve was constructed as in (2).

$$AC = w_s \times SC + w_m \times MC \qquad (2)$$

Where SC represents static attention curve and MC represents motion attention curve. The weight values w_s and w_m are used for a linear combination which satisfies the two conditions:

$$w_s , w_m \geq 0 \quad \text{and} \qquad w_s + w_m = 1$$

In the experimental phase, we will conduct experiments to determine the best values for w_s and w_m. Initially, we determine $w_s = 0.4$ and $w_m = 0.6$.

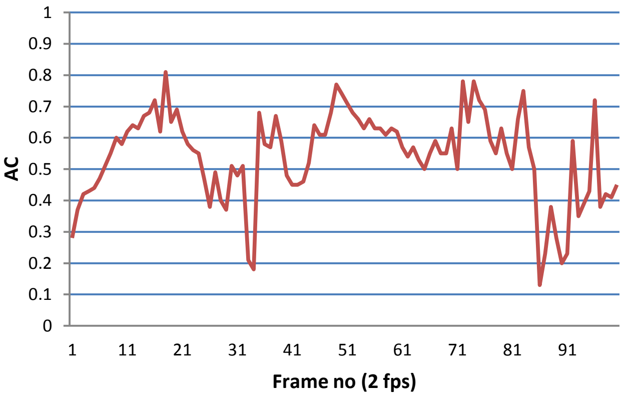


Figure 6. Attention curve of "Big Buck Bunny" video

D. *Scalable keyframes selection*

In the generation stage, the user can determine the number of keyframes to be generated (i.e., temporal scalable summary). Based on the attention curve, the frames with high attention values are selected to form the base-layer video summary (base-layer keyframes). The enhanced summary layers are constructed by selecting the corresponding base-layer keyframes form the enhanced scalable video layers. Finally, the selected scalable key-frames (scalable video summary) are saved in the summary folder.

IV. Experimental Results

A prototype was implemented as web application using a J2EE (JSP and Servlet) technologies. Some third-party libraries are used during the implementations such as FFmpeg [25] and JVSM [26]. Most of the standard video formats (e.g. mp4, flv, avi, etc.) are supported by this prototype in addition to .264 (scalable format).

Table II describes the selected data set videos from the YouTube. All videos are transcoded from in H.264/AVC (.mp4) format to JVSM [26] scalable format (.264) with different layers as described in Table I. The efficiency of the proposed framework is evaluated by comparing the summarization time for both .mp4 and .264 formats.

The experiments were carried out on a PC equipped with an Intel Core i7 and 8 GB of RAM. The experiments are carried out on all videos mentioned in Table II. The results of these experiments are organized in Table III. For each video, we record the processed frames (total number of the processed frames) and Analysis Time (AT) of these frames. Then compute a number of Processed Frames Per Second (PFPS) by dividing the processed frames by AT.

TABLE II. DESCRIPTION OF THE DATASET

Video name	**Duration**	**FPS**	**Frames #**	**Resolution (W× H) (pixels)**
Big Buck Bunny	00:09:56	24	14304	640×360
Tears Of Steel	00:12:14	24	17616	640×360
Of Forests and Men	00:07:33	24	10872	640×360
Beautiful Birds	00:10:46	30	19380	640×360
Bird feeding babies	00:10:59	30	19770	640×360
Bird noises sounds	00:17:55	30	32250	640×360
Final Repechages	00:14:14	25	21350	640×360
High Jump	00:12:49	25	21350	640×360
Land Rover Discovery	00:07:14	29	12586	640×360
Sofia2	00:16:26	25	24650	640×360
Solar System	00:09:28	29	16472	640×360
Strawberry	00:20:59	25	31475	640×360

As shown in Table III, the AT of .264 videos is less than .mp4 videos. The AT includes the feature extraction and motion attention curve extraction. In case of scalable videos, the proposed framework can process an average of 325.7 fps. For other video formats, the proposed approach can process an average of 299.4 fps. It is important to note that those results depend on the computational power of the target environment.

TABLE III. THE PROPOSED FRAMEWORK EFFICIENCY EVALUATION

Video name	**MP4**			**Scalable videos (.264)**		
	Processed frames	***AT (s)***	***PFPS***	***Processed frames***	***AT (s)***	***PFPS***
Big Buck Bunny	1194	3.3	357.4	895	2.2	407.4
Tears Of Steel	2937	7.0	416.7	1102	4.0	274.9
Of Forests and Men	448	3.1	145.3	709	3.4	209.9
Beautiful Birds	1294	8.4	153.2	1213	3.2	376.4
Bird feeding babies	1320	3.5	375.4	1237	3.5	351.4
Bird noises sounds	1570	6.4	247.2	2018	7.3	276.4
Final Repechages	1710	6.6	257.3	1336	5.9	226.8
High Jump	1540	4.3	360.6	1203	3.2	373.7
Land Rover Discovery	870	2.3	379.7	814	1.9	418.3
Sofia2	1974	9.9	198.7	1542	5.3	289.1
Solar System	1138	2.8	413.4	1066	2.4	440.5
Strawberry	1857	6.5	287.4	1968	7.5	263.7
Average	**1487.7**	**5.3**	**299.4**	**1258.6**	**4.2**	**325.7**

V. Conclusions

In this paper, we propose a scalable video summarization framework based on Visual Attention Model (VAM). In this framework, the concept of SVC was extended to the video summarization context. Therefore, the input scalable video will be analyzed once and then generate many summaries with different scalability modalities (such as temporal, quality and/or spatial). VAM is applied to extract the semantic meaning of the low-level video features (color and motion). We carried out experiments to measure the efficiency of the proposed framework. The results show that the proposed framework is an efficient and promising solution. In the future, we intend to conduct more experiments and improve the proposed framework.

Acknowledgments

This research is funded by the Science and Technology Development Fund (STDF), Egypt, project titled “Efficient Media Digital Library Design and Implementation of Summarized Video based on Scalable Video Coding for H.264 (MDLSS)”, (project ID 15269).

References

[1] M. Ajmal, M. H. Ashraf, M. Shakir, Y. Abbas, and F. A. Shah, “Video summarization: techniques and classification,” Computer Vision and Graphics, Springer, vol. 7594, 2012, pp. 1-13, doi: 10.1007/978-3-642-33564-8_1.

[2] Z. Xiong, R. Radhakrishnan, A. Divakaran, Y. Rui, and T. S. Huang, *A unified framework for video summarization, browsing & retrieval: with applications to consumer and surveillance video*: Academic Press, 2006.

[3] B. T. Truong and S. Venkatesh, “Video abstraction: A systematic review and classification,” ACM Transactions on Multimedia Computing, Communications, and Applications (TOMCCAP), vol. 3, no. 1, 2007, pp. 37, doi: 10.1145/1198302.1198305.

[4] H. Farouk, K. ElDahshan, and A. Abozeid, “The State of the Art of Video Summarization for Mobile Devices: Review Article,” Graphics, Vision and Image Processing GVIP, vol. 14, no. 2, 2014, pp. 37-50.

[5] H. Schwarz, D. Marpe, and T. Wiegand, “Overview of the scalable video coding extension of the H. 264/AVC standard,” IEEE Transactions on circuits and systems for video technology, vol. 17, no. 9, 2007, pp. 1103-1120.

[6] J. M. Boyce, Y. Ye, J. Chen, and A. K. Ramasubramonian, “Overview of SHVC: Scalable Extensions of the High Efficiency Video Coding Standard,” IEEE Transactions on Circuits and Systems for Video Technology, vol. 26, no. 1, 2016, pp. 20-34.

[7] A. G. Money and H. Agius, “Video summarisation: A conceptual framework and survey of the state of the art,” Journal of Visual Communication and Image Representation, vol. 19, no. 2, 2008, pp. 121-143, doi: 10.1016/j.jvcir.2007.04.002.

[8] R. Pal, A. Ghosh, and S. K. Pal, "Video Summarization and Significance of Content: A Review," *Handbook on Soft Computing for Video Surveillance*, pp. 79-102: CRC Press, 2012.

[9] L. Herranz and J. M. Martinez, “A framework for scalable summarization of video,” IEEE Transactions on Circuits and Systems for Video Technology, vol. 20, no. 9, 2010, pp. 1265-1270.

[10] Y. Cong, J. Yuan, and J. Luo, “Towards scalable summarization of consumer videos via sparse dictionary selection,” IEEE Transactions on Multimedia, vol. 14, no. 1, 2012, pp. 66-75.

[11] P. Etezadifar and H. Farsi, “Scalable video summarization via sparse dictionary learning and selection simultaneously,” Multimedia Tools and Applications, 2016, pp. 1-25.

[12] R. Panda, S. K. Kuanar, and A. S. Chowdhury, "Scalable Video Summarization Using Skeleton Graph and Random Walk." Year, pp. 3481-3486, doi.

[13] N. Paragios, Y. Chen, and O. Faugeras, *Handbook of mathematical models in computer vision*: Springer Science & Business Media, 2006.

[14] K. R. Perez-Daniel, M. N. Miyatake, J. Benois-Pineau, S. Maabout, and G. Sargent, "Scalable video summarization of cultural video documents in cross-media space based on data cube approach." Year, pp. 1-6, doi.

[15] J. Gray *et al.*, “Data cube: A relational aggregation operator generalizing group-by, cross-tab, and sub-totals,” Data mining and knowledge discovery, vol. 1, no. 1, 1997, pp. 29-53.

[16] "MPEG-DASH Standard," [retrieved: 12, 2017], http://mpeg.chiariglione.org/standards/mpeg-dash.

[17] H. Farouk, Kamal A. ElDahshan, and A. Abozeid, “Effective and Efficient Video Summarization Approach for Mobile Devices,” International Journal of Interactive Mobile Technologies (iJIM), vol. 10, no. 1, 2016, pp. 19-26.

[18] H. Farouk, K. El Dahshan, and A. Abozeid, “Context-Aware Joint Video Summarization and Streaming (CVSS) Approach,” IEEE International Symposium on Multimedia (ISM), 2016, pp. 597-602.

[19] N. Ejaz, I. Mehmood, and S. W. Baik, “Efficient visual attention based framework for extracting keyframes from videos,” Signal Processing: Image Communication, vol. 28, no. 1, 2013, pp. 34-44.

[20] A. B. Mejía-Ocaña *et al.*, “Low-complexity motion-based saliency map estimation for perceptual video coding,” in 2nd National Conference on Telecommunications (CONATEL), 2011, pp. 1-6.

[21] J.-L. Lai and Y. Yi, “Key frame extraction based on visual attention model,” Journal of Visual Communication and Image Representation, vol. 23, no. 1, 2012, pp. 114-125.

[22] N. Ejaz, I. Mehmood, and S. W. Baik, “Feature aggregation based visual attention model for video summarization,” Computers & Electrical Engineering, vol. 40, no. 3, 2014, pp. 993-1005.

[23] Q. Yang, C. LI, and Z. LI, “Motion Navigation System Estimation Algorithm in Mobile Phone Video Learning System,” Journal of Computational Information Systems, vol. 10, no. 16, 2014, pp. 7187-7194.

[24] M. Santamaria and M. Trujillo, “A comparison of block-matching motion estimation algorithms,” in 7th Colombian Computing Congress (CCC), 2012, pp. 1-6.

[25] "FFmpeg," [retrieved: 12, 2017], https://www.ffmpeg.org/.

[26] "JSVM Reference Software," [retrieved: 12, 2017], https://www.hhi.fraunhofer.de/en/departments/vca/research-groups/image-video-coding/research-topics/svc-extension-of-h264avc/jsvm-reference-software.html.

A Sonification Method using Human Body Movements

Felix Albu
Departments of Electronics
Valahia University of Targoviste
Targoviste, Romania
email: felix.albu@valahia.ro

Mihaela Nicolau, Felix Pirvan, Daniela Hagiescu
R&D Department
Advanced Slisys SRL
Bucharest, Romania
email: aslisys.121@gmail.com

***Abstract*—In this paper, we propose a new method to generate piano music starting from body movements recorded with a webcam. The joint coordinates found by using the convolutional pose machine method are used to calculate the pitch and velocity of the produced consonant or dissonant chords.**

Keywords- Sonification; generative music; convolutional pose machine; computer vision.

I. INTRODUCTION

In principle, distinct information or data sets in various fields, such as geology, medical research, or financial markets, can be perceived using the ears by sonification instead of studying large tables or graphics [1]. The principle underlying all sonication techniques is the arbitrary mapping of input data in the auditory domain.

One of the first sonification systems was the Unité Polyagogique Informatique CEMAMu (UPIC) system [6] that generates complex sounds by writing on a screen with a digital pen. Similar systems have been implemented in the AudioSculpt [7] and SPEAR software [8]. These applications reshape modified images in sounds. Other software solutions are EyesWeb [9], and Max/MSP/Jitter [10]. The available technology has allowed the study of sounds without being limited to a system based on symbols, such as the Western music notation [1].

Among first attempts to generate sounds from body movement were those made by using sensors attached to the body or extracting movement from video recordings [1]-[5]. A relevant information can be obtained from the spatial or temporal content of the movement. It is known that the movement extraction using camcorders or webcams are less precise than those using body-attached sensors. On the other hand, this approach is much cheaper, simpler, less obtrusive and offers the ability to make recordings in various locations. Recent approaches of using motion detection to control the real-time sound generation are the Motion Composer [12] or Point Motion Control [13] devices. In [14], an application that translates the perceived movement of the scenery such as passing trains, into music is presented. The continuously changing landscape view outside of the train window was captured with a camera and translated into Musical Instrument Digital Interface (MIDI) events that are replayed instantaneously. A sonification method that incorporates both spatial information and color distribution properties of captured video frames was implemented in [14]. Unfortunately, this method cannot be adapted for body movements because the background is typically static.

One of the most promising techniques for creating music based on the movement of the human body uses "motiongrams" [5]. A "motiongram" is a visual representation of movement based on the difference between successive frames. The visual similarity between motiongrams and spectrograms is exploited by transforming motiongrams into sounds through an inverse Fast Fourier Transform (FFT) [5]. Therefore, an image is treated as the spectrogram with frequency information on the Y axis and time on the X axis as the basis for synthesizing a sound file. Unfortunately, the method is too complex, involving multiple FFT processing [5].

Another promising sonification approach based on heavily numerically complex optical flow computation has been presented in [15]. A musical note is played when a local peak in the optical flow magnitude is higher than a threshold. The pitch corresponds to the location and flow direction of the peak and the velocity (or intensity) of the note corresponds to the magnitude [15]. A fish bowl was filmed and consonant chords were generated when fish were near one another and moved in approximately the same direction [15]. In a proposed alternative for static images the Hue, Saturation and Value (HSV) color space was used.

In this paper, we propose a novel and computationally simpler method based on computer vision techniques that uses the body joint coordinates found by Convolutional Pose Machines (CPM) from [16] and a modified sonification method. To the best of our knowledge, the CPM method has not been previously used for sonification of captured body movements. Another original part of this work is the approach to use computed joint coordinates, avoiding outliers and generate aesthetically pleasing piano sound starting from a layout of pitches using low-level harmonic notions proposed in [15]. Also, the proposed method does not use the HSV color space or optical flow computation methods.

The rest of this paper includes Section II that describes the proposed method, while the acknowledgement, conclusions and future work close the article.

II. The Proposed Method

The scheme of the proposed method is shown in Figure 1.

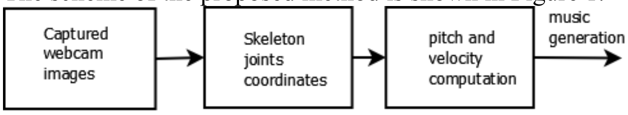


Figure 1. The scheme of the proposed method

A. *Skeleton joints coordinates computation*

The skeleton joints coordinates are found by using the CPM [16] on captured webcam images (see Figure 2). The image from Figure 2 has a width of 491 pixels and a height of 368 pixels.

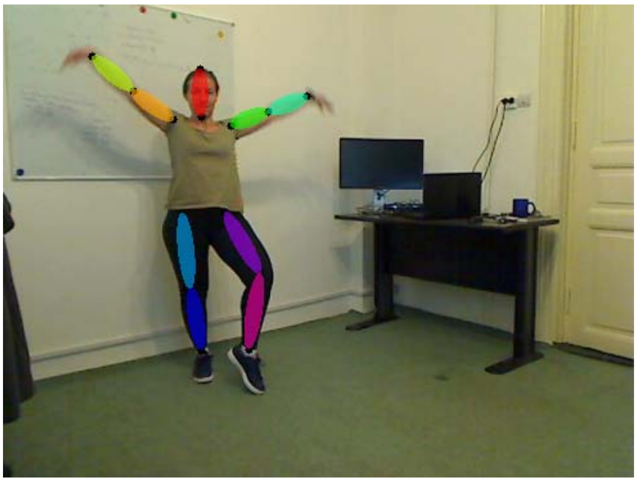


Figure 2. A captured image with overimposed computed joints coordinates.

The convolutional pose machine is a human pose detector. The algorithm produces the coordinates for the following 14 human skeleton joints: head, upper neck, shoulders, elbows, wrists, hips, knees and ankles. An example of skeleton joints is shown in Figure 3. The CPM uses two convolutional neural networks, one to detect the persons present in an image and the other to detect person's skeleton joints. The networks were trained on several public datasets [16]. Each network is composed of a sequence of several stages. Each stage produces belief maps that are supervised within each stage, thus addressing the vanishing gradients problem, inherent to deep neural networks. Each stage is composed of a sequence of convolution and pooling layers. The convolutions capture local features of the size of the convolution kernel (5x5, 9x9, and 11x11 kernels are used), while the pooling layers downscale the image by a factor of 2. The effect of pooling is that the subsequent convolution will operate on a less-detailed version of the image, capturing features on a bigger scale.

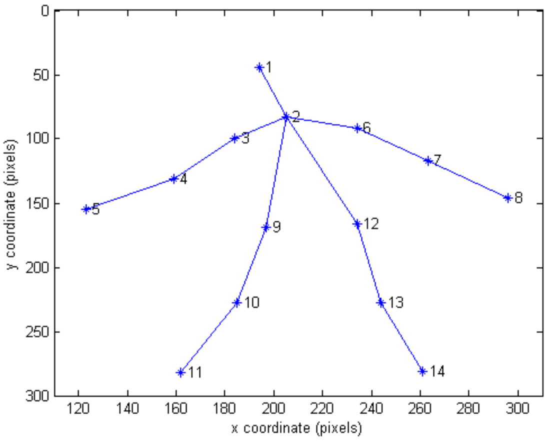


Figure 3. The computed skeleton joints.

The CPM method achieved state of the art accuracy on all primary benchmarks [16] and does not need special expensive equipment like Kinect devices. It has been reported that the CPM has some failure cases when multiple people are in close proximity [16]. However, this is not our case, since only one person is expected to dance in front of the camera. Therefore, the proposed framework is salient enough for the proposed model. More details about the CPM method can be found in [16].

Figure 4 shows the normalized vectors of y-coordinates evolution in time for various body joints. The blue curve shows the head coordinates, the red curve shows the left shoulder coordinates, and the green curve shows the right knee coordinates.

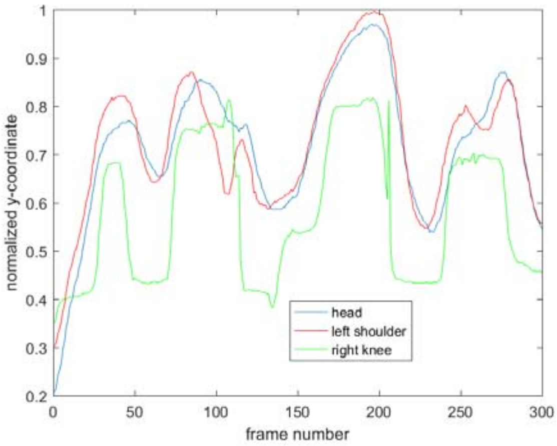


Figure 4. Examples of vectors obtained from the coordinates of the various body joints.

As expected, there is a rather close similarity between the normalized vectors of head and shoulder due to the physical constraints of human body joint movements. It is also obvious that the correlation between the knee coordinates and head/shoulder coordinates is much lower.

The CPM method is able to handle non-standard poses and solve ambiguities between symmetric parts for a variety of different relative camera views [16]. However, there are few failure cases that can appear when there is a sudden change of coordinates (see the green curve from Figure 4). These outlier coordinates can be removed by using the Dynamic Time Warping (DTW) distance between consecutive normalized joint coordinates vectors [17]. The DTW algorithm is a well-known algorithm that computes the optimal alignment between two time-series. It has been used in many applications such as speech recognition [17], handwritten evaluation [18][19], etc. In the classical DTW algorithm, a two-dimensional cost matrix is formed and its elements are the minimum accumulated distances for the sequences time series. More information about the DTW algorithm can be found in [17]. In Figure 5, the histogram of the DTW distances between consecutive vectors of coordinates for a dancing performance is shown. The elements of the vector containing the DTW distances are sorted into 64 equally spaced bins between its minimum and maximum values. It can be easily seen from Figure 5 that most of the time the DTW distances between the coordinates computed from consecutive frames are rather small. This is expected, because in most dance movements, there is not a very fast variability of joints positions in time. It can be noticed that a threshold set to 0.05 can reasonably detect outliers. If the DTW distance between two consecutive vectors is higher than 0.05, the particular vector is not taken into account in order to generate music. Generally, about 5-10% of vectors are ignored and usually these vectors are generated by faulty coordinates provided by the CPM block. The DTW distance was preferred to the Euclidian distance due to its better clustering properties and robustness to outliers.

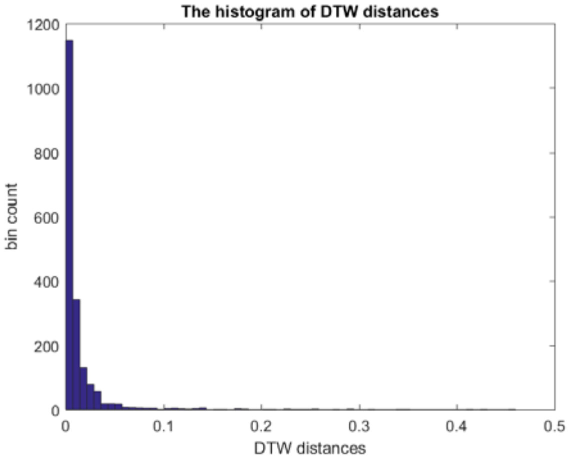


Figure 5. The histogram of DTW distances between consecutive vectors of coordinates.

B. Pitch and velocity computation

The pitch mesh pairs method from [15] is adapted for our approach. The selected normalized joints coordinates are used instead of the pixel values in the RGB or HSV color spaces as proposed in [15]. The pitch mesh consisted on vertically stacked pitch chains in parallel octaves [15]. The musical tone is generated with the "dominance" bit and two real numbers, pitch and register respectively. By increasing the range of pitch, *x*, the amount of chromaticism and dissonance is increased, while increasing the register, *y*, the broadness of the register of the generated notes is increased [15]. There is a correlation between the coordinates vectors and if they move in the same direction, the generated sound is basically consonant, while changes trigger functional changes in harmony. The chromaticism of the generated music is altered by modifying the thresholds for the normalized coordinates [15]. With a very low range of *x*, the notes may all emerge from the same tetrachord, whereas with a very high range, the piece could sound fundamentally atonal [15].

An example of generated pitch and velocity values for 50 frames is shown in Figure 6. The threshold and beat duration were set to 0.5, the pitch range was 7 and the register range was set to 3. The pitch and velocity values were scaled from -1 to 127. The size of the pitch and velocity vectors depends on the result of comparison with the threshold (e.g., it is 896 for Figure 6).

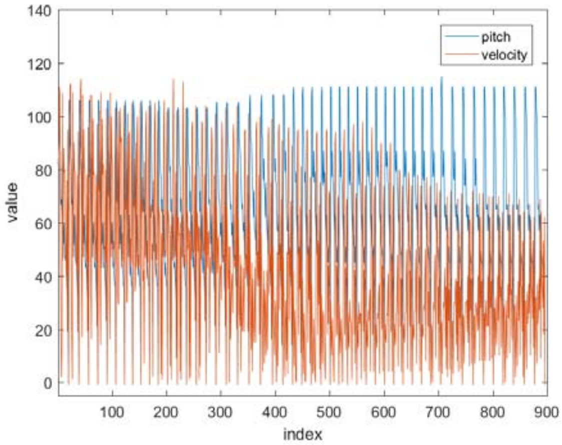


Figure 6. Computed pitch and velocity values.

The generated music contained many times two notes on the same pitch within a distance of each other. Therefore, in this case, the abruptly repeated notes were removed by retaining the note with the higher velocity. The effect of the distance on the filtered pitch and velocity parameters can be seen on Figures 7 and 8. The distance parameter was set to 3 for Figure 7 and 9 for Figure 8, respectively.

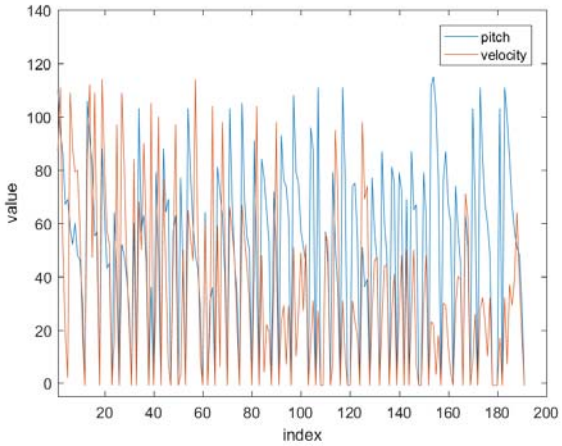


Figure 7. Filtered pitch and velocity for a distance of 3.

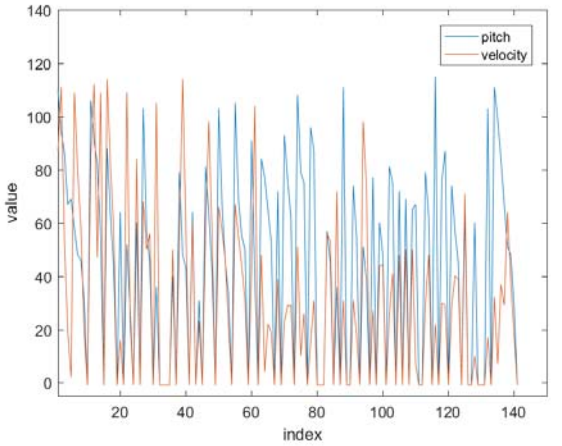


Figure 8. Filtered pitch and velocity for a distance of 9.

The size of the pitch and velocity vectors is reduced a lot (e.g., it is 191 for Figure 7 and 141 for Figure 8). Although the filtered parameters for a distance of 9 seems to be shrinked version of those obtained using a distance of 3, we've found by listening the generated music that a distance of 3 gives slightly more aesthetically pleasing sounds. These filtered parameters were used to generate the piano music in Matlab by employing the Microsoft MIDI Mapper [20] as the midi output device.

The complexity of using the human pose detector based on CPM is much smaller than that of computing the optical flow on successive frames. Also, our proposed sonification approach uses only 14 joints coordinates per frame and simple mathematical operations and comparisons for music generation. The complexity of the methods proposed in [5] [14] and [15] is at least two orders of magnitude higher because they use very numerically intensive and complex operations on frames with full or low resolution.

The pitch and velocity parameters computed using the method from [15] are shown in Figure 9.

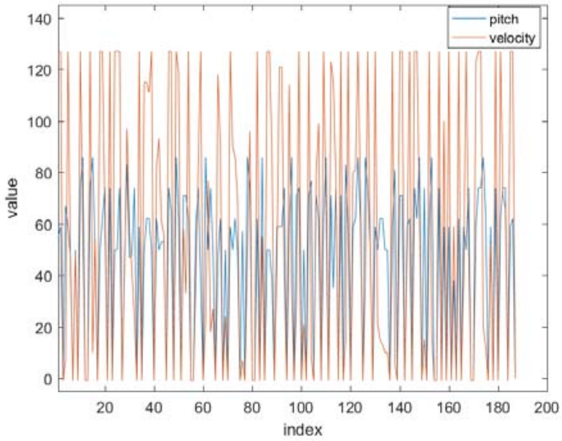


Figure 9. The pitch and velocity parameters using the method from [15]

It can be noticed that the average pitch value of our method is higher than that of [15]. Also, the average velocity value of the proposed method is smaller than that of [15]. The difference from the parameters shown in Figure 7 and Figure 9 can be explained by the fact that the optical flow computation method can get more movement information over the all the body, not only that of the joints tracked by our method. Our generated music sounds differently than that obtained by the method of [15]. However, different notes and piano music feeling can be obtained by varying the threshold, pitch range, register range and beat duration parameters of the proposed skeleton joints coordinates based method.

III. CONCLUSION AND FUTURE WORK

A piano music generating method from dance movement using a human pose detector, dynamic time warping and pitch pair mesh approaches is presented. The proposed technique is simple to implement, does not need special equipment and sensors and has the potential to generate aesthethically pleasing sounds. Future work will be focused on optimization of the parameters of the proposed method in order to open new perspectives of soundtrack generated from body movements.

ACKNOWLEDGMENT

The results presented in this work concerns the research carried out for the "SENTIR" research project, co-financed through the European Regional Development Fund, POC–A.1–A1.2.1–D–2015 grant, research, development and innovation supporting economic competitiveness and the development of businesses.

REFERENCES

[1] T. Hermann, A. Hunt, and J. G. Neuhoff, The Sonification Handbook, Logos Verlag, Berlin, 2011.

[2] M. M. Wanderley, B. W. Vines, N. Middleton, C. McKay, and W. Hatch, “The musical significance of clarinetists’ ancillary gestures: an exploration of the field,” Journal of New Music Research, vol. 34, no. 1, Feb. 2007, pp. 97–113, doi: 10.1080/09298210500124208.

[3] A. R. Jensenius, “Action-Sound: Developing methods and tools to study music-related body movement,” PhD dissertation, University of Oslo, 2007.

[4] R. M. Winters, A. Savard, V. Verfaille, and M. M. Wanderley, “A sonification tool for the analysis of large databases of expressive gesture,” The International Journal of Multimedia and its Applications, vol. 4, No. 6, Dec. 2012, pp. 13-26, doi: 10.5121/ijma.2012.4602.

[5] A. R. Jensenius and R. I. Godøy, “Sonifying the shape of human body motion using motiongrams,” Empirical Musicology Review, 7(3), Aug. 2013, pp. 73-83, doi:10.18061/emr.v8i2.

[6] G. Marino, M. H. Serra, and J. M. Raczinski, “The upic system: Origins and innovations,” Perspectives of New Music, Vol. 31, No. 1, Jan. 1993, pp. 258-269, doi: 10.2307/833053.

[7] AudioSculpt software [Online]. Available from http://anasynth.ircam.fr/home/english/software/audiosculpt 2017.11.06

[8] Spear software [Online]. Available from http://www.klingbeil.com/spear/ 2017.11.06

[9] A. Camurri et al., ”Eyesweb: Toward gesture and affect recognition in interactive dance and music systems,” Computer music Journal, vol. 24, No. 1, Mar. 2000, pp. 57-69, doi: 10.1162/014892600559182.

[10] M. Wright, R. Dudas, S. Khoury, R. Wang, and D. Zicarelli, "Supporting the Sound Description Interchange Format in the Max/MSP Environment", Proc. of the Int. Computer Music Conference (ICMC), Oct. 1999, pp. 1-4, doi: 10.1.1.30.6737.

[11] A. R. Jensenius, “Some video abstraction techniques for displaying body movement in analysis and performance,” Leonardo, Vol. 46, No. 1, Jan. 2013, pp. 53-60, , doi: 10.2307/23468117.

[12] Motioncomposer device [Online]. Available from http://motioncomposer.de/ 2017.11.06

[13] Pointmotioncontrol software [Online]. Available from http://www.pointmotioncontrol.com/ 2017.11.06

[14] T. Pohle and P. Knees, “Real-Time Synaesthetic Sonification of Traveling Landscapes” Proc. of 5th Int’l Mobile Music Workshop (MMW), May 2008, pp. 1-3, doi:10.1145/1459359.1459592.

[15] A. M. Taylor and J. Altosaar, "Sonification of Fish Movement Using Pitch Mesh Pairs" Proc. of the Int. Conf. on New Interfaces for Musical Expression (NIME), Jun. 2015, pp. 28-29.

[16] S. Wei, V. Ramakrishna, T. Kanade, and Y. Sheikh. “Convolutional Pose Machines” Proc. of IEEE Conf. on Computer Vision and Pattern Recognition (CVPR), IEEE Press, Jun. 2016, pp. 1-9, doi: 10.1109/CVPR.2016.511.

[17] C. S. Myers and L. R. Rabiner, "A Comparative Study of Several Dynamic Time-Warping Algorithms for Connected-Word Recognition," Bell System Technical Journal, vol. 60, No. 7, Sep. 1981, pp. 1389–1409, doi:10.1002/j.1538-7305.1981.tb00272.

[18] F. Albu, D. Hagiescu, M. A. Puica, and L. Vladutu, "Intelligent tutor for first grade children’s handwriting application", Proc. of 9th International Technology, Education and Development Conference (INTED), Mar. 2015, pp. 3708–3717, doi: 10.13140/RG.2.1.2591.7607.

[19] F. Albu, D. Hagiescu, and M. A. Puica, “Quality evaluation approaches of the first grade children’s handwriting”, Proc. of the 10th International Scientific Conference on eLearning and software for Education (ELSE), Apr. 2014, pp. 17-23, doi: 10.12753/2066-026X-17-055.

[20] Microsoft Windows MIDI Mapper Help [Online]. https://support.microsoft.com/en-us/help/84817/using-the-midi-mapper 2017.11.06.

ISBN: 978-1-61208-611-8