Matei Mancas Nicolas d'Alessandro
Xavier Siebert Bernard Gosselin
Carlos Valderrama Thierry Dutoit (Eds.)

LNICST 124

Intelligent Technologies for Interactive Entertainment

5th International ICST Conference, INTETAIN 2013
Mons, Belgium, July 2013
Revised Selected Papers

ICST
Springer

Lecture Notes of the Institute
for Computer Sciences, Social Informatics
and Telecommunications Engineering 124

Matei Mancas Nicolas d'Alessandro
Xavier Siebert Bernard Gosselin
Carlos Valderrama Thierry Dutoit (Eds.)

Intelligent Technologies for Interactive Entertainment

5th International ICST Conference, INTETAIN 2013
Mons, Belgium, July 3-5, 2013
Revised Selected Papers

Volume Editors

Matei Mancas
Nicolas d'Alessandro
Xavier Siebert
Bernard Gosselin
Carlos Valderrama
Thierry Dutoit

University of Mons (UMONS)
Numediart Institute
for Creative Technologies
Mons, Belgium
E-mail:
{matei.mancas, nicolas.dalessandro,
xavier.siebert, bernard.gosselin,
carlos.valderrama, thierry.dutoit}
@umons.ac.be

ISSN 1867-8211 e-ISSN 1867-822X
ISBN 978-3-319-03891-9 e-ISBN 978-3-319-03892-6
DOI 10.1007/978-3-319-03892-6
Springer Cham Heidelberg New York Dordrecht London

Library of Congress Control Number: 2013955939

CR Subject Classification (1998): K.8, I.2.10, H.5, I.5, I.4, J.5, J.7, H.1

Typesetting: Camera-ready by author, data conversion by Scientific Publishing Services, Chennai, India

Printed on acid-free paper

Springer is part of Springer Science+Business Media (www.springer.com)

Preface

After Genoa (2011), Amsterdam (2009), Cancun (2008), and Madonna di Campiglio (2005), the 5th edition of the International Conference on Intelligent Technologies for Interactive Entertainment (INTETAIN) was held in Mons, Belgium, and hosted by the NUMEDIART Institute of the University of Mons (UMONS).

Mons is a charming medieval city in the heart of Europe that was chosen to become the European Capital of Culture in 2015. Prepared as early as 2004, the "MONS2015 - *Where Technology Meets Culture*" event has boosted the involvement of the city in digital culture for the past decade, with the hosting of digital artists, the building of new rehearsal and performance theaters, and the development of the Mons Digital Innovation Valley, a scientific and industrial park that favors the outgrowth of creative SMEs. In this framework, it was natural for Mons to host INTETAIN 2013 (`http://www.intetain.org/2013/`), a conference that aims at fostering creativity and mixes science and technology with artistic creations and creative industries.

This year the conference focused on three main topics spread over the three days of the conference. The main topic of the first day was linked media, with applications to the future of television. It was concluded by a discussion/debate panel on the future of television and gathered both academic experts in television technologies and people from the broadcast industry. David Geerts, director of the user experience center of the KUL, and Lyndon Nixon, technical leader of the LinkedTV EU Project, discussed the issue with Fabrice Massin, director of new media at RTBF, and Thierry Piette, technical manager at RTL-TVI. The debate was moderated by Roger Roberts from TITAN-RTBF. The second day was more focused on gaming technologies, while the last day was dedicated to live entertainment.

Along with the paper sessions, two demonstration sessions were organized. One of them was held in conjunction with the LinkedTV EU Project and provided both technological and artistic demonstrations from several visual artists. The second session mainly showed technical demos of the main track papers.

We were very pleased to welcome three outstanding keynote speakers at INTETAIN 2013. David Geerts shared his enlightening views about the future of television, second screens and beyond, on the first day. He was followed by Gilles Pinault, co-founder of SoftKinetics S.A., on the second day, who talked about the natural and gestural interaction interfaces of the future. Our final speaker was Indy Saha, Director of Strategy of Google Creative Labs in London, who talked about new real-time interaction possibilities through the Web and especially through the Chrome navigator.

The conference coincided with CUTE 2013, a free one-day hands-on workshop with four tracks: Behavior tracking and advanced MOCAP, Smart rooms,

Performative speech synthesis in MAGE, and Sense & Perform. Several artistic installations were also shown in the framework of this event.

The organizers want to thank all the sponsors who made INTETAI 2013 become a reality: the European Alliance for Innovation (EAI), Create-Net research consortium, the University of Mons, the 175th anniversary program of its Faculty of Engineering, the FNRS, the EU-LinkedTV Project, the Twist Industry Cluster, Arts2 (the art school of Mons) and the NUMEDIART Institute. INTETAIN 2013 was a real melting pot between artists, scientists, and people from industry showing the potential of creativity in our society.

July 2013

Thierry Dutoit
Matei Mancas

Organization

The 5th International ICST Conference on Intelligent Technologies for Interactive Entertainment (INTETAIN 2013) was organized in Mons, Belgium, by the NUMEDIART Institute (`www.numediart.org`) of the University of Mons (UMONS).

Steering Committee

Imrich Chlamtac	CREATE-NET Research Consortium, Italy
Anton Nijholt	University of Twente, The Netherlands
Antonio Camurri	University of Genoa, Italy

Organizing Committee

Thierry Dutoit	University of Mons (UMONS)
Bernard Gosselin	University of Mons (UMONS)
Carlos Valderrama	University of Mons (UMONS)
Matei Mancas	University of Mons (UMONS)
Nicolas d'Alessandro	University of Mons (UMONS)
Xavier Siebert	University of Mons (UMONS)
Alexis Moinet	University of Mons (UMONS)
Stephane Dupont	University of Mons (UMONS)
Christian Frisson	University of Mons (UMONS)
Guy Vanden Bemden	University of Mons (UMONS)
François Zajega	University of Mons (UMONS)
Michel Cleempoel	ARTS2 School of Arts
Benoit Macq	Catholic University of Louvain (UCL)

Program Committee

Ben Falchuk	Ericsson
Steven Feiner	Columbia University, USA
Alois Ferscha	Johannes Kepler University, Austria
Matthew Flagg	Georgia Tech
Matei Mancas	University of Mons (UMONS), Belgium
Jaap van den Herik	University of Tilburg, The Netherlands
Dirk Heylen	University of Twente, The Netherlands
Hlavacs Helmut	University of Vienna, Austria
Herman van der Kooij	University of Twente, The Netherlands
Tsvi Kuflik	University of Haifa

Markus Loeckelt	DFKI GmbH, Germany
Isaac Rudomin	Tecnologico de Monterrey, Mexico
Evert van Loenen	Philips
Henry Lowood	Stanford University, USA
Maic Masuch	University of Duisberg-Essen, Germany
Oscar Mayora	CREATE-NET, Italy
John-Jules Meijer	University of Utrecht, The Netherlands
Imrich Chlamtac	CREATE-NET, Italy
Anton Nijholt	University of Twente, The Netherlands
Antonio Camurri	University of Genoa, Italy
Gualtiero Volpe	University of Genoa, Italy
Catherine Pelachaud	CNRS, Telecom ParisTech, France
Christophe De Vleeschouwer	Catholic University of Louvain (UCL), Belgium
Olivier Debeir	Free University of Brussels (ULB), Belgium
Albert Ali Salah	Bogazici University, Turkey
Radu-Daniel Vatavu	Stefan cel University Mare of Suceava, Romania
Florian Mueller	Stanford University, USA
Lyndon Nixon	STI GmbH
Paolo Petta	Medical Uniersity of Vienna, Austria
Fabio Pianesi	Fondazione Bruno Kessler (FBK), Italy
Louis-Philippe Morency	Institute for Creative Technologies, USC
Matthias Rauterberg	Eindhoven University of Technology, The Netherlands
Charles Rich	Worcester Polytechnic Institute, USA
Mark Riedl	Georgia Institute of Technology, USA
Stefan Agamanolis	Akron Children's Hospital, USA
Ulrike Spierling	FH/University of Applied Sciences
Pieter Spronck	Tilburg University, The Netherlands
Oliviero Stock	Fondazione Bruno Kessler (FBK), Italy
Bill Swartout	University of Southern California, USA
Mariet Theune	University of Twente, The Netherlands
Thanos Vasilakos	University of Western Macedonia, Greece
Woontack Woo	Gwangju Institute of Science and Technology (GIST), Korea
Wijnand IJsselstein	University of Eindhoven, The Netherlands
Massimo Zancanaro	Fondazione Bruno Kessler (FBK), Italy
Véronique Moeyart	University of Mons (UMONS), Belgium
Sébastien Bette	University of Mons (UMONS), Belgium
Bruno Quoitin	University of Mons (UMONS), Belgium
Elisabeth Andre	Augsburg University, Germany
Pierre Manneback	University of Mons (UMONS), Belgium
Tilde Bekker	University of Eindhoven, The Netherlands
Regina Bernhaupt	IRIT-CNRS, France
Kim Binsted	University of Hawaii, USA
Anthony Brooks	Aalborg University, Denmark

Yang Cai	Carnegie Mellon University, USA
Marc Cavazza	Teesside University, UK
Tat Jen Cham	Technological University
Kieth Cheverst	Lancaster University, UK
Arjan Egges	University of Utrecht, The Netherlands
Anton Eliëns	Vrije Universiteit Amsterdam, The Netherlands
Catholijn Jonker	Delft University of Technology, The Netherlands
Christian Jacquemin	LIMSI-CNRS, France
Sylvain Marchand	University of Brest, France
Jérôme Idier	Ecole Centrale Nantes, France
Richard Kronland-Martinet	LMA - CNRS, France
François Xavier Coudoux	University of Valenciennes, France
Sylvie Merviel	University of Valenciennes, France
Rudi Giot	ISIB Bruxelles, Belgium
Frank Kresin	Waag Society
Thierry Dutoit	University of Mons (UMONS), Belgium
Patrick Olivier	Newcastle University, UK
Berry Eggen	University of Eindhoven, The Netherlands
Andreas Butz	University of Munich, Germany
Christophe d'Alessandro	LIMSI-CNRS, France
Mark Maybury	MITRE
Donald Glowinski	University of Genoa, Italy

Sponsoring Institutions

European Alliance for Innovation and CREATE-NET Consortium

Numediart Institute of the University of Mons

University of Mons (UMONS) and the 175th Anniversary of the Engineering Faculty of Mons Program

Fonds National pour
la Recherche Scientifique

LinkedTV EU Project

High School for the Arts of Mons

TWIST Industry Cluster

Table of Contents

Linked Media

Gaming Technologies

Technologies for Live Entertainment

Personalized Summarization of Broadcasted Soccer Videos with Adaptive Fast-Forwarding

Fan Chen[1] and Christophe De Vleeschouwer[2]

[1] Japan Advanced Institute of Science and Technology, Nomi 923-1211, Japan
chen-fan@jaist.ac.jp
[2] Université catholique de Louvain, B-1348 Louvain-la-Neuve, Belgium
christophe.devleeschouwer@uclouvain.be

Abstract. We propose a hybrid personalized summarization framework that combines adaptive fast-forwarding and content truncation to generate comfortable and compact video summaries. We formulate video summarization as a discrete optimization problem, where the optimal summary is determined by adopting Lagrangian relaxation and convex-hull approximation to solve a resource allocation problem. Subjective experiments are performed to demonstrate the relevance and efficiency of the proposed method.

Keywords: Personalized Video Summarization, Adaptive Fast forwarding, Soccer Video Analysis.

1 Introduction

Video summarization techniques address different purposes, including fast browsing [6], retrieval [14], behaviour analysis [15], and entertainment. We intend to generate from the source video(s) a concise version with well organized story-telling, from which the audience can enjoy the contents that best satisfy their interest. Two kinds of information are essential for producing semantically relevant and enjoyable summaries: *Semantic information* of the scene directly evaluates the importance of frames for producing semantically relevant summaries; *Scene activity* is associated to the changes of the scene presented to the audience. Conventional content-truncation-based methods mainly maximize the semantic information associated to the content played during the constraint browsing period, e.g. using fast-browsing of highlighted moments [10]. However, semantic information extracted from individual images/segments fails to model a complicated story-telling with strong dependency in its contents. In contrast, conventional fast-forwarding-based methods mainly sample the video frames at a rate that increases with the measured scene activity, defined via optical flow [13] or the histogram of pixel differences [8]. By only evaluating changes in the scene, it is difficult to assure the semantic relevance of the summary. The application of pure fast-forwarding based methods is also constrained by the fact the highest tolerable playback speed is bounded due to the limitation of visual perception [9].

We thus propose an approach that truncates contents with intolerable playback speeds and saves time resources for better rendering the remaining contents.

M. Mancas et al. (Eds.): INTETAIN 2013, LNICST 124, pp. 1–11, 2013.

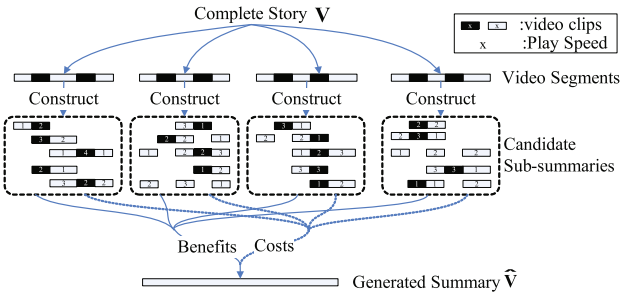


Fig. 1. Conceptual diagram of the overall proposed summarization process envisioned in a divide and conqure paradigm

We design a hybrid summarization method with both content truncation and adaptive fast-forwarding to provide continuous as well as semantically relevant summaries with improved visual comfort. We select playback speeds from a set of discrete options, and introduce a hierarchical summarization framework to find the optimal allocation of time resources into the summary, which enables various story-telling patterns for flexible personalized video summarization. Our resource allocation summarization in [2] only considers content truncation. In [1], we only considered the semantic information of pre-defined video segments in evaluating the benefit of the summary, and scene activity was only adopted for heuristically determining the maximum tolerable speed, in an independent process from the resource allocation optimization. In the present paper, we determine both the truncation actions and the playback speeds in a soft way through a unified resource allocation process that considers both semantic information and scene activity. Furthermore, subjective tests are now performed to validate the adaptive fast-forwarding principle.

The paper is organized as follows. In Section 2 we introduce the proposed summarization framework. In Section 3, we present experimental results. Finally, we conclude the paper in Section 4.

2 Resource Allocation Framework

Our resource-allocation-based framework interprets the summarization problem as finding the optimal allocation of duration resources u^{L} into video segments, according to various user preferences. We design the whole process using the divide and conquer paradigm (Fig.1(a)). The whole video is first cut into short clips by using a shot-boundary detector. These short clips are then organized into video segments. A sub-summary or local story defines one way to select clips within a segment. Several sub-summaries can be generated from a segment: not only the content, but also the narrative style of the summary can be adapted to user requirements. By tuning the benefit and the cost of sub-summaries, we balance -in a natural and personal way- the semantics (what is included in the summary) and the narrative (how it is presented to the user) of the summary.

The final summary is formed by collecting non-overlapping sub-summaries to maximize the overall benefit, under the user-preferences and duration constraint.

Let the video be cut into N^{C} clips, with the i^{th} clip $\mathcal{C}_i$ being $\mathcal{C}_i = \{t|t = t_i^S, \cdots, t_i^E\}$. t_i^S and t_i^E are the index of its starting and ending frames. These video clips are grouped into M segments. A set of candidate sub-summaries is considered for each segment, from which at most one sub-summary can be selected. We denote the k^{th} sub-summary of the m-th segment $\mathcal{S}_m$ as $\mathbf{a}_{mk}$, which is a set of playback speeds for all its clips, i.e., $\mathbf{a}_{mk} = \{v_{ki}|i \in \mathcal{S}_m\}$. v_{ki} is the playback speed assigned to the i^{th} clip if the k^{th} sub-summary $\mathbf{a}_{mk}$ is adopted.

Let $\mathbf{b}_m = \{b_i|i \in \mathcal{S}_m\}$ be the list of base benefits for all clips in $\mathcal{S}_m$. Our major task is to find the set of sub-summaries that maximizes the total payoff

$$\hat{\mathbf{V}}^* = \arg\max_{\hat{\mathbf{V}}} \mathcal{B}(\{\mathbf{a}_{mk}\}|\{\mathbf{b}_m\}), \tag{1}$$

subject to $\sum_{m=1}^{M} |\mathbf{a}_{mk}| \leq u^{\mathrm{L}}$. We define $|\mathbf{a}_{mk}|$ as the length of summary $\mathbf{a}_{mk}$,

$$|\mathbf{a}_{mk}| = \sum_{i \in \mathcal{S}_m} \frac{t_i^E - t_i^S}{v_{ki}}. \tag{2}$$

The overall benefit of the whole summary is defined as accumulated benefits of all selected sub-summaries:

$$\mathcal{B}(\{\mathbf{a}_{mk}\}|\{\mathbf{b}_m\}) = \sum_{m=1}^{M} \mathcal{B}_m(\mathbf{a}_{mk}) \tag{3}$$

with $\mathcal{B}_m(\mathbf{a}_{mk})$ being defined as a function of the user preferences, of the highlighted moments, and of the playback speeds as described in the following.

2.1 Video Segmentation

We divide the soccer video into clips, according to the detected production actions, such as position of replays, shot-boundaries and view types. We detect replays from producer-specific logos [12], extract shot-boundaries with a detector proposed in [7] to better deal with smooth transitions, and recognize the view-type by using the method in [4]. We segment the video based on the monitoring of production actions by analysing the view-structure [2] instead of using (complex) semantic scene analysis tools.

2.2 Local Story Organization

One major advantage of the resource allocation framework is that it allows highly personalized story organization, which is achieved via flexible definition of benefits. We define the benefit of a sub-summary as

$$\mathcal{B}_m(\mathbf{a}_{mk}) = \sum_{i \in \mathcal{S}_m} \mathcal{B}_i(v_{ki}) \mathcal{B}_{mi}^{\mathrm{P}}(\mathbf{a}_{mk}), \tag{4}$$

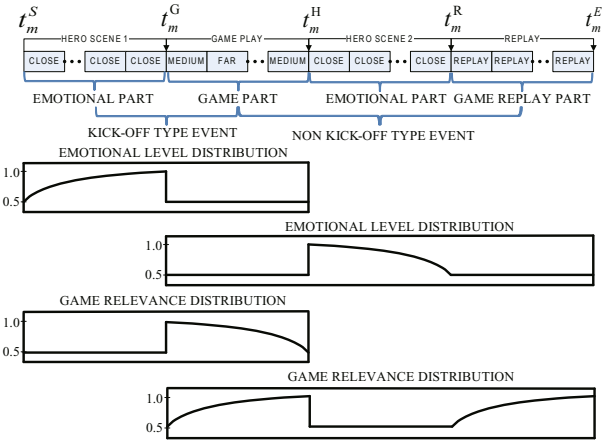


Fig. 2. The base benefit of a clip is evaluated from the game relevance and emotional level, defined as functions of clip view-types. The decaying process is modelled by hyperbolic tangent function. t_m^G, t_m^H, t_m^R are starting times of game play, hero scene, and replay in the m-th segment, respectively.

which includes accumulated benefits of selected clips. $\mathcal{B}_i(v_{ki})$ computes the base benefit of clip i at playback speed v_{ki},

$$\mathcal{B}_i(v_{ki}) = b_i(1/v_{ki})^{\beta}. \tag{5}$$

b_i is the clip benefit, defined as

$$b_i = |t_i^E - t_i^S| \left(\overline{f_t}\,\overline{a_t}\right)^{\alpha}, \tag{6}$$

which consider the average semantic information $\overline{f_t}$ and scene activity $\overline{a_t}$. As in [1], we automatically locate hot-spots by analyzing audio signals [3], whose (change of) intensity is correlated to the semantic importance of each video segment. The benefit of each frame t within each segment is further evaluated from its relevance to the game f_t^G and its level of emotional involvement f_t^E. The frame information f_t is computed as

$$f_t = 0.25 f_t^E + 0.75 f_t^G. \tag{7}$$

f_t^G mainly evaluates the semantic relevance of a clip in presenting the game progress, while f_t^E evaluates the importance of a clip in revoking the emotional involvement of the audience, e.g. via closeup view of a player. Hence, the above fixed weight favours game related contents in the summary. We define f_t^G and f_t^E by propagating the significance of the detected hot-spot event according to the view type structure of the segment, as depicted in Fig.2. The decaying process was modelled by using the hyperbolic tangent function, because it is bounded and is integrable thus simplifying the computation of $\overline{f_t}$.

Scene activity a_t is defined on the fluctuation of the camera view or the diversified movement of multiple players. Given a clip, the fluctuation of its

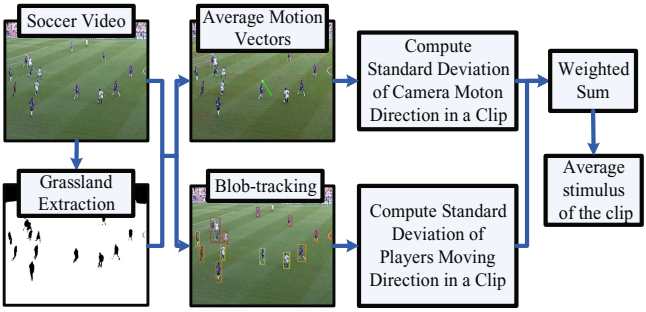


Fig. 3. We evaluate the average stimulus in a far-view clip by estimating information associated to scene activity from camera motion and player motion, which are computed on average motion vector in the grassland region and tracked player positions

camera view $\overline{\tau^{\mathrm{M}}}$ is evaluated by the average standard deviation of the motion vectors in the clip, while the complexity of diversified player movements $\overline{\tau^{\mathrm{P}}}$ is defined as the average standard deviation of players' moving speeds in the clip. As shown in Fig.3, the average information $\overline{a_t}$ is then defined as a weighted sum of the above two terms,

$$\overline{a_t} \propto \begin{cases} \overline{\tau^{\mathrm{M}}} + \overline{\tau^{\mathrm{P}}}, & \text{far view} \\ \overline{\tau^{\mathrm{M}}}, & \text{otherwise} \end{cases} \tag{8}$$

which is normalized to [0 1] for far-view and non-far-view clips independently. Using the standard deviation avoids the need of accurate compensation of player speed with respect to camera motions. $\mathcal{B}^{\mathrm{P}}_{mi}(\mathbf{a}_{mk})$ evaluates the extra benefits by satisfying specific preferences:

$$\mathcal{B}^{\mathrm{P}}_{mi}(\mathbf{a}_{mk}) = \mathcal{P}^{\mathrm{O}}(v_{ki}, u^{\mathrm{O}})\mathcal{P}^{C}_{mki}(u^{C})\mathcal{P}^{F}_{mk}. \tag{9}$$

$\mathcal{P}^{\mathrm{O}}(v_{ki}, u^{\mathrm{O}})$ is the extra gain obtained by including user's favorite object u^{O} specified through an interactive interface,

$$\mathcal{P}^{\mathrm{O}}(v_{ki}, u^{\mathrm{O}}) = \begin{cases} 1.5,\ v_{ki} < \infty, \exists t \in \mathcal{C}_i, u^{\mathrm{O}} \text{exists in} I_t, \\ 1.0, \text{otherwise.} \end{cases} \tag{10}$$

We favour a continuous story-telling by defining $\mathcal{P}^{C}_{mki}(u^{C})$

$$\mathcal{P}^{C}_{mki}(u^{C}) = 1 + u^{C}(2 - \delta_{\frac{1}{v_{ki}v_{k(i+1)}},0} - \delta_{\frac{1}{v_{ki}v_{k(i-1)}},0}), \tag{11}$$

where $\delta_{a,b}$ is the Kronecker delta function, and u^{C} is fixed to 0.1 in our experiments. Satisfaction of general production principles is also evaluated through $\mathcal{P}^{F}_{mk}$, which takes 1 for normal case and 0.001 for forbidden cases (or a value that is small enough to suppress this case from being selected), to avoid unpleasant visual/story-telling artifacts (e.g. too-short/incomplete local stories). We only allow normal speed for a replay clip in local story organization. If time resources to render a replay are available, we present the action in the clearest way.

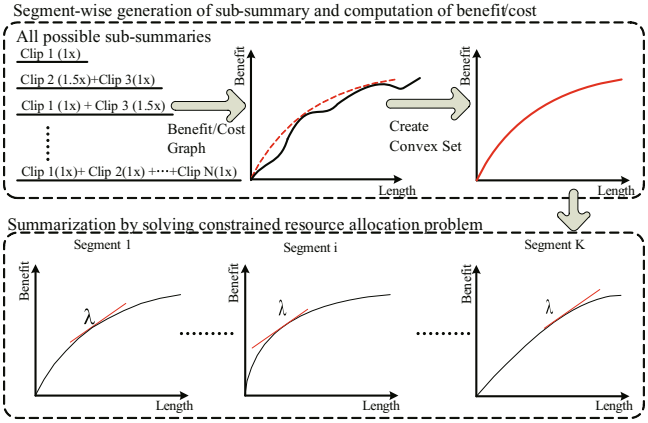


Fig. 4. Lagrangian relaxation and convex-hull approximation are adopted to solve the resource allocation problem, which restrict the eligible summarization options to the convex hulls of benefit-to-cost curves of the segments, where the collection of points from all convex-hulls with a same slope λ produces one optimal solution under the corresponding summary length

2.3 Global Story Organization

The global-duration resource is allocated among the available sub-summaries to maximize the aggregated benefit (Eq.1). When relaxation of constraints are allowed, Lagrangian optimization and convex-hull approximation can be considered to split the global optimization problem in a set of simple block-based decision problems [11]. The convex-hull approximation restricts the eligible summarization options for each sub-summary to the (benefit, cost) points sustaining the upper convex hull of the available (benefit, cost) pairs of the segment. Global optimization is obtained by allocating the available duration among the individual segment convex-hulls [5], which results in a computationally efficient solution. Fig.4 summarizes the summarization process.

We solve this resource allocation problem by using the Lagrangian relaxation [5]: if λ is a non-negative Lagrangian multiplier and $\{k^*\}$ is the optimal set that maximizes

$$\mathcal{L}(\{k\}) = \sum_{m=1}^{M} \mathcal{B}_m(\mathbf{a}_{mk}) - \lambda \sum_{m=1}^{M} |\mathbf{a}_{mk}| \tag{12}$$

over all possible $\{k\}$, then $\{\mathbf{a}_{mk^*}\}$ maximizes $\sum_{m=1}^{M} \mathcal{B}_m(\mathbf{a}_{mk})$ over all $\{\mathbf{a}_{mk}\}$ such that $\sum_{m=1}^{M} |\mathbf{a}_{mk}| \leqslant \sum_{m=1}^{M} |\mathbf{a}_{mk^*}|$. Hence, if $\{k^*\}$ solves the unconstrained problem in Eq.12, then it also provides the optimal solution to the constrained problem in Eq.1, with $u^{\mathrm{L}} = \sum_{m=1}^{M} |\mathbf{a}^*_{mk}|$. Since the contributions to the benefit and cost of all segments are independent and additive, we can write

$$\sum_{m=1}^{M} \mathcal{B}_m(\mathbf{a}_{mk}) - \lambda \sum_{m=1}^{M} |\mathbf{a}_{mk}| = \sum_{m=1}^{M} (\mathcal{B}_m(\mathbf{a}_{mk}) - \lambda |\mathbf{a}_{mk}|). \tag{13}$$

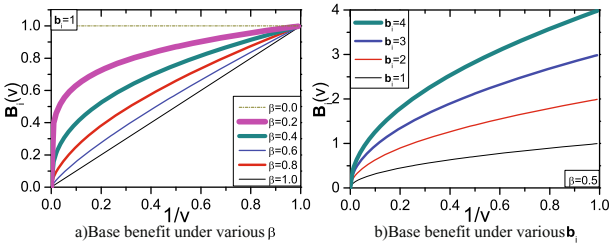


Fig. 5. Clip benefit complies with convex-hull approximation and the greedy algorithm adopted for solving the resource allocation problem

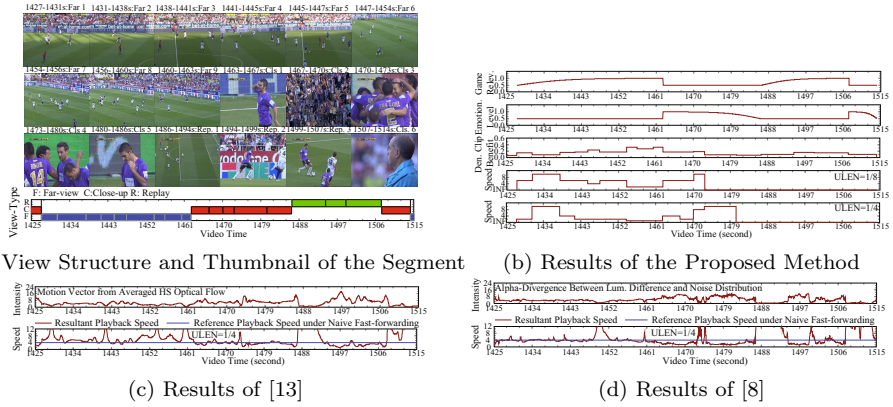
From the curves of $\mathcal{B}_m(\mathbf{a}_{mk})$ with respect to their corresponding summary length $|\mathbf{a}_{mk}|$, the collection of points maximizing $\mathcal{B}_m(\mathbf{a}_{mk}) - \lambda|\mathbf{a}_{mk}|$ with a same slope λ produces one unconstrained optimum. Different choices of λ lead to different summary lengths. If we construct a set of convex hulls from the curves of $\mathcal{B}_m(\mathbf{a}_{mk})$ with respect to $|\mathbf{a}_{mk}|$, we can use a greedy algorithm to search for the optimum under a given constraint u^{L}. The approach is depicted in Fig.4 and explained in details in [11]. In short, for each point in each convex hull, we first compute the forward (incremental) differences in both benefits and summary-lengths. We then sort the points of all convex-hulls in decreasing order of λ, i.e., of the increment of benefit per unit of length. Ordered points are accumulated until the summary length gets larger or equal to u^{L}.

Fig.5 shows the clip benefit $\mathcal{B}_i(v)$ w.r.t. $1/v$ under various β and b_i values, so as to analyse the behaviour the clip interest defined in Eq.5 in the above optimization process. Fig.5(a) reveals that the whole curve is convex when $0 < \beta < 1$, which thus enables various options of playback speeds to appear in the benefit/cost convex hulls. In Fig.5(b), we found that the clip with a higher base interest b_i has the same slope value at a slower playback speed. Accordingly, in the above greedy algorithm, slower playback speed will be first assigned to semantically more important clips in the sense of high information.

3 Experimental Results

The proposed framework aims at focusing on summarization with adaptive fast-forwarding and semantically relevant and personalized story telling. Those properties are explored through a comparative analysis with state of the art methods. The soccer video used for performance evaluation is 3 hours long with a list of 50 automatically extracted audio hot-spots. Seven different speed options, i.e., 1x, 2x, 4x, 6x, 8x, 10x, and $+\infty$(for content truncation), are enabled in the current implementation, so as to provide comparative flexibility in fast-forwarding control to those methods with continuous playback speeds. Here, ax stands for the a times of the normal playback speed. We compared the behavior of our proposed method to the following two methods:

– Peker et al. [13] achieve the adaptive fast-forwarding via constant activity sub-sampling.



(a) View Structure and Thumbnail of the Segment (b) Results of the Proposed Method

(c) Results of [13] (d) Results of [8]

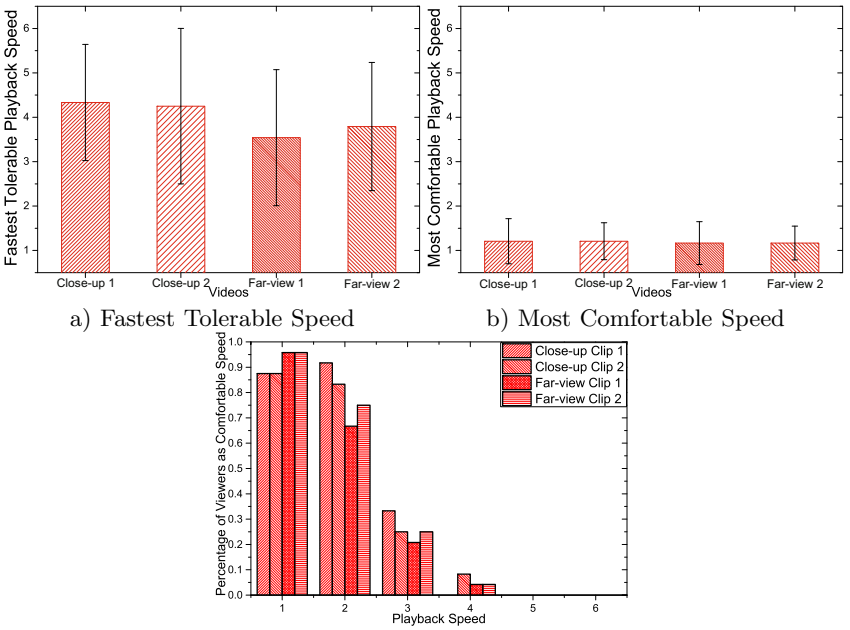
Fig. 6. Summaries produced for the broadcasted soccer video. The first subgraph presents the view-structure of segments and clip thumbnails. Resultant playback speeds from three methods are plotted with the corresponding clip benefit (ULEN for u^{L}).

– Höferlin et al. [8] determine the activity level by computing the alpha-divergence between the luminance difference of two consecutive frames and the estimated noise model. The adjusted sampling interval is then set to be linearly proportional to the activity level.

The results of the proposed and comparison methods are shown in Figs.6. We use $\alpha = \beta = 0.5$ and plot the results from different methods. We made the following major observations: both the optical flow and the alpha divergence failed to correctly measure the intensity of scene activities or the importance of the events; compared to the linear playback speed control in [13] and [8], our framework allows flexible personalization of story organization. We can suppress redundant contents in the replays for higher compaction, consider story continuity, and remove very short clips to avoid flickering; playback speeds of different clips in [13] and [8] maintain the same ratio, when the length of target summary changes, while our method performs non-linear time allocation under different target summary lengths, owing to the flexible definition of clip benefit.

We first subjectively evaluate the suitable playback speeds (Fig.7). 25 participants (including 11 females and 14 male, age from 20-40) were asked to specify their highest tolerable playback speed, comfortable playback speeds and the most comfortable playback speed when presented four groups of video samples with various playback speeds. The highest tolerable speed for far views is lower than that of the close-up views. We consider this as a result that understanding far-view need attentional perception to follow the players. Audiences still feel comfortable in faster playback speeds, which is the base of adaptive fast-forwarding. The most comfortable speed is selected to be the original speed that was produced by experts.

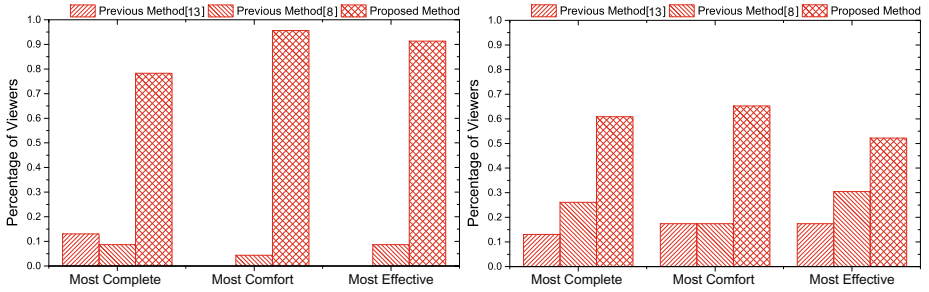
We then collect the global impression of the audiences in comparatively evaluating the generated summaries. We asked 23 participants (including 10 females and 13 males, age from 20-40) to give their opinions on the most preferred result



a) Fastest Tolerable Speed b) Most Comfortable Speed

c) Comfortable Speeds

Fig. 7. Results of the first subjective evaluation from 25 participants on their feedback under various fast-forwarding speeds when browsing the broadcasted soccer videos



a) High Compaction Ratio (1/8) b) Low Compaction Ratio (1/4)

Fig. 8. Results of the second subjective evaluation test from 23 viewers, by collecting their global impression on the summaries, in the sense of completeness, comfort and the efficiency of time allocation

when presented a group of three summaries generated by the above different methods (in the random order), from their *completeness*, *comfort*, and *effectiveness* of time allocation. We plot the results of evaluating summaries under two different compaction ratios (i.e. 1/8 and 1/4) in Fig.8. We make the following observations: our method outperforms the other two methods in generating more complete summaries for highly compact summarization (1/8), which supports our idea of introducing content truncation to save time resources for presenting

key events in a clearer way; our method produces more comfortable summaries from the broadcasted soccer video, where both 1/8 and 1/4 are too high for an adaptive fast-forwarding method to produce a comfortable video without truncating some contents. In order to slow down a key event, we have to raise the playback speed of other contents to a much higher level in exchange for the equivalent time resource, which results in flickering and lowers the visual comfort of the summary; our method is evaluated to be the most effective in allocating playback speeds for presenting the actions of interest, especially under a high compaction ratio.

4 Conclusions

We proposed a framework for producing personalized summaries that enables both content truncation and adaptive fast-forwarding. Instead of a rigid determination of the fast-forwarding speed, we efficiently select the optimal combination from candidate summaries, which is solved efficiently as a resource-allocation problem. Subjective experiments demonstrate the proposed system by evaluating summaries from broadcasted soccer videos. We will further extend our hybrid method of content truncation and adaptive fast-forwarding. Both semantic information and scene activity are important in producing a semantically relevant and visually comfort summary. We will thus consider both types of information in our future work.

References

1. Chen, F., De Vleeschouwer, C.: Automatic summarization of broadcasted soccer videos with adaptive fast-forwarding. In: ICME 2011, pp. 1–6 (July 2011)
2. Chen, F., De Vleeschouwer, C.: Formulating team-sport video summarization as a resource allocation problem. TCSVT 21(2), 193–205 (2011)
3. Duxans, H., Anguera, X., Conejero, D.: Audio based soccer game summarization. In: BMSB 2009, pp. 1–6 (2009)
4. Ekin, A., Tekalp, A., Mehrotra, R.: Automatic soccer video analysis and summarization. TIP 12(7), 796–807 (2003)
5. Everett, H.: Generalized lagrange multiplier method for solving problems of optimum allocation of resources. Operations Research 11(3), 399–417 (1963)
6. Ferman, A., Tekalp, A.: Two-stage hierarchical video summary extraction to match low-level user browsing preferences. TMM 5(2), 244–256 (2003)
7. Fernandez, I., Chen, F., Lavigne, F., Desurmont, X., De Vleeschouwer, C.: Browsing sport content through an interactive h.264 streaming session. In: MMEDIA 2010, vol. 1, pp. 155–161 (June 2010)
8. Höferlin, B., Höferlin, M., Weiskopf, D., Heidemann, G.: Information-based adaptive fast-forward for visual surveillance. Multimedia Tools Appl. 55, 127–150 (2011)
9. Holcombe, A.O.: Seeing slow and seeing fast: Two limits on perception. Trends in Cognitive Sciences 13(5), 216–221 (2009)
10. Li, Z., Schuster, G.M., Katsaggelos, A.K.: Minmax optimal video summarization. TCSVT 15, 1245–1256 (2005)
11. Ortega, A.: Optimal bit allocation under multiple rate constraints. In: DCC 1996, pp. 349–358 (1996)

12. Pan, H., van Beek, P., Sezan, M.I.: Detection of slow-motion replay segments in sports video for highlights generation. In: ICASSP 2001, vol. 3, pp. 1649–1652 (2001)
13. Peker, K.A., Divakaran, A., Sun, H.: Constant pace skimming and temporal subsampling of video using motion activity. In: ICIP 2001, vol. 3, pp. 414–417 (2001)
14. de Silva, G.C., Yamasaki, T., Aizawa, K.: Evaluation of video summarization for a large number of cameras in ubiquitous home. In: ACM MM 2005, pp. 820–828 (2005)
15. Zhu, G., Huang, Q., Xu, C., Rui, Y., Jiang, S., Gao, W., Yao, H.: Trajectory based event tactics analysis in broadcast sports video. In: ACM MM 2007, pp. 58–67 (2007)

Real-Time GPU-Based Motion Detection and Tracking Using Full HD Videos

Sidi Ahmed Mahmoudi[1], Michal Kierzynka[2], and Pierre Manneback[1]

[1] University of Mons, Faculty of Engineering, Mons, Belgium
[2] Poznań Supercomputing and Networking Center, Poznań, Poland

Abstract. Video processing algorithms present a necessary tool for various domains related to computer vision such as motion tracking, videos indexation and event detection. However, the new video standards, especially those in high definitions, cause that current implementations, even running on modern hardware, no longer respect the needs of real-time processing. Several solutions have been proposed to overcome this constraint, by exploiting graphic processing units (GPUs). Although, they present a high potential of GPU, any is able to treat high definition videos efficiently. In this work, we propose a development scheme enabling an efficient exploitation of GPUs, in order to achieve real-time processing of Full HD videos. Based on this scheme, we developed GPU implementations of several methods related to motion tracking such as silhouette extraction, corners detection and tracking using optical flow estimation. These implementations are exploited for improving performances of an application of real-time motion detection using mobile camera.

Keywords: GPU, CUDA, video procesing, motion tracking, real-time.

1 Introduction

In recent years, the CPU power has been capped, essentially for thermal reasons, to less than 4 GHz. A limitation that has been circumvented by the change of internal architecture, with multiplying the number of integrated computing units. This evolution is reflected in both general (CPU) and graphic (GPU) processors, as well as in recent accelerated processors (APU) which combine CPU and GPU on the same chip [1]. Moreover, GPUs have larger number of computing units, and their power has far exceeded the CPUs ones. Indeed, the advent of GPU programming interfaces (API) has encouraged many researchers to exploit them for accelerating algorithms initially designed for CPUs.

Video processing and more particularly motion estimation algorithms present the core of various methods used in computer vision. They have been used, for example, in surveillance systems tracking humans in public places, such as metro or airports, to identify possible abnormal behaviors and threats [2,3]. Motion estimation algorithms serve therefore as a common building block of some more complex routines and systems. However, these algorithms are hampered by their high consumption of both computing power and memory. The exploitation

M. Mancas et al. (Eds.): INTETAIN 2013, LNICST 124, pp. 12–21, 2013.

of graphic processors can present an efficient solution for their acceleration. Indeed, they can present prime candidates for acceleration on GPU by exploiting its processing units in parallel, since they consist mainly of a common computation over many pixels. Nevertheless, the new standards, especially those in high resolutions cause that current implementations even running on modern hardware, no longer meet the needs of real-time processing. Moreover, modern surveillance systems are nowadays more commonly equipped with high definition cameras that expect to be treated in real-time. Furthermore, the treatment of TV broadcast images, which cannot be down sampled, require an accelerated object detection and recognition. Therfore, a fast processing of videos is needed to ensure the treatment of 25 high definition frames per second (25 fps). To overcome these constraints, several GPU computing approaches have recently been proposed. Although they present a great potential of a GPU platform, any one is able to process high definition video sequences efficiently. Thus, a need arose to develop a tool being able to address the outlined problem.

In this paper, we propose a development scheme enabling an effective exploitation of GPUs for accelerating video processing algorithms, and hence achieving real-time treatment of high definition videos. This scheme allows an efficient management of GPU memories and a fast visualization of results. Based on this scheme, we developed CUDA [4] implementations of methods related to motion tracking domain such as silhouette extraction, corners detection and tracking using optical flow estimation. These implementations are exploited for accelerating a method of real-time motion detection using mobile camera.

The remainder of the paper is organized as follows: related works are described in section 2. Section 3 presents our development scheme for video processing on GPU. Section 4 describes our GPU implementations of silhouette extraction, features detection and tracking methods. Section 5 presents the use of these implementations for impoving performance of motion detection using mobile camera. Finally, section 6 concludes and proposes further work.

2 Related Works

Unlike algorithms requiring a high dependency of computation between the input data and hence a complicated parallelization, most of image and video processing algorithms consist of similar computations over many pixels. This fact makes them well adapted for acceleration on GPU by exploiting its processing units in parallel. Otherwise, these algorithms require generally a real-time treatment of video frames. We may find several methods in this category such as human behavior understanding, event detection, camera motion estimation. These methods are generally based on motion tracking algorithms that can exploit several techniques such as optical flow estimation [6], block matching technique [7] and SIFT [8] descriptors.

Motion tracking methods consist on estimating the displacement and velocity of features in a given video frame with respect to the previous one. In this work, we are more focused on optical flow methods since they present a promising

solution for tracking even in noisy and crowded scenes or in case of small motions. In case of GPU-based optical flow motion tracking algorithms, one can find two kinds of related works. The first presents so called dense optical flow which tracks all pixels without selecting features. In this context, [9] presented a GPU implementation, using the API CUDA [4], of the Lucas-Kanade method used for optical flow estimation. The method computes dense and accurate velocity field at 15 fps with 640×480 video resolution. Authors in [11] proposed the CUDA implementation of the Horn-Schunck optical flow algorithm with a real-time processing of low resolution videos (316×252). The second category consists of methods that enable to track selected image features only. Sinha *et al.* [12] developed a GPU implementation of the KLT feature tracker [13] and the SIFT feature extraction algorithm [8]. This allowed to detect 800 features from 640×480 video at 10 fps which is around 10 times faster than the CPU implementation. However, despite their high speedups, none of the abovementioned GPU-based implementations can provide real-time processing of high definition videos. Otherwise, OpenCL [5] proposed a framework for writing programs which execute across hybrid platforms consisting of both CPUs and GPUs. There are also some GPU works dedicated to medical imaging for parallel [22] and heterogeneous [15,25] computation for vertebra detection and segmentation in X-ray images.

Our contribution focuses on the conception of a scheme development that enables an efficient exploitation of GPUs for high definition video processing in real-time. This scheme is based upon CUDA for parallel constructs and OpenGL [14] for visualization. It enables also an effective management of GPU memories that allows a fast access to pixels within video frames. Based on this scheme, we developed GPU implementations of three methods : silhouette extraction, features detection and tracking using optical flow estimation. These implementations enabled a real-time processing of Full HD videos, they were exploited for improving performance of real-time motion detection using camera in move.

3 Video Processing on GPU

As pointed out in previous sections, a GPU presents an effective tool for accelerating video processing algorithms. This section is presented in two parts: the first one describes our development scheme for video processing on GPU, showing also the employed GPU optimization techniques. The second part is devoted to describe our GPU implementations of silhouette extraction, features detection and tracking algrithms that exploit optical flow measures.

3.1 Development Scheme for Video Processing on GPU

The proposed scheme is based upon CUDA for parallel computing and OpenGL for visualization. This scheme is based on the three following steps :

1. **Loading of video frames on GPU:** we start with reading and decoding the video frames using the OpenCV library [16]. We copy the current frame on a device (GPU) that processes it in the next step.

2. **CUDA parallel processing:** before launching the parallel processing of the current frame, the number of GPU threads has to be defined, so that each thread can perform its processing on one or a group of pixels. This enables the program to treat the image pixels in parallel. Note that the number of threads depends on the number of pixels.
3. **OpenGL visualization:** the current image can be directly visualized on the screen through the video output of GPU. Therefore, we use the OpenGL library that allows for fast visualization, as it can operate buffers already existing on GPU, and thus requires less data transfer between host and device memories. Once the visualization of the current image is completed, the program goes back to the first step to load and process next frames. Otherwise and in case of multiple videos processing, the OpenGL visualization will be impossible using one video output only. So, a transfer of the processed video frames from GPU to CPU memory is required, which represents an additional cost for the application.

For a best exploitation of GPUs, we employed two optimization techniques. The first one consists on exploiting texture and shared memories. Indeed, video frames are loaded on texture memory in order to have a fast access to pixels values. The pixel neighbors are loaded on shared memory for a fast processing of pixels using their neighbors' values. The second optimization that we propose is the exploitation of four CUDA streams in order to overlap kernels executions by images transfers. Each stream consists of three instructions :

1. Copy of the current frame from host to GPU memory
2. Computations performed by CUDA kernels
3. Copy of the current frame (already processed) from GPU to host memory

3.2 GPU Implementations

Based on the scheme described in section 3.1, we propose the GPU implementation of silhouette extraction, features detection and tracking methods, which enabled to obtain both efficient results in terms of the quality of detected and tracked motions, and improved performance thanks to the exploitation of GPU.

3.2.1 GPU-Based Silhouette Extraction

The computation of difference between frames presents a simple and efficient method for detecting the silhouettes of moving objects. Based on the scheme presented in section 3.1, we propose the GPU implementation of this method using three steps. We start by loading the two first frames on GPU in order to compute the difference between them during the CUDA parallel processing step. Once the first image displayed, we replace it by the next video frame in order to apply the same treatment. Fig. 1(a) presents the obtained result of silhouette extraction. This figure shows two silhouettes extracted, that present two moving persons. In order to improve the quality of results, a threshold of 200 was used for noise elimination.

3.2.2 GPU-Based Features Detection and Tracking

In this section, we propose the GPU implementation of both features detection and tracking methods. The first one enables to detect features that are good to track, i.e. corners. To achieve this, we have exploited the Bouguet's corners extraction technique [17], which is based on the principle of Harris detector [24]. Our GPU implementation of this method is detailled in [18,19,20].

The second step enables to track the features previously detected using the optical flow method, which presents a distribution of apparent velocities of movement of brightness pattern in an image. It enables to compute the spatial displacements of images pixels based on the assumption of constant light hypothesis which supposes that the properties of consecutive images are similar in a small region. For more detail about optical flow computation, we refer readers to [6]. In literature, several optical flow methods exist such as Horn-Shunck [21], Lucas-Kanade [23] and block matching [7]. In this work, we propose the GPU implementation of the Lucas-Kanade algorithm, which is well known for its high efficiency, accuracy and robustness. This algorithm disposes of six steps:

1. **Step 1: Pyramid construction :** In the first step, the algorithm computes a pyramid representation of images I and J which represent two consecutive images from the video. The other pyramid levels are built in a recursive fashion by applying a Gaussian filter. Once the pyramid is constructed, a loop is launched that starts from the smallest image (the highest pyramid level) and ends with the original image (level 0). Its goal is to propagate the displacement vector between the pyramid levels.
2. **Step 2: Pixels matching over levels :** For each pyramid level (described in the previous step), the new coordinates of pixels (or corners) are calculated.
3. **Step 3: Local gradient computation :** In this step, the matrix of spatial gradient G is computed for each pixel (or corner) of the image I. This matrix of four elements (2×2) is calculated based on the horizontal and vertical spatial derivatives. The computation of the gradient matrix takes into account the area (window) of pixels which are centered on the point to track.
4. **Step 4: Iterative loop launch and temporal derivative computation:** A loop is launched and iterated until the difference between the two successive optical flow measures (calculated in the next step), or iterations, is higher than a defined threshold. Once the loop is launched, the computation of the temporal derivatives is performed using the image J (second image). This derivative is obtained by the subtraction of each pixel (or corner) of the image I (first image) and its corresponding corner in the image J (second image). This enables to estimate the displacement estimations which is then propagated between successive pyramid levels.
5. **Step 5: Optical flow computation:** The optical flow measure $\bar{g}$ is calculated using the gradient matrix G and the sum of temporal derivatives presented by shift vector $\bar{b}$. The measure of optical flow is calculated by multiplying the inverse of the gradient matrix G by the shift vector $\bar{b}$.
6. **Step 6: Result propagation and end of the pyramid loop:** The current results are propagated to the lower level. Once the algorithm reaches the

lowest pyramid level (the original image), the pyramid loop (launched in the first step) is stopped. The vector $\bar{g}$ presents the final optical flow value of the analyzed corner. For more detail, we refer readers to [17].

Upon matching and tracking pixels (corners) between frames, the result is a set of vectors as shown in Equation (1):

$$\Omega = \{\omega_1 \ ... \ \omega_n \mid \omega_i = (x_i, y_i, v_i, \alpha_i)\} \tag{1}$$

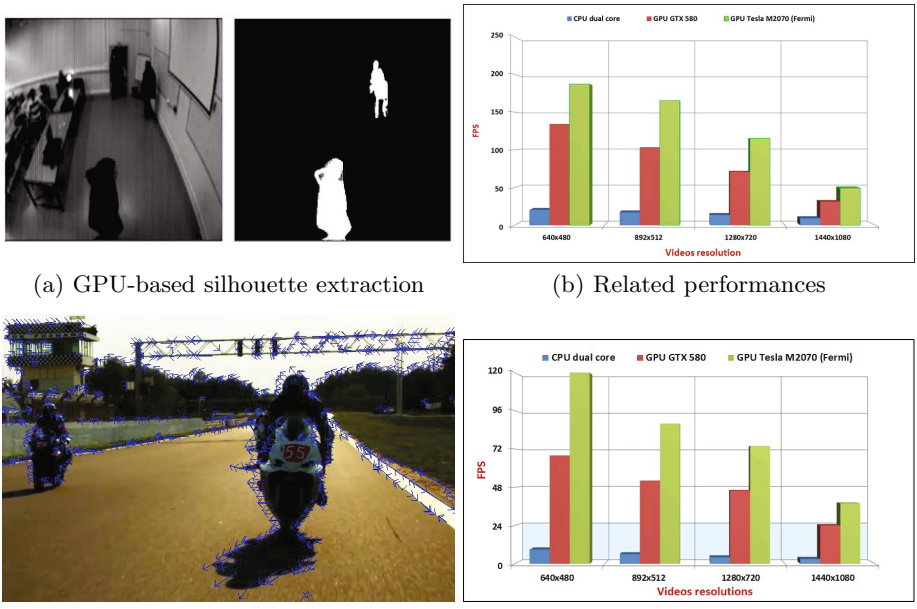
where:

- x_i, y_i are the x a y coordinates of the feature i;
- v_i represents the velocity of the feature i;
- α_i denotes motion direction of the feature i.

Based on the scheme presented in section 3.1, we propose the GPU implementation of the Lucas-Kanade otpical flow method by parallelizing its steps on GPU. These steps are executed in parallel using CUDA such that each GPU thread applies its instructions (among the six steps) on one pixel or corner. Therefore, the number of GPU threads is equal to the number of pixels or corners. Since the algorithm looks at the neighboring pixels, for a given pixel, the images, or pyramid levels are kept in the texture memory. This allows a faster access within the 2-dimensional spatial data. Other data, e.g. the arrays with computed displacements, are kept in the global memory, and are cached in the shared memory if needed. Notice that the quality of results remains identical since the process has not changed. Fig. 1(b) presents the comparison between CPU and GPU implementations of silhouette extraction method, while figures 1(c) and 1(d) present, respectively, the quality and performance of our GPU implementation of features detection and tracking method using optical flow estimation. These performances are compared with a CPU solution developed with OpenCV [16]. Notice that the constraint of real-time processing can be achieved with high definition videos thanks to the efficient exploitation of high computing power of GPUs. Notice also that the transfer time of video frames between CPU and GPU memories is included. This transfer time presents about 15 % from the total time of the application.

4 GPU for Real-Time Motion Detection Using Mobile Camera

The abovementioned GPU implementations are exploited in an application that consists of real-time motion detection within moving camera. In this category, motion detection algorithms are generally based on background subtraction which presents a widely used technique in computer vision domain. Typically, a fixed background is given to the application and new frames are subtracted from this background to detect the motion. The difference will give the objects or motion when the frame is subtracted from the fixed background. This difference in resulting binary image is called foreground objects. However, some



(a) GPU-based silhouette extraction

(b) Related performances

(c) GPU-based motion tracking

(d) Related performances

Fig. 1. Real-time treatment of Full HD videos on GPU

scenarios present a dynamic background which can changes due the movement of cameras. In this context, we propose an application for real-time background subtraction, which enables to detect automatically background and foreground using a moving camera. This application can be summarized in four steps :

1. **Corners detection:** The Harris corner detector [24] is applied to extract good features to track and examine for camera motion.
2. **Optical flow computation:** The Lukas-Kanade optical flow method [17] is applied to track the corners, detected previously.
3. **Camera motion Inhibition:** The camera motion is estimated by computing the dominant values of optical flow vectors. This enables to extract the common area beween each two consecutive images and focus only on motions related to objects in the scene.
4. **Motion detection:** This step consists of detecting movements based on computing the difference between each two consecutive frames.

In order to achieve a real-time treatment of high definition videos, the most intensive steps of this method are ported on GPU: corners detection, optical flow computation, motion detection. The GPU implementation of these steps is described in section 3.2, following the steps of loading of video frames on GPU, CUDA parallel processing and OpenGL visualization. Fig. 2.(a) shows a scene of camera motion. Dotted and dashed line presents the first image, dotted line presents the second frame and solid line shows the joint area of

two frames. Once, the camera motion is estimated. The joint area between 2 consecutive frames is determined by cropping the incoming and outgoing areas as seen in the white area of Fig. 2.(a). Fig. 2.(b) shows the resulting image of background subtraction. White areas represent the difference around moving objects. Table 1 presents a comparison between CPU and GPU performances of the abovementioned method. Notice that the use of GPU enabled a real-time processing for Full HD videos (1920×1080), which is 20 times faster than the corresponding CPU version.

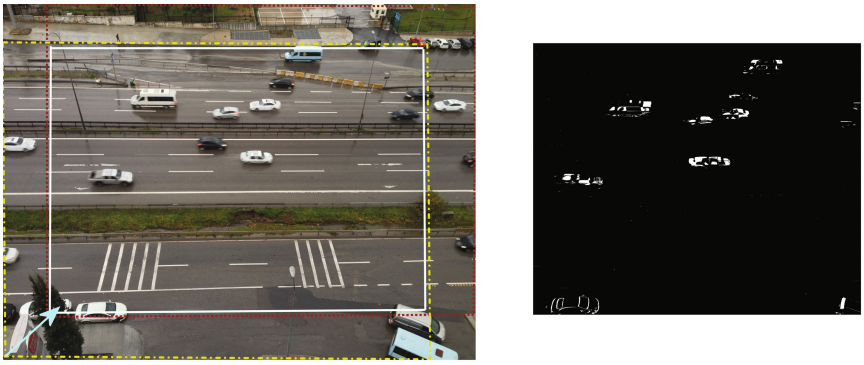


Fig. 2. (a). Camera motion estimation (b). Motion detection

Table 1. GPU performances of motion detection using mobile camera

Resolution	**CPU dual-core**	**GPU**	**Acceleration**
512×512	5 fps	79 fps	15,80 ×
1280×720	2,9 fps	51 fps	17,59 ×
1920×1080	1,7 fps	35 fps	20,59 ×

5 Conclusion

We proposed in this paper a development scheme for video processing on GPUs. Based on this scheme, we proposed an efficient implementation of the optical flow algorithm for the sparse motion tracking. More precisely, we developed a GPU based software that applies Lucas-Kanade tracking method to the previously detected corners. A GPU implementation of the silhouette extraction, based on frames difference, was also developed. These implementations were exploited for improving performance of an application that requires a real-time processing of high definition videos. This application consists of motion detection using a camera in move. As future work, we plan to develop a smart system for real-time processing of high definition videos in multi-user scenarios. This system

could exploit nvidia and ATI graphic cards thanks to the exploitation of CUDA and OpenCL APIs, respectively. The idea is to provide a dynamic platform enabling to facilitate the implementation of new advanced monitoring and control systems, effectively, that exploit parallel and heterogeneous architectures, with minimum energy consumption.

References

1. AMD Fusion, Family of APUs. The Future brought to you by AMD introducing the AMD APU Family, http://sites.amd.com/us/fusion/au/Pages/fusion.aspx
2. Fonseca, A., Mayron, L., Socek, D., Marques, O.: Design and im-plementation of an optical flow-based autonomous video surveillance system. In: Proceedings of the IASTED, p. 209 (2008)
3. Mahmoudi, S.A., Sharif, H., Ihaddadene, N., Djerabe, C.: Abnormal event detection in real time video. In: 1st International Workshop on Multimodal Interactions Analysis of Users in a Controlled Environment, ICMI (2008)
4. NVIDIA, NVIDIA CUDA: Compute Unified Device Architecture (2007), http://www.nvidia.com/cuda
5. Khronos-Group, The Open Standard for Parallel Programming of Heterogeneous Systems (2009), http://www.khronos.org/opencl
6. Bimbo, A.D., Nezi, P., Sanz, J.L.C.: Optical flow computation using extended constraints. IEEE Transaction on Image Processing, 720 (1996)
7. Kitt, B., Ranft, B., Lategahn, H.: Block-matching based optical flow estimation with reduced search space based on geometric constraints. In: 13th International Conference on Intelligent Transportation Systems, p. 1140 (2010)
8. Lowe, D.G.: Distinctive image features from scale-invariant keypoints. International Journal of Computer Vision (IJCV) 60(2), 91 (2004)
9. Marzat, J., Dumortier, Y., Ducrot, A.: Real-time dense and accurate parallel optical flow using CUDA. In: In Proceedings of WSCG, p. 105 (2009)
10. Mizukami, Y., Tadamura, K.: Optical Flow Computation on Compute Unified Device Architecture. In: ICIAP'14, p. 179 (2007)
11. Mizukami, Y., Tadamura, K.: Optical Flow Computation on Compute Unified Device Architecture. In: ICIAP'14, p. 179 (2007)
12. Sinha, S.N., Fram, J.-M., Pollefeys, M., Genc, Y.: Gpu-based video feature tracking and matching. In: Edge Computing Using New Commodity Architectures (2006)
13. Tomasi, C., Kanade, T.: Detection and tracking of point features. Technical Report CMU-CS-91-132, CMU, p. 1 (1991)
14. OpenGL, OpenGL Architecture Review Board: ARB vertex program, Revision 45 (2004), http://oss.sgi.com/projects/ogl-sample/registry/
15. Lecron, F., Mahmoudi, S.A., Benjelloun, M., Mahmoudi, S., Manneback, P.: Heterogeneous Computing for Vertebra Detection and Segmentation in X-Ray Images. International Journal of Biomedical Imaging (2011)
16. OpenCV, OpenCV computer vision library, http://www.opencv.org
17. Bouguet, J.Y.: Pyramidal Implementation of the Lucas Kanade Feature Tracker, Description of the algorithm. Intel Corporation Microprocessor Research (2000)
18. Mahmoudi, S.A., et al.: Traitements d'images sur architectures parallèles et hétérogènes. Technique et Science Informatiques 31, 1183 (2012)
19. Mahmoudi, S.A., Manneback, P., Augonnet, C., Thibault, S.: Détection optimale des coins et contours dans des bases d'images volumineuses sur architectures multicœurs hétérogènes. 20ème Rencontres Francophones du Parallélisme (2012)

20. Mahmoudi, S.A., Manneback, P.: Efficient Exploitation of Heterogeneous Platforms for Images Features Extraction. In: International Conference on Image Processing Theory, Tools and Applications, IPTA (2012)
21. Horn, B.K.P., Schunk, B.G.: Determining Optical Flow. Artificial Intelligence 2, 185 (1981)
22. Mahmoudi, S.A., Lecron, F., Manneback, P., Benjelloun, M., Mahmoudi, S.: GPU-Based Segmentation of Cervical Vertebra in X-Ray Images. In: IEEE International Conference on Cluster Computing, p. 1 (2010)
23. Lucas, B.D., Kanade, T.: An iterative image registration technique with an application to stereo vision. In: Imaging Understanding Workshop, p. 121 (1981)
24. Harris, C.: A combined corner and edge detector. In: Alvey Vision Conference, p. 147 (1988)
25. Mahmoudi, S.A., Lecron, F., Manneback, P., Benjelloun, M., Mahmoudi, S.: Efficient Exploitation of Heterogeneous Platforms for Vertebra Detection in X-Ray Images. In: Biomedical Engineering International Conference, Biomeic 2012, Tlemcen, Algeria, p. 1 (2012)

Feeling Something without Knowing Why: Measuring Emotions toward Archetypal Content

Huang-Ming Chang[1,2], Leonid Ivonin[1,2], Wei Chen[2], and Matthias Rauterberg[2]

[1] CETpD Research Center, Universitat Politècnica de Catalunya
[2] Dept. Industrial Design, Eindhoven University of Technology
{H.M.Chang,L.Ivonin,W.Chen,G.W.M.Rauterberg}@tue.nl

Abstract. To enhance communication among users through technology, we propose a framework that communicates 'pure experience.' This framework can be achieved by providing emotionally charged communication. To initiate this undertaking, we propose to explore materials for communicating human emotions. Research on emotion mainly focuses on emotions that are relevant to utilitarian concerns. Besides the commonly-known emotions like joy and fear, there are non-utilitarian emotions, such as aesthetic emotions, which are essential to our daily lives. Based on Jung's theory of collective unconsciousness, we consider archetypal content as a new category of affective stimuli of non-utilitarian emotions. We collected pictures and sounds of the archetype of the self, and conducted an experiment with existing affective stimuli of utilitarian emotions. The results showed that archetypal content is potential to be a new category of affective content. It is promising to explore other affective content for further studies.

Keywords: affective computing, non-utilitarian emotion, archetypal content.

1 Introduction

As a vision of affective computing, the new class of intelligent system is expected to be capable of capturing emotional responses from users. By analyzing these responses, this system should also be able to generate corresponding information and communicate it expressively, either back to the user herself or to another user [1]. In our previous work [2], we proposed a conceptual framework of communication among users and computers through expanding the richness of the emotional information (see figure 1.). This framework follows the paradigms in psychological experiments, which can be simply broken down to two main processes: emotion recognition and emotion elicitation [3]. In this sense, the driving question appears to be what kind of content can be used to deliver emotional information. As a starting point of this direction, we propose to follow the paradigms in psychological studies on emotion to explore affective content for communicating emotions.

Emotion has been intensively discussed in the field of psychology. However, the mainstream of psychological studies mostly focuses on explicit emotions or utilitarian emotions [4]. These types of emotions can be considered utilitarian in the sense of

M. Mancas et al. (Eds.): INTETAIN 2013, LNICST 124, pp. 22–31, 2013.

facilitating our adaptation to events that have important consequences for our wellbeing, such as anger, fear, joy, disgust, and sadness. On the other hand, there are also other types of emotions that occur without attention or intention [5]. Non-utilitarian emotions are rarely discussed but still play important roles in our daily lives, e.g. aesthetic emotion. This kind of emotions is more delicate and difficult to describe. To enrich the diversity of affective content, we attempt to explore new content for non-utilitarian emotions. In this paper, we introduce the theory of the collective unconscious proposed by psychologist Carl Jung [6]. Based on this theory, we develop a new category of affective content that are considered to be non-utilitarian. An experiment was performed to investigate how people feel about this new category of affective content. The results are discussed in the later section. Then we provide our conclusion and future work in this direction.

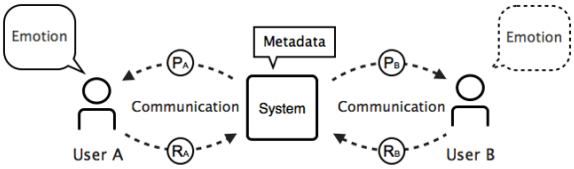


Fig. 1. An intelligent system enables emotionally charged interaction (adapted from [2]). R_A and R_B are emotion recognition; P_A and P_B are presenting stimuli as triggers to induce emotions in the receiver.

2 Psychological Approach to Explore Affective Content

There are many ways to elicit specific emotions under laboratory settings, e.g. hypnosis, imagery, and presenting affective stimuli [7]. For our purpose of exploring affective content for communicating emotions, presenting affective stimuli would probably be the most feasible and straightforward approach. The essence of this technique is to present selected audiovisual material to participants and measure their responses to these stimuli. Unlike the approaches involving confederate interaction procedures, this method may not provide psychological responses of high intensity but it ensures high degree of standardization [3]. As a benchmark for exploring new content, we look for reliable resources that provide affective stimuli with well documented results. Bradley and Lang developed the International Affective Picture System (IAPS) [8], and the International Affective Digital Sound System (IADS) [9], which are two of the broadly-used databases to investigate the correlation between subjects' self-reported emotions and the presented affective stimuli. IAPS and IADS are being developed to provide dimensional ratings of emotions for a large set of emotionally-evocative, internationally-accessible stimuli that include content across a wide range of semantic categories [8, 9].

The critical part of our research question is how we can define the new category of the affective content outside of the scope of existing psychological models of emotion. Psychologist Carl Jung proposed the idea of *collective unconsciousness*, saying that in contrast to the personal psyche, the unconsciousness has some content

and modes of behavior that are identical in all human beings [6]. The collective unconsciousness constitutes a common psychic substrate of a universal nature which is present in every human being. Jung further argued that the collective unconsciousness contains archetypes: ancient motifs and predispositions to patterns of behavior that manifest symbolically as archetypal images in dreams, art or other cultural forms [10]. Jung's theory was of great interests to some scholars. They therefore built an database — the Archive for Research in Archetypal Symbolism (ARAS) [11], which is a pictorial and written archive of mythological, ritualistic, and symbolic pictures from all over the world and from all epochs of human history.

Mandala, a cultural symbol originated in India, was considered as a typical archetypal symbol of *Self* [6]. The very basic form of Mandala is composed of one circle with one dot in its center, which shows the same pattern with the Celtic cross, the halo, the aureole, and rose windows (see Figure 2). Contemporary psychotherapists use Mandala drawing as a basic tool for self-awareness, self-expression, conflict resolution, and healing [12, 13]. Furthermore, it could also be an assessment tool for patients to communicate their physical condition in a non-verbal manner [14]. It seems that the pictures of Mandala are potential to be the new category of affective content especially Mandala does not contain any utilitarian concerns due to the fact that the content of it is merely pure symmetric pattern in a circle.

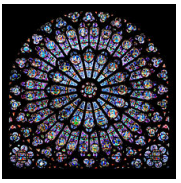
(a)  (b) 

Fig. 2. (a) traditional Tibetan Mandala (book cover of [10]) and (b) the Rose Window in western cultures (the window of the north rose of Notre Dame church, Paris [15])

As for archetypal sounds, very few resources can be found. 'Om' [16] and Solfeggio Frequencies [17] would probably the only resource can be considered as archetypal sounds. 'Om' or 'Aum' is a sacred syllable in Indian religions [16]. 'Om' is the reflection of the absolute reality without beginning or the end and embracing all that exists [18]. Moreover, Solfeggio frequencies are a set of six tones that were used centuries ago in Gregorian chants and Indian Sanskrit chants (i.e. 'Om'.) These chants contained special tones that were believed to impart spiritual blessings during religious ceremonies [17]. Solfeggio frequencies introduce the common fundamental sound that is both used in western Christianity and eastern Indian religions, which strongly resonate on Jung's theory of archetypal symbols. Therefore we included them as archetypal sounds in our study.

In order to perform our experiment in a limited period of time, we needed to categorize the existing stimuli and select proper number of stimuli for each category. We followed a previous study [19], which classified the stimuli in IAPS into four categories: Positive-Arousing (PA), Positive-Relaxing (PR), Neutral (NT), and

Negative (NG). These four categories serve as the benchmark of stimuli of utilitarian emotions for comparison. We added Mandala pictures and Om sounds as the fifth category in our study.

3 Experiment

The experiment followed within-subjects design. Therefore each session only accommodated one participant, and each participant viewed all the stimuli in a random order. At the beginning each participant was asked to sit in front of a monitor for displaying visual stimuli and two speakers for playing audio stimuli. The experiment was built with web-based system. All the experimental data were stored online in the database for further analysis. Before the real experiment started, each participant went through a tutorial to get familiar with the controls and the interface. After the introduction, two sessions started one after the other: picture session and sound session. During each session, the screen or the speakers presented one stimulus at a time for six seconds in a random order. After presenting each stimulus, the interface would pause for five seconds and then show the self-report screen to the participant. Our study followed the method used by IAPS [8] and IADS [9], utilizing the Self-Assessment Manikin scales (SAM) [20] for reporting their current emotional feelings, which consisted of three dimensions: valence, arousal, and dominance. No time limit for the participant to report his or her emotional feelings. Another 5-second pause appeared after the self-report, which was meant to let participants calm down and recover from the previously induced emotion. Then the next picture or sound clip was shown or played. All of the participants ran through the whole procedure individually.

We recruited 37 healthy participants, including 17 males (Mean age=, 26.00, Std. Deviation = 4.062) and 20 females (Mean age=, 27.35, Std. Deviation = 7.322). Most of the participants are students and researchers associated with Eindhoven University of Technology. The participants had diverse nationalities: 17 from Asia (China, India, Indonesia, and Taiwan), nine from Europe (Belgium, the Netherlands, Russia, Spain, and Ukraine), eight from Middle East (Turkey and United Arab Emirates), and three from South America (Colombia and Mexico).

IAPS and IADS contain huge amounts of visual and audio stimuli, including 1194 pictures and 167 sound clips. However, we needed to select a reasonable number of stimuli for conducting a within-subject experiment. Therefore we decided to pick six stimuli for each category to control the duration of each session within one hour. The selection of the stimuli took into account not only the embedded emotional value provided by IAPS and IADS but also the content of the stimuli. This is to ensure the diversity of the stimuli that would enhance the validity of our experiment. For example, the most positive and arousing content (which is PA category) are usually relevant to erotic pictures or sounds. If we had solely selected erotic content as PA category, the validity of this category would have become questionable. We have to include other featured content such as adventure, sports, and delicious food. Although we had tried to pick stimuli in a reasonable manner, the selection might still be biased because the selection work relied on our own judgment.

The same criteria were used to select the materials for the fifth category archetypal content (AR), with the main difference that the distribution of archetypal content in the affective space is not defined yet. To sum up, there were two kinds of media, which were pictures and sound clips; each media type contained five categories, which were mentioned above as PR, PA, NT, NG, and AR; each category comprised of 6 stimuli. In total, 30 pictures and 30 sound clips were selected and included as experimental materials in our experiment (see table 1).

Table 1. An overview of the stimuli used in our experiment. Each medium consists of five categories: Positive-Arousing (PA), Positive-Relaxing (PR), Neutral (NT), Negative (NG), and Archetypal (AR).

Media	Category	Stimuli (Code Number or Name)	Source
Picture	PA	4652, 4668, 7405, 8080, 8179, 8490	IAPS
	PR	1605, 1610, 2060, 2260, 5760, 5891	IAPS
	NT	5530, 7000, 7050, 7090, 7175, 7705	IAPS
	NG	3053, 3170, 6230, 6350, 9321, 9412	IAPS
	AR	3Hc.041, 3Pa.208, 5Ef.007, 7Ao.014, Mandala001 [10], the Wheel of Life [21]	ARAS[11]
Sound	PA	110, 201, 215, 352, 360, 367	IADS
	PR	150, 151, 172, 230, 726, 810	IADS
	NT	171, 262, 376, 377, 708, 720	IADS
	NG	115, 255, 260, 277, 286, 424	IADS
	AR	SF396Hz, SF417Hz, SF528Hz, SF639 Hz, SF741 Hz, Om	[22][23]

4 Results and Discussion

An appropriate statistical test for the design of our experiment would be multivariate analysis of variance (MANOVA) for repeated measures. It showed significant main effects on the variable of 'media type' ($F(3, 34) = 3.596$, $p = 0.023$, Wilks' Lambda) and 'category' ($F(12, 375.988) = 67.870$, $p < 0.001$, Wilks' Lambda). There also exists significance on the interaction between 'media type' and 'category' ($F(12, 375.988) = 4.629$, $p < 0.001$, Wilks' Lambda). Next, we proceeded to look into the test of (univariate) repeated measures ANOVA (Huynh-Feldt) on the variable 'category', the three affective ratings all show significance: valence ($F(3.181, 114.514) = 257.641$, $p < 0.001$), arousal ($F(3.321, 119.546) = 81.302$, $p < 0.001$), dominance ($F(2.414, 86.898) = 28.025$, $p < 0.001$).

We performed the tests of within-subject contrasts on affective ratings to see if the emotions induced by archetypal category are different from other four categories of utilitarian emotions (see table 2). Archetypal content was set as the reference category to be compared. The tests on valence dimension between archetypal content (including both pictures and sounds) and other four categories all show significance. Then we look into the descriptive statistics. For both media (pictures and sounds), the

rating of archetypal content on valence was lower than 'positive relaxing' and 'positive arousing' categories, higher than 'neutral' and 'negative' categories. The explained variance of 'positive relaxing' (η^2=0.780) and 'negative' (η^2=0.912) are remarkably high, whereas 'positive arousing' (η^2=0.275) and 'neutral' (η^2=0.103) are relatively low. Same tests were performed on arousal and dominance dimensions. Along these two dimensions, the results only show significance among 'positive arousing' and 'negative' categories. Then we looked into the descriptive statistics (see table 2). For both media (pictures and sounds), the arousal rating of archetypal category is lower than 'positive arousing' and 'negative' categories; the dominance rating of archetypal category is lower than 'positive arousing' category, but higher than 'negative' category.

Table 2. Statistical results of the affective ratings (valence, arousal, and dominance) on each category of pictures and sounds. Specification of effect column shows the results of the tests of within-subject contrasts on affective ratings, comparing archetypal category (AR) with each of the four categories (PR, PA, NT, and NG).

Rating	Category	Picture		Sound		Specification of effect		
		Mean	Std.Er	Mean	Std.Er	F value	P	η^2
Valence	AR	0.914	0.154	0.446	0.167	-	-	-
	PR	2.279	0.120	1.743	0.127	F(1,36) = 127.905	<0.001***	0.780
	PA	1.509	0.162	1.293	0.159	F(1,36) = 13.629	0.001 ***	0.275
	NT	0.631	0.122	0.248	0.102	F(1,36) = 4.151	0.049 *	0.103
	NG	-2.766	0.151	-2.243	0.119	F(1,36) = 373.370	<0.001***	0.912
Arousal	AR	-0.527	0.193	-0.586	0.196	-	-	-
	PR	-1.095	0.242	-0.608	0.174	F(1,36) = 1.962	0.170	0.052
	PA	1.261	0.180	1.189	0.120	F(1,36) = 130.282	<0.001***	0.784
	NT	-0.833	0.188	-0.203	0.131	F(1,36) = 0.082	0.776	0.002
	NG	1.649	0.214	1.694	0.155	F(1,36) = 188.628	<0.001***	0.840
Dominance	AR	0.324	0.148	0.432	0.185	-	-	-
	PR	0.617	0.216	0.604	0.151	F(1,36) = 2.296	0.138	0.060
	PA	0.901	0.182	0.730	0.145	F(1,36) = 6.191	0.018 *	0.147
	NT	0.423	0.143	0.009	0.090	F(1,36) = 2.929	0.096	0.075
	NG	-1.243	0.281	-1.054	0.205	F(1,36) = 36.969	<0.001 ***	0.507

(* means p value < 0.05, which shows significance; ** means p value < 0.01, which shows high significance; *** means p value <= 0.001, which shows very high significance.)

Scatterplots of the ratings on valence and arousal (see Figure 3) provides a general overview of archetypal content and plot the other four categories in affective space. It needs to be noticed that the distribution of the 'archetypal' category (both pictures and sounds) in the affective space is very close to the 'neutral' category. Moreover, the significance appears to be very weak (p = 0.049) and the explained variance is

relatively low (η^2=0.103). Therefore, we performed an in-depth analysis specifically on the comparison between the 'archetypal' category and the 'neutral' category. It still showed significant main effects on the variable of 'media type' (F (3, 34)=4.218, p = 0.012, Wilks' Lambda) and the interaction between 'media type' and 'category' (F(3, 34)=6.162, p=0.002, Wilks' Lambda). However, the significance on the variable 'category' disappeared (F (3, 34)=1.946, p=0.141, Wilks' Lambda). In order to have a clear view of the difference between the archetypal content and the neutral content, we proceeded to conduct the same test on different media type (pictures and sounds) separately. The results can be found in table 3.

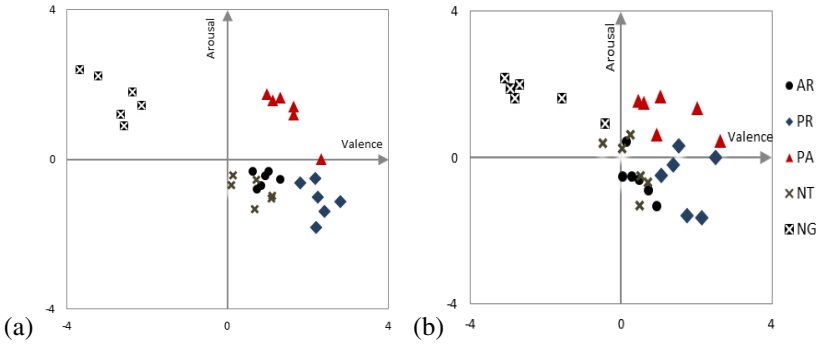


Fig. 3. Scatterplots for participants' rating of valence and arousal on affective pictures (a) and sounds (b). Archetypal (AR), Positive-Relaxing (PR), Positive-Arousing (PA), Neutral (NT), Negative (NG).

Table 3. Statistical results of the affective ratings (valence, arousal, and dominance) for the Archetypal (AR) and the Neutral (NT) category on different media types

Media	Valence	Arousal	Dominance
Picture	F(1,36) = 5.151 p = 0.029 *	F(1,36) = 5.820 p = 0.021 *	F(1,36) = 0.911 p = 0.346
Sound	F(1,36) = 1.321 p = 0.258	F(1,36) = 2.764 p = 1.05	F(1,36) = 7.040 p = 0.012 *

(* means p value < 0.05, which shows significance)

For the media type 'picture', it demonstrated significant effects on the valence dimension (F(1,36) = 5.151, p = 0.029) and the arousal dimension (F(1,36) = 5.820, p = 0.021). For the other media type 'sound', only the dominance value shows significance (F(1,36) = 7.040, p = 0.012). Bringing all the above analysis together, it can be argued that the emotions induced by archetypal content are distinctive from most of the utilitarian emotions. Although the differences between archetypal content and the neutral content are relatively minor, they can still be differentiated if we compare them only with the same media type.

Since our main purpose is to enable an intelligent system to recognize human emotions, we need to build and evaluate predictive models. Because we have already known the previous tests have shown significant main effects on the media type, we

therefore fed the SAM ratings for affective pictures and sounds separately into Linear Discriminant Analysis (LDA), and obtained two predictive models. We present the confusion matrices generated by LDA to evaluate how well the model can predict stimuli of each category based on the data of the SAM scale. Only the cross-validated results are reported. The predictive model derived from LDA on the data for affective pictures obtains 55.8% accuracy and the effect size is large (canonical correlation = 0.770). On the other hand, the predictive model derived from LDA on the data for affective sounds obtains an accuracy of 49.4% and the effect size is also large (canonical correlation = 0.679). Based on the obtained confusion matrices (see table 4 and table 5), we can see that all the four stimuli of utilitarian emotions can be easily recognized, which means that the selected stimuli are nicely chosen to be a benchmark. However, archetypal pictures are more likely to be recognized as the neutral or positive relaxing pictures. Meanwhile, archetypal sounds can be correctly recognized up to 31.1% accuracy. To summarize, although archetypal pictures and sounds are significantly different from other four categories of stimuli of utilitarian emotions, the predictive models seem still not robust enough for emotion recognition.

Table 4. The confusion matrix of the model obtained from LDA on the SAM ratings for the affective pictures [count (percentage)]. Archetypal (AR), Positive-Relaxing (PR) Positive-Arousing (PA), Neutral (NT), Negative (NG). Canonical Correlation = 0.770, Effect Size = Large, 55.8% of the cross-validated grouped cases are correctly classified. The cell with bold numbers means the percentage where the category was correctly predicted.

Category	Predicted Group Membership					Total
	AR	PR	PA	NT	NG	
AR	**27(12.2%)**	54(24.3%)	28(12.6%)	105(47.3%)	8(3.6%)	222(100%)
PR	15(6.8%)	**123(55.4%)**	52(23.4)	27(12.2%)	5(2.3%)	222(100%)
PA	15(6.8%)	32(14.4%)	**137(61.7%)**	15(6.7%)	23(10.4%)	222(100%)
NT	21(9.5%)	35(15.8%)	20(9.0%)	**138(62.2%)**	8(3.6%)	222(100%)
NG	5(2.3%)	2(0.9%)	6(2.7%)	15(6.8%)	**194(87.4%)**	222(100%)

Table 5. The confusion matrix of the model obtained from LDA on the SAM ratings for the affective sounds [count (percentage)]. Archetypal (AR), Positive-Relaxing (PR) Positive-Arousing (PA), Neutral (NT), Negative (NG). Canonical Correlation = 0.679, Effect Size = Large, 49.4% of the cross-validated grouped cases are correctly classified.

Category	Predicted Group Membership					Total
	AR	PR	PA	NT	NG	
AR	**69(31.1%)**	49(22.1%)	21(9.5%)	45(20.3%)	38(17.1%)	222(100%)
PR	29(13.1%)	**102(45.9%)**	53(23.9)	31(14.0%)	7(3.2%)	222(100%)
PA	7(3.2%)	37(16.7%)	**123(55.4%)**	29(13.1%)	26(11.7%)	222(100%)
NT	64(28.8%)	28(12.6%)	31(14.0%)	**67(30.2%)**	32(14.4%)	222(100%)
NG	3(1.4%)	5(2.3%)	17(7.7%)	10(4.5%)	**187(84.2%)**	222(100%)

According to the results of our study, it is clear that archetypal stimuli (AR) are distinctive from most of the stimuli of utilitarian emotions (e.g. PA, PR, and NG). Although we can still differentiate AR and NT, it appears that the emotional responses to these two categories seemed to be very similar to each other. Nevertheless, we do not consider it as a minor finding. Instead, this can lead to many interesting discussions. First, since the SAM scale was designed focusing on the investigation of utilitarian emotions, it is unclear if it is capable of differentiating non-utilitarian emotions. Moreover, IAPS contains some stimuli of 'abstract art' that are also considered to be neutral (e.g. No.7192 in IAPS). In this sense, it seems that we need more dimensions in affective space for representing non-utilitarian emotions. Second, Jung claims that archetypes are hidden in the collective unconsciousness and cannot be accessed consciously. Although Mandala and Om are embodied the archetype of self as visual and audio stimuli, it is probably difficult for subjects to consciously introspect their emotions toward archetypal symbols. Regarding this issue, some previous study has suggested using physiological measurement for recognizing implicit emotions [24]. It seems promising to apply this approach for future studies in this direction.

5 Conclusion

Our study aims at exploring new affective content to enhance the richness of emotional information that can be used for human-computer interaction. Our preliminary findings demonstrated that archetypal symbolism could be a new resource for developing new affective content for designing emotionally charged communication. Besides the archetype of the 'self', many other kinds of archetypal content is still available, e.g. hero and shadow. Emotions that induced by these contents are still unknown. It would be a promising direction to investigate the emotional qualities induced by archetypal content, and utilize the findings to design a better media to communicate 'pure experiences' for mental well-being.

Acknowledgement. This work was supported in part by the Erasmus Mundus Joint Doctorate in Interactive and Cognitive Environments, which is funded by the EACEA Agency of the European Commission under EMJD ICE FPA n 2010-0012.

References

1. Scheirer, J., Picard, R.W.: Affective Objects. MIT Media Laboratory Perceptual Computing Section Technical Report No. 524 (1999)
2. Chang, H.-M., Ivonin, L., Chen, W., Rauterberg, M.: Lifelogging for hidden minds: Interacting unconsciously. In: Anacleto, J.C., Fels, S., Graham, N., Kapralos, B., Saif El-Nasr, M., Stanley, K. (eds.) ICEC 2011. LNCS, vol. 6972, pp. 411–414. Springer, Heidelberg (2011)
3. Rottenberg, J., Ray, R.D., Gross, J.J.: Emotion elicitation using films. In: Coan, J.A., Allen, J.J.B. (eds.) Handbook of Emotion Elicitation and Assessment, pp. 9–28. Oxford University Press, USA (2007)

4. Scherer, K.R.: What are emotions? And how can they be measured? Social Science Information 44, 695–729 (2005)
5. Bargh, J.A., Morsella, E.: The unconscious mind. Perspectives on Psychological Science 3, 73–79 (2008)
6. Jung, C.G.: The Archetypes and the Collective Unconscious. Princeton University Press, Princeton (1981)
7. Gross, J.J., Levenson, R.W.: Emotion elicitation using films. Cognition and Emotion 9, 87–108 (1995)
8. Lang, P.J., Bradley, M.M., Cuthbert, B.N.: International affective picture system (IAPS): Affective ratings of pictures and instruction manual. Technical Report A-8. pp. 1–12. University of Florida, Gainesville, FL (2008)
9. Bradley, M.M., Lang, P.J.: The international affective digitized sounds (IADS-2): Affective ratings of sounds and instruction manual. Technical Report B-3. pp. 29–46. Oxford University Press, USA, Gainesville, FL (2007)
10. Jung, C.G.: Man and His Symbols. Doubleday, Garden City (1964)
11. Gronning, T., Sohl, P., Singer, T.: ARAS: Archetypal Symbolism and Images. Visual Resources 23, 245–267 (2007)
12. Kim, S., Kang, H.S., Kim, Y.H.: A computer system for art therapy assessment of elements in structured mandala. The Arts in Psychotherapy 36, 19–28 (2009)
13. Schrade, C., Tronsky, L., Kaiser, D.H.: Physiological effects of mandala making in adults with intellectual disability. The Arts in Psychotherapy 38, 109–113 (2011)
14. Elkis-Abuhoff, D., Gaydos, M., Goldblatt, R., Chen, M., Rose, S.: Mandala drawings as an assessment tool for women with breast cancer. The Arts in Psychotherapy 36, 231–238 (2009)
15. Wikipedia: Rose window - Wikipedia, the free encyclopedia, `http://en.wikipedia.org/w/index.php?title=Rose_window&oldid=510855813`
16. Wikipedia contributors: Om, `http://en.wikipedia.org/wiki/Om`
17. Wikipedia contributors: Solfeggio frequencies, `http://en.wikipedia.org/wiki/Solfeggio_frequencies`
18. Maheshwarananda, P.S.S.: The hidden power in humans: Chakras and Kundalin. Ibera Verlag (2004)
19. Ribeiro, R.L., Teixeira-Silva, F., Pompéia, S., Bueno, O.F.A.: IAPS includes photographs that elicit low-arousal physiological responses in healthy volunteers. Physiology & Behavior 91, 671–675 (2007)
20. Bradley, M.M., Lang, P.J.: Measuring emotion: The self-assessment manikin and the semantic differential. Journal of Behavior Therapy and Experimental Psychiatry 25, 49–59 (1994)
21. Bhavacakra - Wikipedia, the free encyclopedia, `http://en.wikipedia.org/wiki/Wheel_of_Existence`
22. Welch, D.: mountainmystic9's Channel, `http://www.youtube.com/user/mountainmystic9`
23. ONEMIND4U: Om meditation, `http://www.youtube.com/watch?v=imWRQpY0P58`
24. Bechara, A., Damasio, H., Tranel, D., Damasio, A.R.: The Iowa Gambling Task and the somatic marker hypothesis: Some questions and answers. Trends in Cognitive Sciences 9, 159–162 (2005)

Web and TV Seamlessly Interlinked: LinkedTV*

Lyndon Nixon**

STI International GmbH
Neubaugasse 10/15, 1070 Vienna, Austria
lyndon.nixon@sti2.org

Abstract. This paper reports on the vision of LinkedTV driven by the EU project of the same name[1], and the work done in its first year. LinkedTV is a new type of television (or audio-visual) experience where Web and TV content can be seamlessly interlinked based on the concepts present within that content. The project addresses how the Web and TV is converging in end devices, and particularly this paper focuses on how we intend to answer the research challenges that the LinkedTV vision raises.

Keywords: Smart TV, Networked Media, semantic multimedia, media annotation, Connected TV, future TV.

1 Introduction

Networked Media will be a central element of the Next Generation Internet. Online multimedia content is rapidly increasing in scale and ubiquity, yet today it remains largely still unstructured and unconnected from related media of other forms or from other sources. This cannot be clearer than in the current state of the Digital "Smart" TV market. The full promise and potential of Web and TV convergence is not reflected in offerings which place the viewer into a closed garden, or expect PC-like browsing of the Web on a distant TV screen, or extend the television with new functionalities which however lack any relation to the currently viewed TV programming.

Our vision of future **Television Linked To The Web (LinkedTV)** is of a ubiquitously online cloud of Networked Audio-Visual Content decoupled from place, device or source. Accessing audio-visual programming will be "TV" regardless whether it is seen on a TV set, smartphone, tablet or personal computing device, regardless of whether it is coming from a traditional or new media broadcaster, a Web video portal or a user-sourced media platform. Television existing in the same

* The other LinkedTV consortium partners are: RBB Rundfunk Berlin Brandenburg (Germany), Sound and Vision (Netherlands), University of Mons (Belgium), CONDAT (Germany), Noterik (Netherlands), Fraunhofer IAIS (Germany), CERTH-ITI (Greece), EURECOM (France), University of Economics Prague (Czech Rep), CWI (Netherlands), University of St Gallen (Switzerland).

** The LinkedTV Consortium.

[1] www.linkedtv.eu, Twitter@linkedtv

M. Mancas et al. (Eds.): INTETAIN 2013, LNICST 124, pp. 32–42, 2013.

ecosystem as the Web means that television content and Web content should and can be seamlessly connected, and browsing TV and Web content should be so smooth and interrelated that in the end even "surfing the Web" or "watching TV" will become as meaningless a distinction as whether the film is coming live from your local broadcaster, as VOD from another broadcaster, or from an online video streaming service like Netflix. As a result, not only commercial opportunities but also opportunities for education, exploration and strengthening European society and cultural heritage arise. Imagine browsing from your local news to Open Government Data about the referenced location to see voting patterns or crime statistics, or learning more about animals and plants shown in the currently viewed nature documentary without leaving that show, or jumping from the fictional film to the painting the character just mentioned to virtually visiting the museum when it can be seen, or seamlessly accessing additional information that has been automatically aggregated from multiple sources in order to get better informed on an important event that was just mentioned in the news.

Technologically, this vision requires systems to be able to provide networked audio-video information usable in the same way as text based information is used today in the original Web: interlinked with each other at different granularities, with any other kind of information, searchable, and accessible everywhere and at every time. Ultimately, this means creating hypermedia at the level of the Web whose original success was the underlying hypertext paradigm built into HTML. Hypermedia has been pursued for quite a while as an extension of the hypertext approach towards video information. This requires suitable descriptive models of media that allow for its interlinking, as well as client applications able to process and play out hypermedia based on those descriptions, but to avoid a fully manual and hence not scalable approach for the scale of the Web, it needs complex media analysis algorithms and is still an open issue of research. The **Television Linked To The Web (LinkedTV)** project aims at a novel *practical* approach to Future Networked Media based on four phases: annotation, interlinking, search, and usage (including personalization, filtering, etc.).

The rest of the paper is structured as follows: Section 2 introduces the scenarios of LinkedTV, which motivate our vision and will act as the basis for prototypes. Section 3 outlines the LinkedTV architecture and player, while Section 4 references the research challenges within the project and how we aim to answer them. Finally Section 5 concludes with the outlook for the realisation of LinkedTV as part of every citizen's future experience of television.

2 LinkedTV Scenarios

LinkedTV will demonstrate its vision of weaving of television and the Web through three scenarios, each of which representing different aspects of the value and potential of the future Networked Media Web. These are a current affairs scenario, a documentary scenario, and a media artist scenario.

Current Affairs Scenario

In general, the envisaged service targets a broader audience. For the sake of a convincing scenario, however, we have sketched a few fictional users of the LinkedTV news service and their motivations to use it. For example, socially active retiree **Peter** watches the news show "rbb AKTUELL". One of the spots is about a fire at famous Café Keese in Berlin. Peter is shocked. He used to go there every once in a while, but that was years ago. As he hasn't been there for ages, he wonders how the place may have changed over the years. In the news spot, smoke and fire engines was almost all one could see, so he watches some older videos about the story of the famous location where men would call women on their table phones – hard to believe nowadays, now that everyone carries around mobile phones! Memories of these good old days make him happy and sad at the same time. After checking these very nice clips on the LinkedTV service, he returns to the main news show and watches the next spot on a new Internet portal about rehabilitation centres in Berlin and Brandenburg. He knows an increasing number of people who need such facilities. He follows a link to a map of Brandenburg showing the locations of these centres and bookmarks the linked information to check again later.

Documentary Scenario

In the documentary scenario, we have storyboarded with the persona Rita, an administrative assistant at the Art History department of the University of Amsterdam. She didn't study art herself, but spends a lot of her free time on museum visits, creative courses and reading about art. One of her favourite programmes is the Antiques Roadshow (Dutch title: Tussen Kunst & Kitsch), which she likes to watch because, on the one hand, she learns more about art history, and on the other hand because she thinks it's fun to guess how much the objects people bring in are worth. She's also interested in the locations where the programme is recorded, as this usually takes place in a historically interesting location, such as a museum or a cultural institute.

Rita is watching the latest episode of the Roadshow. The show's host, Nelleke van der Krogt, gives an introduction to the programme. Rita sees the show has been recorded in the Hermitage Museum in Amsterdam. She always wanted to visit the museum as well as finding out what the link is between the Amsterdam Hermitage and the Hermitage in St. Petersburg. She sees a shot of the outside of the museum and notices that it was originally a home for old women from the 17th century. Intriguing! Rita wants to know more about the Hermitage location's history and see images of how the building used to look. After expressing her need for more information, a bar appears on her screen with additional background material about the museum and the building in which it is located. While Rita is browsing, the programme continues in a smaller part of her screen. After the show introduced the Hermitage, a bit of its history and current and future exhibitions, the objects brought in by the participants are evaluated by the experts. One person has brought in a golden, filigree box from France in which people stored a sponge with vinegar they could sniff to stay awake

Fig. 1. Chi-Ro symbol in the TV program

during long church sermons. Inside the box, the Chi Ro symbol has been incorporated (Figure 1). Rita has heard of it, but doesn't really know much about its provenance and history.

Again, Rita uses the remote to access information about the Chi Ro symbol on Wikipedia and to explore a similar object, a golden pyx with the same symbol, found on the Europeana portal. Since she doesn't want to miss the expert's opinion, Rita pauses the programme only to resume it after exploring the Europeana content. The final person on the show (a woman in her 70s) has brought in a painting that has the signature 'Jan Sluijters'. This is in fact a famous Dutch painter, so she wants to make sure that it is indeed his. The expert - Willem de Winter - confirms that it is genuine. He states that the painting depicts a street scene in Paris, and that was made in 1906. Rita thinks the painting is beautiful, and wants to learn more about Sluijters and his work. She learns that he experimented with various styles that were typical for the era: including fauvism, cubism and expressionism. She'd like to see a general overview of the differences of these styles and the leaders of the respective movements.

During the show Rita could mark interesting fragments by pressing a button on her remote control. While tagging she continued watching the show but afterwards these marked fragments are used to generate a personalized extended information show based on the topics Rita has marked as interesting. She can watch this related / extended content directly after the show on her television or decide to have this playlist saved so she can view it later. This is not only limited to her television but could also be a desktop, second screen or smartphone, as long as these are linked together. She's able to share this information on social networks, allowing her friends to see highlights related to the episode.

The Media Art Scenario

This scenario has focused on the other hand on "personalized remixing of TV" using several feature dimensions. Features can be extracted from two main sources. The first one is the automatic analysis of the video itself. We use state-of the art frameworks to segment the videos into scenes. For each scene, three kinds of features can be extracted concerning object detection, event detection and emotion detection. On the other side, features can be extracted from behavioural observation, especially

by using RGBD cameras (like the Microsoft Kinect). Those features can bring information about viewer's attention or change in behaviour (was not moving but now he is moving, was talking but now he stopped and looks towards the TV, etc.). During playout of video content with a structure it is not possible to adapt the video itself to the viewer as parts of the video cannot be avoided or moved without breaking the scenario continuity. In this context automatically segmented scenes can be stored during playout to be used after the content visualisation or during a pause (advertisement or user defined program pause). The scenes to be stored can be either automatically stored (they contain objects of interest for the user as described in his profile, the user had a sudden behaviour change, …) or be manually stored (by using a specific gesture recognized by an RGBD camera).

A mock-up of the interface could be the one in Figure 2 with the main content in the middle and the band of scenes with their characteristics (activity detection, objects, viewer reaction). This menu could also be located on a second screen to avoid distracting too much the viewer. During pauses or after the content is finished, the user can access to all the scenes which were saved or the ones he manually saved during playout. The viewer has access to the enrichment of those videos (hyperlinks, links to other videos…) based on the three kinds of features which were extracted (object-based, event-based and emotion-based). The viewer can than keep a subset of scenes which summarizes in a personal way the video he saw with additional content he added from the enrichment links. This can be sent towards social networks. The information coming from these summaries can be used to augment the personal profile for more efficient enrichment personalisation.

The scenes coming from already seen content and additional content from automatic enrichment can finally be used as a scene database for semi-automatic remixes and mash-ups. An initial video scene is selected and placed in the centre of the screen while 3 clusters are formed around using object-based, event-based and emotion-based features to compute the similarity to the seed video scene. The viewer can than mix the current seed with another video from one of the three clusters as in the figure below.

Fig. 2. Mock interface for the media art scenario

3 LinkedTV Architecture

To realize all these scenarios an architecture for a LinkedTV infrastructure, integrating different media technologies into an agreed workflow, has been developed based on analysis of the use case requirements, technical feasibility and identification of the project research goals.

The LinkedTV platform analyses and annotates external videos, generates media fragments and enriches them semantically with external information from the Web and Linked Open Data sources. The annotated tags and links can be adjusted and enhanced by an editor through tools for annotation and hyperlinking. Based on the enriched and interlinked media fragments various user applications with personalized user interfaces for clickable video allow the user to access the provided videos. Within LinkedTV three different user scenarios demonstrate the possibilities enabled by the LinkedTV approach. This includes a Web client variant using a Browser with the full potential of HTML5 for clickable video based on a two-way HTTP communication and a second client variant, using HbbTV, HTTP and a TV-set with reduced functionality.

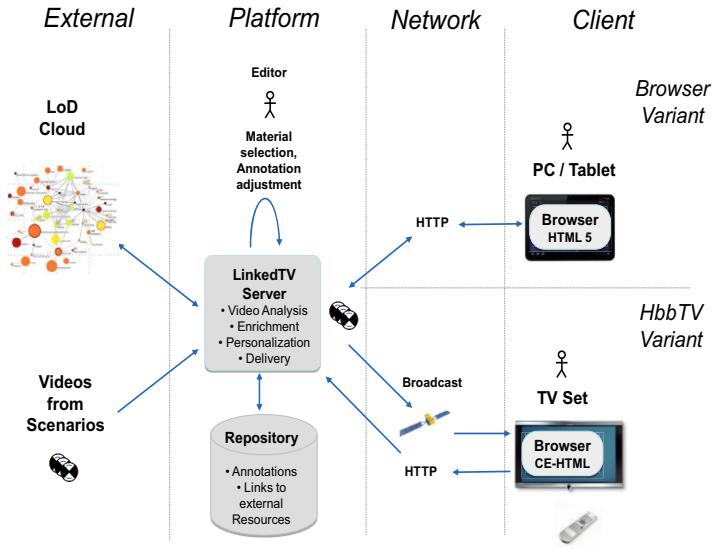


Fig. 3. LinkedTV System Overview

The LinkedTV platform employs a Service Oriented Architecture (SOA) with a division in three layers each consisting of several components. The use of REST Services ensures an efficient, flexible and fault tolerant communication between all components. The SOA architecture allows together with the adoption of standards for formats, interfaces and protocols the exchange of platform components by third party software, distributed development of components, scalability and multi-lingualism.

The LinkedTV Platform is divided in three main layers: 1) the Analysis and Annotation Layer containing the components and interfaces for the analysis, annotation and enrichment components, 2) the Presentation Layer containing the components for developing personalized LinkedTV applications for end users and 3) the Linked Media Layer containing all the metadata generated including the services to access them, as well as management tools.

Connected to the Analysis and Annotation Layer there are editorial tools for 1) the Media Selection and Analysis Tool to select new videos for analysis and include them into the platform 2) the Annotation Tool to adjust automatically generated annotations and 3) the Hyperlinking Tool for the manual insertion of links associated to certain objects in the video.

The Presentation Layer provides the basis for the development of specific end user components and applications: 1) a Hypervideo Player for HTML5 to retrieve and view hyperlinked video, 2) a HbbTV compliant player for TV applications and 3) Specific applications to perform the LinkedTV Scenarios with individual user interfaces and features such as the media use case.

The first step has been to develop the HTML5 based Hypervideo Player. It combines media fragments together with annotations from the Annotation Layer and displays these on the video canvas. Video hotspots are used to allow editors and end users to find objects using layered technology. The visual hotspots also provide access to related content and other available media on the Web. The HTML5 technology allows the Hypervideo Player to be used on a broad range of systems, including PCs, Smart TVs and mobile devices.

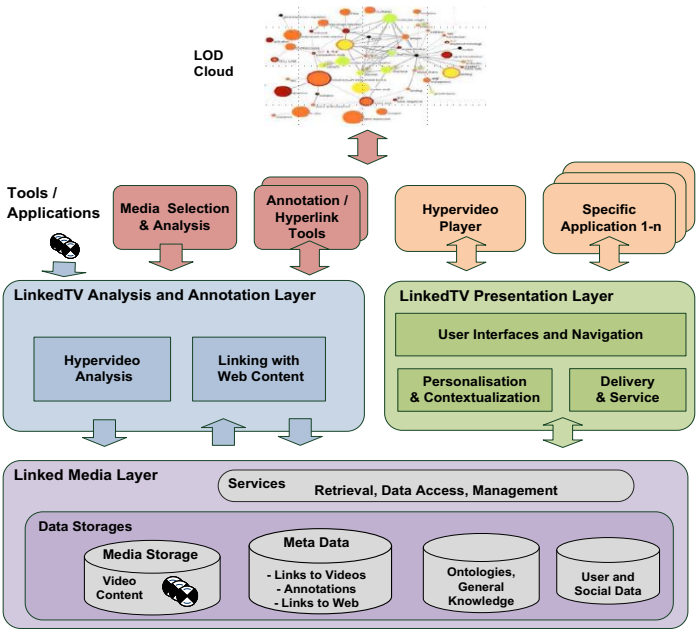


Fig. 4. Architecture overview

Fig. 5. Hypervideo player with hotspot

4 Research Challenges

The described architecture incorporates a number of components that must provide a certain level of quality and efficiency around specific functionalities that become the subject of project research challenges. These challenges can be categorized into the domains of (i) media analysis, (ii) cross-media interlinking, (iii) LinkedTV user interfaces and navigation, and (iv) LinkedTV personalisation and contextualisation.

Media Analysis [1]. Based on the scenarios envisioned within LinkedTV, we need to be (semi-)automatically deriving from the A/V content information about the concepts being present in that content. For example, persons could be identified via face and/or voice, and objects of interest should be detected and tracked. Due to the broad target domains it cannot be guaranteed that established (domain specific) databases of annotated media (for helping classifiers identify objects in new video material) contain enough instances of concepts present in new video material to analyse. This is why we need strong clustering and re-detection techniques so that an editor only needs to label a concept once and can automatically find other instances within the video itself, or within a larger set of surrounding videos. For keywords needed to tag the videos, both more abstract concepts and events should be recognized. Also needed for keywords, as well as for the named entity recognition, is an automatic speech recognizer whenever there are no subtitles available, or, if subtitles are given, forced alignment techniques to match the timestamps to the video on a word level rather than on an utterance level which might be to coarse-granular for our needs. Finally, larger videos should be temporally segmented based on their content to provide reasonable time-stamp limits for hyperlinks.

The partners of the LinkedTV consortium have access to state-of-the-art techniques that can pursue these goals. The main challenge will therefore be to interweave the individual analysis results into refined high-level information. For example, person detection can gain information from automatic speech recognition, speaker recognition and face recognition. Also, in order to find reasonable story segments in a larger video, one can draw knowledge both from speech segments, topic classification, and video shot segments. As a final example, video similarity can be

estimated with feature vectors carrying information from the concept detection, the keywords, the topic classification and the entities detected within the video.

In a first step, we already collected all semi-automatically produced analysis results into a single annotation file that can be viewed with a dedicated tool, and will use this in order to establish ground truth material on the scenario data. We already validated and extended various methods – object re-detection, shot segmentation, face detection and keyword extraction - and increased their accuracy in LinkedTV scenarios[2]. A project deliverable summarizes all first year achievements [7].

Cross-Media Interlinking. LinkedTV will use an ontology-based data model in order to represent all the information about television programs that needs to be managed by the system. This data model will be based on various well-established vocabularies from the multimedia and television domain.

First, LinkedTV receives legacy data either coming from the content itself or produced by the broadcaster. This data generally includes properties such as title, short description, format, duration, etc. that will be represented using the Ontology for Media Resources [2]. It includes as well broadcasting information that can be represented according to the BBC ontology [3] and the SchemaDotOrgTV vocabulary [4]. Second, the results of the analysis performed over the video in previous stages (shot detection, face recognition, ASR, etc.), which are currently available in XML like formats will be RDF-ized using appropriate additional ontology properties. The decomposition of media content into pieces will be directly addressed by the use of Media Fragment URIs [5]. Finally, the Open Annotation Core Data Model [6] will enable to associate real world concepts as annotations to media fragments and to state more information about this annotation such as its provenance or level of trust. In most of the cases, this information is scarce, incomplete and sometimes inaccurate. It is then necessary to retrieve additional contextual information from different external sources. These sources will mainly be datasets from the LOD cloud such as the BBC Programmes, LinkedMDB or Geonames, as well as various Social Networks where fresh media, reactions and opinions are available.

We have established the following workflow for managing the metadata in the LinkedTV system: first, the legacy metadata and the analysis results are serialized into RDF by performing certain translation processes between the original files and the corresponding vocabularies[3]. During this phase, some NER techniques are also applied over the ASR files in order to extract named entities from the video[4]. Then, once this data is already available in the knowledge base, LinkedTV iteratively executes background operations for retrieving missing information, enrich existing data, and interlink local entities with similar ones in other external datasets. At the end of the complete process, a complete RDF graph can be accessed by the LinkedTV player in order to show the viewer the information he needs. The LinkedTV Ontology was defined in a project deliverable [8].

---

[2] Several video demos are posted at `http://www.linkedtv.eu/demos-materials/online-demos/#core-media`

[3] See the online RDF Metadata Generator `http://linkedtv.eurecom.fr/tv2rdf`

[4] Using NERD `http://nerd.eurecom.fr`

User Interfaces and Navigation. Users are familiar with interaction with video in terms of using text to search for video fragments on the internet and of navigating the timeline of video using concepts such as "fast-forward". Users are also used to navigating the plethora of web pages on the internet, using search engines and bookmarks to find information they are looking for. The goal of LinkedTV from the user perspective is that s/he should be able to manipulate video material combining these two interaction paradigms. The challenge for the project is to ensure that the information, and entertainment, experience is enhanced, rather than frustrated by being presented with a bewildering assembly of vaguely related videos, where they lose both the advantages of passively watching edited video or being in control of finding specific clips of their own choice. Our research questions are to what extent we can provide links to usefully related video content without disrupting the viewing experience. At the same time we need to be able to provide unobtrusive interaction tools that allow the material to be explored freely. This led to a particular focus on second screen solutions. The Antiques Roadshow and news scenarios were used to explore the commonalities of the tools needed and specific ways of presenting the combination of functionalities to users in an appropriate LinkedTV UI [9].

As TV becomes more interactive, it moves closer to a gaming environment. While the goals of LinkedTV are specifically related to providing information to users, interaction devices more familiar in the gaming environment can be used to enhance the exploration experience. We also explore where these interaction devices, such as Kinect, can help the user retain a sense of control in a complex linked and time-based information environment. An example is a dance exploration scenario where we will explore the potential of interacting with the video space by dancing.

Personalisation and Contextualisation. The provision of recommendations for Media Fragments to end users requires the enhancement of current recommender technology. Innovative methods to derive semantic fingerprints from the fragment properties, surrounding objects, current scene and spoken text are needed to provide users with personalized related information while watching TV. LinkedTV has published a first user schema for capturing viewer's interests and a set of approaches to implicitly gather those interests via user interaction with the LinkedTV content as well as tracking attention and emotions during watching via a Kinect installation [10].

LinkedTV plans to show hyperlinked video on tablets and TV Sets. After a first demonstrator for TVs and tablets using HTML5[5] we want to show clickable video for TV devices based on hbbTV. This will require innovative solutions for features not initially provided by HTML5 (e.g. access to resources such as a broadcast video stream and its metadata) or hbbTV 2.0 (e.g. hyperlinks placed over dynamically moving objects or alternative interaction models such as pointing devices based on gesture control or second screens). This summary of research challenges reflects the need for the cross-European collaborative work of expert organisations in the respective domains, which is being enabled by the LinkedTV EU project.

5 See a screencast of the 1st year demo at `http://www.linkedtv.eu/demos-materials/online-demos/#scenarios`

5 Outlook for LinkedTV

Television is changing. However, uptake of the new Smart TVs will increase at a gradual rate and currently the Web-TV offering is not capturing the interest of the critical mass of viewers. Key growth is seen presently in the younger adult demographic with use of social TV apps in a second screen device. Finally, industry lock-in is largely seeking to limit third party OTT (over the top) services that would take revenue from existing customers (e.g. cable subscribers). In the next years connected TV platforms will not only become more present but the quality and intuitiveness of their UI experience will improve. The legacy media – TV and video content - industry will either co-opt the new technology within their own controlled services or be pushed out by innovative OTT services (cf. the history of the music industry and the Web). The LinkedTV vision – seamlessly interlinking Web and TV content in an unified interactive, audio-visual experience on the end device – has both technological and business barriers to overcome. The LinkedTV project is working on overcoming the technological barriers, and will also monitor the shifting TV business landscape. As the project comes to an end in April 2015, TV will already be very different from it has been traditionally, and LinkedTV will be ready to promote its vision for television, connected to the Web, its metadata and content, and to innovative services for analysis, annotation, linking, personalisation and presentation of such LinkedTV content.

Acknowledgements. This work is supported by the Integrated Project LinkedTV (www.linkedtv.eu) funded by the European Commission through the 7th Framework Programme (FP7-287911).

References

1. Stein, D., Apostolidis, E., Mezaris, V., de Abreu Pereira, N., Müller, J.: Semi-Automatic Video Analysis for Linking Television to the Web. In: Proc. FutureTV Workshop, at EuroITV Conference 2012, Berlin, Germany (July 2012)
2. http://www.w3.org/TR/mediaont-10/
3. http://www.bbc.co.uk/ontologies/programmes/2009-09-07.shtml
4. http://www.w3.org/wiki/SchemaDotOrgTV
5. http://www.w3.org/TR/media-frags/
6. http://www.openannotation.org/spec/core/
7. Mezaris, V., et al.: State of the Art and Requirements Analysis for Hypervideo. LinkedTV project deliverable D1.1 published at http://linked.tv/deliverables
8. García, J.L., Tröncy, R., Vacura, M.: Specification of lightweight metadata models for multimedia annotation. LinkedTV project deliverable D2.2 at http://linked.tv/deliverables
9. Leyssen, M., Traub, M., Hardman, L., van Ossenbruggen, J.: LinkedTV user interfaces sketch. LinkedTV project deliverable D3.3 at http://linkedtv.eu/deliverables
10. Tsatsou, D., Mezaris, V., Kliegr, T., Kuchar, J., Mancas, M., Nixon, L., Klein, R., Kober, M.: User profile schema and profile capturing. LinkedTV project deliverable D4.2 at http://linkedtv.eu/deliverables

VideoHypE: An Editor Tool for Supervised Automatic Video Hyperlinking

Lotte Belice Baltussen[1], Jaap Blom[1], and Roeland Ordelman[1,2]

[1] Netherlands Institute for Sound and Vision, Hilversum, The Netherlands
[2] University of Twente, Enschede, The Netherlands

Abstract. Video hyperlinking is regarded as a means to enrich interactive television experiences. Creating links manually however has limitations. In order to be able to automate video hyperlinking and increase its potential we need to have a better understanding of how both broadcasters that supply interactive television and the end-users approach and perceive hyperlinking. In this paper we report on the development of an editor tool for supervised automatic video hyperlinking that will allow us to investigate video hyperlinking in a real-life scenario.

Keywords: video hyperlinking, interactive television, video analysis, user studies, information extraction.

1 Introduction

Aiming towards richer interactive television experiences, broadcast companies are becoming increasingly interested in enriching television content with hyperlinks that connect the primary content to other data sources that could enhance the attractiveness of watching television in either a linear fashion or on-demand. By default, such links are currently created manually by broadcast companies' editorial departments. However, the identification of media-fragments [8] that can be used as *anchors* in the primary content, and the selection of appropriate link *targets* completely manually is labour intensive. Automating at least parts of the hyperlink generation process for television content is essential to be able to provide the users with a rich set of links that significantly enhance the user experience. A second argument for pulling technology into the link generation process is that manual hyperlink generation is inherently limited by a human editor's subjective view on anchor selection and limited view on possible target data sources. Note that the latter in return also confines the anchor selection process.

A baseline approach towards the automation of link generation is to deploy available time-labelled content descriptors such as subtitles, and automatically generated annotations (e.g., automatic speech recognition, visual concept detection). Entities such as names, people, places and objects can be extracted from the textual annotations to serve as anchor "candidates" that can be linked to other content sources. A curated white list of resources may define the domains that are considered for the link targets. The final step would then be to connect

M. Mancas et al. (Eds.): INTETAIN 2013, LNICST 124, pp. 43–48, 2013.

the link anchors to the link targets. For example, consider a programme that is about *Rembrandt* and his painting *Night Watch*, which is displayed in the *Rijksmuseum* in *Amsterdam*. The entities "Rembrandt", "Night Watch", "Rijksmuseum" and "Amsterdam" can be extracted from the subtitles and could be identified unambiguously using Semantic Web URIs (Linked Data) [10] and linked to the Wikipedia articles about Rembrandt and the Night Watch, assuming that Wikipedia is part of the white list of link sources.

Although the use of a white list restricts the amount of anchors that can practically be linked, it is not the case that every anchor in a white list that *can* be linked is also a useful candidate. One can imagine that the usefulness of a candidate anchor depends among others on the content itself, user context, and the characteristics of the link target, such as its specificity and relevance. Ideally, the relevance of candidates for linking is determined automatically on the basis of context features. However, we need to have a better understanding of user behaviour in a video linking scenario to model the relevance of link candidates appropriately. We investigate user behaviour in a scenario in which an editorial department of a broadcast company creates links for an interactive television application, and evaluate how automatic link suggestion is controlled and perceived by the editors of the programme and how end-users that watch the programme in the interactive television application appreciate the links. In this paper we describe the first stage of the development of video hyperlink editor (*VideoHypE*) that uses rich, and partly automatically extracted content annotations for *supervised* hyperlink generation. The tool will be used and evaluated in real-life editorial link generation scenarios at broadcast companies.

The rest of this paper is structured as follows. In the next section (section 2) we briefly describe the underlying technology of the *VideoHypE* tool. In section 3 we report on a requirements elicitation session with editorial staff of a broadcaster interested in video hyperlinking and close with a future work section (section 4).

2 Link Suggestion Workflow

The envisaged *VideoHypE* tool will use the input of a processing chain incorporating audio and visual analysis and information extracting. The chain consists of shot detection for video segmentation into shots [6], spatial-temporal and visual concept detection [2], face detection [7], face clustering and face recognition for respectively detecting, grouping and tagging faces of persons in a video, and object re-detection for retrieving instances of pre-selected images within video frames. With respect to audio analysis technologies we will use speaker identification [4] for identifying certain pre-selected persons of interest related to the user scenarios, automatic speech recogintion (ASR) [5], and audio fingerprinting for media synchronisation, e.g. for synchronising a second screen application with the main screen.

Based on ASR transcripts, existing programme metadata and subtitles, named entities are extracted, disambiguated and finally enriched with related content

(generally linked data resources). Combined with the previously compiled annotations the end-result of a video being processed is a number of annotated media fragments stored as RDF triples in a Virtuoso semantic repository[1]. The VideoHypE tool is built on top of this semantic repository and includes mechanisms to assess the relevance of the automatically generated links and push them to a user interface [1].

3 Hyperlink Editor Requirements

In order to develop a *VideoHypE* tool that is consistent with the needs of broadcasters engaging into video hyperlinking, we need user requirements from the editors of the interactive television applications that are going to use the tool in a real-life scenario. We are currently looking at two use scenarios for hyperlink suggestion: an interactive news scenario based on news content from a German broadcaster[2], and interactive documentary scenario on a Dutch programme called "Tussen Kunst en Kitch" (TKK), an antiques roadshow from Dutch public broadcaster AVRO. In this paper we focus on the scenario for the TKK programme.

To elicit requirements for the tool, we organised an interview session with the editorial members and technical staff of the TKK programme (further referred to as TKK staff) in which we explained the concept of extracting media fragments and suggesting hyperlinks and asked them about their view on addressing hyperlinking in such a manner. Point of departure was a mutual agreement on the fact that automatic video hyperlinking could be a helpful tool in an interactive television application scenario, either fully automated or in an editorial setting. With an eye on the current experimental state of technology, the TKK staff expects that some kind of supervision on the links that are created will always be required. One important reason for keeping control over the links is that broadcast companies fear that providing links that are not appreciated by end-users may harm the appreciation of the television programme.

The corrections of the hyperlinks can be done on several levels. First of all, the editor could merely accept and reject the hyperlinks. However, an extra level of control could be added that allows editors to also *add* hyperlinks not provided by the system. Another important aspect of the editor tool will be the granularity on which the work will take place. The system can provide hyperlinks on the level of shots, scenes or even spatial elements within shots and scenes (e.g. a painting within a shot). The TKK editors indicated that they would most likely prefer to work on chapter level, i.e. a scene of a TKK episode in which a specific art object is discussed. The reason for this is that it is easier for them to work in increments, instead of having to describe an episode as a whole. Furthermore, the TKK staff suggests to allow selection of entities for linking on a global level to increase efficiency. For example, they would like to select a specific entity

[1] `http://virtuoso.openlinksw.com/`

[2] We use a news show programme from the German broadcaster Rundfunk Berlin Brandenburg (RBB).

type, such as *Person: Rembrandt van Rijn* or *Location: Amsterdam* and accept or reject this type, which automatically results in the hyperlinks connected to these entities either being accepted or rejected for the entire chapter.

Obviously, the TKK staff are very interested in the expected precision of the link suggestion process, and whether this would help to provide better links using less human resources. They suggest that it could be useful to be able to control the cut-off point of a ranked list of links as a function of the amount of supervision that can be provided. In general, editorial staff would have about 20-30 minutes to work on this editor tool, but this can differ per episode. For example, when there is only limited time for supervision, the amount of suggested links should be small (low recall is fine) and have a high probability of being relevant (precision is important). When there is more time to supervise the suggested links it would be interesting to explore the link suggestions more abundantly for a bit more icing on the cake.

3.1 On the Definition of Relevance

The definition of precision and relevance is in this context a crucial issue. Although the concept of hyperlinking in general –based on people's experience with linking in webpages on the World Wide Web– is well-known, the unfamiliarity with the concept of video hyperlinking (see also [9]) complicates the discussion. During the interview session with the TKK staff, different perspectives on link relevance could be distinguished. For example, relevance can be regarded from a television production point-of-view where links should add to the intended "message" of the television programme. Clearly, producers have a detailed view on how they should reach specific audiences and providing video hyperlinks can be regarded as just an extension of the original medium. This perspective can have many gradations, from really having the creator in the loop to a broadcast company that demands control over their perceived identity.

Heading more towards the user playing a role in the definition of relevance, it was suggested that links could be typified as being relevant for a more general audience and relevant for a specific (type of) user. Interestingly, this typology may align well with the amount of supervision that is required. Obviously, it is less difficult to define guidelines on what an audience in general may be interested in with respect to links which would allow to focus supervision on a limited set of highly relevant "general purpose" links. Moving towards more specific (types of) users, supervision could be reduced to global levels (e.g., by defining acceptable links on the basis of user/audience types) or even omitted and left to application-side personalisation approaches.

Next to general and user-specific link relevance, also programme specific relevance was mentioned. For an antiques roadshow type of programme for example, context information on the time-period in which an art object was made is regarded as being very relevant. Note that looking at the programme level for assessing the relevance of a link is different from a content-based assessment as the former typically applies to all the programmes of this type whereas the latter applies to a specific programme item. On another level related to link relevance is

the difference between how links are perceived while watching a live programme or when watching it using an on-demand service. However, one can argue that the boundaries between live broadcast and on-demand are fading.

3.2 Link Targets

The choice for a set of link targets or link target domains, defined beforehand as being useful for linking in the context of a certain programming, plays a significant role in the process. When the set is small and confined in terms of entities that serve as anchors for linking (e.g., person names like "Rembrandt van Rijn" with biographic information available in a white-listed source like Wikipedia), the precision of automatic link generation can be expected to be higher than would be the case when there are several link target domains that allow different entity types for establishing links, as the limitations can serve as a restriction on the variability in the processes of anchor selection (e.g., only focus on automatic person detection) and link target suggestion (e.g., disambiguation needs to be done only at person level).

Hence, both from a technical point of view and from the viewpoint of a broadcaster, the definition of a "white-list" of one or more link target domains is an important step. In the discussions with the TKK staff the possibility to use self-curated content sets, such as a photo database of art objects taken on the location where the antiques roadshow programme was recorded, was also addressed. Another suggestion was to use the frequency of a single link suggested within a certain time frame as a way to collect evidence on the probability that a certain link is relevant. Finally, from the interview session some ideas on addressing the user were made such as the possibility to bookmark links and/or media fragments, and to have the user in the feedback loop for assessing relevance and/or rank links (e.g., by collecting thumbs-up/thumbs-down information, or use implicit feedback via the clicks). Note however, that the latter would be less feasible for video hyperlinking in live programmes. Using different colours for different link types was another suggestion.

4 Future Work

Using the link extraction workflow as described in section 2 and the requirements provided by the broadcaster's editorial team, the *VideoHypE* tool will be built in close collaboration with the TKK staff. A working version of the tool will be presented to them, and the tool will be formally evaluated with TKK editorial members. The output of the tool –the links suggested by the link extraction workflow and validated by the TKK editors using the tool– will be evaluated with end-users watching the antiques roadshow programme in an interactive application featured with video hyperlinks. The evaluation with editors will provide us with more insight into a supervised approach with respect to video hyperlink generation and, in its slipstream, with the performance levels of automatic link extraction. The evaluation with end-users will give us the opportunity to develop

a better understanding about user preferences, personalisation and appreciation of video hyperlinking.

Acknowledgements. The work reported here is done within the scope of the LinkedTV project (FP7-287911), which is funded by the European Commission through the 7th Framework Programme. We thank the Dutch public broadcaster AVRO for the re-use of the Tussen Kunst & Kitsch content and their valuable input.

References

1. Baltussen, L.B., Leyssen, M.H.R., Ossenbruggen, J., van, O.J., Blom, J., Leeuwen, P., van, H.L.: Antiques Interactive. In: EuroITV 2012 – Adjunct Proceedings, pp. 23–24. Fraunhofer Institute for Open Communication Systems, FOKUS, Berlin (2012)
2. Stein, D., Apostolidis, E., Mezaris, V., Patz, N., Müller, J.: Semi-Automatic Video Analysis for Linking Television to the Web. In: EuroITV 2012 – Adjunct Proceedings: FutureTV Workshop, pp. 154–161. Fraunhofer Institute for Open Communication Systems, FOKUS, Berlin (2012)
3. Romero, L.P., Traub, M.C., Leyssen, H.R., Hardman, L.: Second Screen interactions for Automatically Web-enriched Broadcast video. In: Submitted to: ACM SIGCHI Conference on Human Factors in Computing Systems (CHI 2013) "Exploring And Enhancing the User Experience for Television" Workshop, Paris, France (2013)
4. Reynolds, D.A., Quatieri, T.F., Dunn, R.B.: Speaker Verification Using Adapted Gaussian Mixture Models. Digital Signal Processing 10, 19–41 (2000)
5. Schneider, D., Schon, J., Eickeler, S.: Towards Large Scale Vocabulary Independent Spoken Term Detection: Advances in the Fraunhofer IAIS Audiomining System. In: Proc. SIGIR, Singapore, pp. 34–41. CTIT, Enschede (2008)
6. Tsamoura, E., Mezaris, V., Kompatsiaris, I.: Gradual transition detection using color coherence and other criteria in a video shot meta-segmentation framework. In: 15th IEEE International Conference on Image Processing, ICIP 2008, pp. 45–48 (2008)
7. Lienhart, R., Maydt, J.: An extended set of Haar-like features for rapid object detection. In: International Conference on Image Processing Proceedings, vol. 1, pp. I-900–I-903 (2002)
8. Troncy, R., Leeuwen, P., van, G.J.: Specification of the Media Fragment URI scheme, D2.1, LinkedTV (2012), `http://www.linkedtv.eu/wp/wp-content/uploads/2012/11/LinkedTV_D2.1_Specification_of_the_Media_Fragment_URI_scheme.pdf`
9. Aly, R., McGuinness, K., Kleppe, M., Ordelman, R., O'Connor, N.E., de Jong, F.: Link Anchors in Images: Is there Truth? In: Proceedings of the 12th Dutch Belgian Information Retrieval Workshop (DIR 2012), pp. 1–4 (2012)
10. Lašek, I., Kliegr, T., Dojchinovski, M., Sahuguet, M., Rizzo, G., Huet, B., Troncy, R.: Specification of Web Mining Process for Hypervideo Concept Identification (2012)

Interactive TV Potpourris: An Overview of Designing Multi-screen TV Installations for Home Entertainment

Radu-Daniel Vatavu[1] and Matei Mancaş[2]

[1] University Ştefan cel Mare of Suceava, Romania
[2] University of Mons, Belgium
vatavu@eed.usv.ro, matei.mancas@umons.ac.be

Abstract. Home entertainment systems comprising multiple TV screens offer new opportunities to display more content, accommodate more viewers, and deliver enriched user experiences. In many cases, such installations take the form of mixed-reality environments, in which video projections coexist with physical TV sets. We refer to such installations as *interactive TV potpourris*, due to their composite nature of hybridizing individual TV screens of different natures, form factors, and potential to render different multimedia types. This work discusses current implementations for interactive TV potpourris, identifies technical and interaction challenges, and pinpoints future research and development directions. It is our hope that this work will encourage new explorations and developments of TV potpourris.

Keywords: interactive TV, TV potpourris, interaction techniques, multiple displays, home entertainment.

1 Introduction

Users' increasing demands for sophisticated TV technology in terms of more content [4], integration with personal devices [6], and simple and effective control of the TV set (e.g., gestural interfaces [14]), have recently started to be accommodated by the TV industry manufacturers. However, fully matching users' expectations with today's single-screen TV sets represents a considerable challenge. For instance, although web browsing is supported by the software platform of all Smart TVs, it cannot take place at the same time with TV channel watching because of the limited real-estate of the physical TV screen. Consequently, users frequently resort to employing second-screen devices [4,6] to compensate the shortcomings of the main TV screen.

We address in this work collections of TV screens (i.e., more than two screens that are co-located on the same living room wall) that are part of a single, unified home entertainment system. We refer to such systems as *interactive TV potpourris*, due to their composite nature of hybridizing individual screens of different form factors (i.e., display size and aspect ratio, such as 16:9 or 4:3), that display different content, occupy distinct locations in space, and present different interfaces (e.g., remote controls or gestural interfaces). A potpourri of interactive TVs (iTVs) is therefore

M. Mancas et al. (Eds.): INTETAIN 2013, LNICST 124, pp. 49–54, 2013.

composed of more than two screens that hybridize into the same unified entertainment system in the attempt to provide more content and functionality beyond traditional single and two-screen TVs. In this paper we describe existing technical solutions that implement such systems, summarize interaction challenges for interactive TV potpourris, and point to interesting questions regarding the human capacity to manage the visual information flow that iTV potpourris are able to deliver. We hope this paper will contribute towards a better understanding of today's technical and interaction challenges for these installations and, consequently, encourage the community to further investigate the opportunities delivered by such systems.

1.1 Technical Implementation of Interactive TV Potpourris

A simple way to explore scenarios with multiple TV screens, without recurring to complex hardware installations involving physical TV sets, is to employ video projections [12,13]. The result resembles the output of augmented reality systems that project digital content onto the real environment in order to deliver enriched user experiences for various application scenarios [11]. Previous research outside the interactive TV community has already explored such installations. For example, Cotting and Gross [5] experimented with displays adaptable in form and size projected on the surfaces of tabletops; Vermeulen et al. [15] enhanced the walls of a room containing technical equipment with visual feedback designed to help users understand the technical intricacies of the room and guide them into using the various technical facilities of the room; and Wilson et al. [16] introduced the Beamatrom, a steerable video projector that can display images at any location inside a room. The Ambilight lighting effects installation for Philips TVs[1], the NDS Surfaces concept [9], and the IllumiRoom prototype of Jones et al. [8] for Microsoft XBox are all part of the same attempt to enrich the user experience with elaborated visual effects.

The interactive TV wall system of Vatavu [12] was the first installation to combine multiple video projected TV screens, independently controlled in terms of location, size, and the content they deliver (Figure 1a). The goal of the TV wall was to go beyond current limitations of the physical TV set, such as its post-purchase inability to be customized in terms of desired size and limited possibility to install it at different locations in the living room. In order to facilitate users' adoption of a complex TV environment comprising many controllable TVs, the authors reused the viewers' existing expertise of employing point & click interaction metaphors on windows operating systems. The AROUND-TV system [13] went further and explored interactive TV potpourris that combine both video-projected and physical TV screens in a hybrid, digital-physical augmented TV space (Figure 1b). Viewers are still able to control each individual screen with point & click techniques, but more sophisticated interactions are also possible, such as "dragging" TV content running on the physical TV set outside its bounding area, or employing widgets projected on the wall, next to the screen, in order to control the TV transmission (e.g., "go to next channel" or "increase volume").

[1] http://www.research.philips.com/technologies/ambilight/index.html

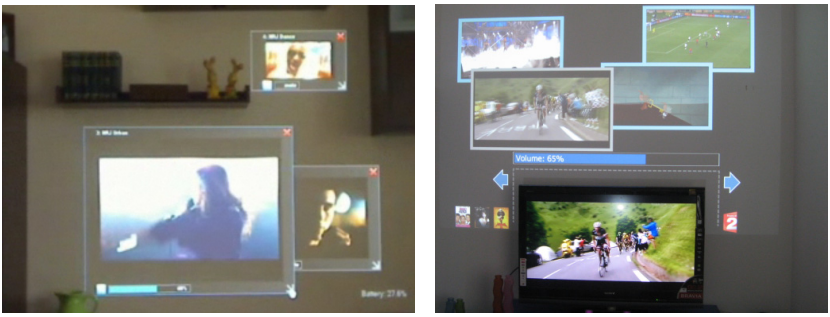


Fig. 1. Two interactive TV potpourris installations: (a) the TV wall [12] with multiple TV screens video projected on the living room wall and (b) the AROUND-TV system [13] in which projected screens co-exist with the physical TV set

Technical challenges reported by these works are high processing power demands needed to display multiple video streams, accommodating video projectors to work at resolutions that enable full-HD quality for the projected TV screens, and video projecting on walls and furniture of different colors (e.g., dark colors). In order to guide further development of such installations, Vatavu [13] provided a series of design principles for augmented-TV spaces that include: (1) context-dependent augmentation of home entertainment, (2) multi-tasking home entertainment spaces, (3) sophisticated control, (4) natural forms of interaction and gradual transition to new input modalities, (5) seamless integration with personal devices, and (6) scalability to more viewers. We can further add to these challenges the need to design and develop new infrastructures and software architectures for TV potpourris to accommodate the complexity of delivered content from multiple sources. In this direction, Falchuk et al. [7] discussed the motivation, design, and uses of high experience coalition-based services and described how such services fit an architecture for co-located devices.

2 Interacting with TV Potpourris

Interactive TV potpourris need appropriate interfaces that would go beyond the interaction possibilities offered by standard remote controls. Indeed, the research community has many times warned against the complexity and ineffectiveness of such control devices in many cases (such as in the living room ethnographic study of Bernhaupt et al. [3]). Consequently, hybrid solutions combing remote controls and gestures have started to be proposed and investigated in terms of performance [2].

Vatavu [14] addressed the problem of interacting with collections of TV screens and conducted the first comparative study on two of today's emerging gestural interfaces for interactive TVs: (1) hand-held remote devices with embedded motion-sensing features and (2) free-hand gestures captured by ambient video cameras. (The TV industry currently provides solutions for both scenarios, such as the LG's Magic

Remote[2] or Samsung's free-hand gesture control system[3].) The study employed a participatory design methodology [17], in which participants were asked to suggest commands for a large set of twenty-two frequent TV control tasks. Commands were proposed in terms of buttons, motion gestures, and combinations of buttons and gestures. The study reports several findings that can be employed to guide the design of gestural interfaces for collections of TV screens, such as: (a) recommendations to reuse point & click and drag & drop interaction metaphors when working with graphical items; (b) recommendations to prefer buttons to motion gestures for cases in which buttons are available and intuitive, without agglomerating the design of the remote control; and (c) empirical findings that recommend to explore culture-specific and full body gestures. We rely on these findings to address in the following a current debate on the use of gesture commands for iTVs, applicable to iTV potpourris as well: remote controls vs. free-hand gestures.

Buttons vs. Motion Commands. One of the findings of [14] was that people prefer button commands when they can choose between using a button and performing a motion gesture. The reason is that buttons are simpler to use, require minimum effort, and are easier to remember than motion gestures. This finding was supported by a large majority of participants' suggestions, with 76% recommending buttons and 65% using buttons exclusively. Vatavu reports that the first intention of the participants when suggesting a command was to think of a button with an intuitive meaning that could be used to execute that command. However, participants agreed that not too many buttons should exist (with a total of 15 buttons resulting from the experiment). Buttons were preferred when the task to perform had an abstract nature (e.g., Mute), for which they were not able to associate a suitable gesture. Instead, gesture commands were preferred in about 20% cases. These findings show that people are familiar with remote controls and they tend to prefer them in 4 out of 5 cases, but motion gestures are still perceived as useful (e.g., when left/right or up/down motions can be intuitively mapped to commands such as "go to next channel"). Also, these findings suggest that TV remotes implementing a combination of buttons and motion commands represent a good compromise for today's TV viewers.

Technical vs. Non-technical People. Vatavu [14] reports that people with non-technical backgrounds preferred button commands, while technical people were able to reuse interaction metaphors from desktop computing including point & click and mouse drawing (e.g., drawing a rectangle to create a new TV screen and drawing the question mark symbol to invoke "help"). Obviously, non-technical people will face many challenges with a TV interface relying on motion gestures exclusively. This finding supports our previous recommendation of a hybrid design of the TV remote, including buttons and motion input. We expect that such a hybrid design will allow a smooth transition to gestural interfaces for different population groups.

[2] http://www.lg.com/global/magicremote/
[3] http://www.samsung.com/us/2013-smart-tv

Human Factors and Visual Attention. Although collections of interactive TVs can bring many benefits in terms of control and content, we expect some limitations to occur in terms of viewers' capacity to manage the increased visual complexity of the information being delivered. Attention is the capacity to selectively process information subsets instead of taking into account at each moment the entire amount of information that is available [1]. By definition, attention leads people to focus on a single task at a time. However, attention can be divided between several information sources, especially when the senses associated to the different sources are uncorrelated, e.g., when one drives and talks to the phone at the same time. This divided attention of uncorrelated senses works well, even if individual performances of each sense are decreased in comparison to sustained attention on a single task [10]. In the case of interactive TV potpourris, the same senses (vision and hearing) are used for all screens: while ears focus on the audio signal of the main TV transmission, eyes are likely to be drawn by other content displayed on different screens. We can therefore suspect that a multi-screen TV environment will increase the cognitive load of the viewer in such a way that, when improperly designed, the TV experience might not always be perceived as relaxing. These hypotheses could be true if the screens arrangement does not follow a proper structure (i.e., all screens having the same size, without a clearly identified main TV screen) or if all screens display the same type of information. However, studies are needed in order to fully understand the human capacity to adapt to the complex visual information that multi-screen TVs provide.

3 Conclusion

We described in this work existing solutions for implementing interactive TV potpourris, for which we identified technical and interaction challenges, reported preliminary experimental findings, and pointed to some design recommendations. TV potpourris are made available by the advances in interaction and visualization technologies; however, the balance between the number of screens, what content to display, and how to design interaction techniques are all crucial for their adoption. We hope this work will benefit the iTV community and will encourage the practitioners of interactive TV to further develop on top of existing potpourris scenarios in order to deliver new interactive TV applications and enriched user experiences.

References

1. Anderson, J.R.: Cognitive psychology and its implications, 6th edn., p. 519. Worth Publishers (2004)
2. Bailly, G., Vo, D.B., Lecolinet, E., Guiard, Y.: Gesture-aware remote controls: Guidelines and interaction technique. In: Proceedings of the 13th International Conference on Multimodal Interfaces (ICMI 2011), pp. 263–270. ACM, New York (2011)
3. Bernhaupt, R., Obrist, M., Weiss, A., Beck, E., Tscheligi, M.: Trends in the living room and beyond: results from ethnographic studies using creative and playful probing. Comput. Entertain. 6(1), Article 5, 23 pages (2008)

4. Cesar, P., Bulterman, D.C., Jansen, A.J.: Usages of the Secondary Screen in an Interactive Television Environment: Control, Enrich, Share, and Transfer Television Content. In: Tscheligi, M., Obrist, M., Lugmayr, A. (eds.) EuroITV 2008. LNCS, vol. 5066, pp. 168–177. Springer, Heidelberg (2008)
5. Cotting, D., Gross, M.: Interactive environment-aware display bubbles. In: Proceedings of the 19th Annual ACM Symposium on User Interface Software and Technology (UIST 2006), pp. 245–254. ACM, New York (2006)
6. Courtois, C., D'heer, E.: Second screen applications and tablet users: Constellation, awareness, experience, and interest. In: Proceedings of the 10th European Conference on Interactive TV and Video (EuroITV 2012), pp. 153–156. ACM, New York (2012)
7. Falchuk, B., Zernicki, T., Koziuk, M.: Towards streamed services for co-located collaborative groups. In: Proc. of the 8th Int. Conf. on Collaborative Computing: Networking, Applications and Worksharing, pp. 306–315 (2012)
8. Jones, B., Benko, H., Ofek, E., Wilson, A.D.: IllumiRoom: Peripheral Projected Illusions for Interactive Experiences. In: Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI 2013). ACM Press, New York (2013)
9. Parnall, S.: NDS Surfaces. In: 10th European Interactive TV Conference (EuroITV 2012). iTV in Industry (2012)
10. Strayer, D.L., Drews, F.A., Johnston, W.A.: Cell phone-induced failures of visual attention during simulated driving. Journal of Experimental Psychology Applied 9(1), 23–32 (2003)
11. Thomas, B.H.: A survey of visual, mixed, and augmented reality gaming. Comput. Entertain. 10(3), Article 3, 33 pages (2012)
12. Vatavu, R.D.: Point & Click Mediated Interactions for Large Home Entertainment Displays. Multimedia Tools and Applications 59(1), 113–128 (2012)
13. Vatavu, R.D.: There's a World outside Your TV: Exploring Interactions beyond the Physical TV Screen. In: 11th European Interactive TV Conference. ACM Press, New York (2013)
14. Vatavu, R.D.: A Comparative Study of User-Defined Handheld vs. Freehand Gestures for Home Entertainment Environments. Journal of Ambient Intelligence and Smart Environments 5(2), 187–211 (2013)
15. Vermeulen, J., Slenders, J., Luyten, K., Coninx, K.: I Bet You Look Good on the Wall: Making the Invisible Computer Visible. In: Tscheligi, M., de Ruyter, B., Markopoulus, P., Wichert, R., Mirlacher, T., Meschterjakov, A., Reitberger, W. (eds.) AmI 2009. LNCS, vol. 5859, pp. 196–205. Springer, Heidelberg (2009)
16. Wilson, A., Benko, H., Izadi, S., Hilliges, O.: Steerable augmented reality with the beamatron. In: Proceedings of the 25th Annual ACM Symposium on User Interface Software and Technology (UIST 2012), pp. 413–422. ACM, New York (2012)
17. Wobbrock, J.O., Morris, M.R., Wilson, A.D.: User-defined gestures for surface computing. In: Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI 2009), pp. 1083–1092. ACM, New York (2009)

3D Head Pose Estimation for TV Setups

Julien Leroy, Francois Rocca, Matei Mancaş, and Bernard Gosselin

University of Mons (UMONS), Faculty of Engineering (FPMs),
20, Place du Parc, 7000 Mons, Belgium

Abstract. In this paper, we present an architecture of a system which aims to personalize the TV content to the viewer reactions. The focus of the paper is on a subset of this system which identifies moments of attentive focus in a non-invasive and continuous way. The attentive focus is used to dynamically improve the user profile by detecting which displayed media or links have drawn the user attention. Our method is based on the detection and estimation of face pose in 3D using a consumer depth camera. Two preliminary experiments were carried out to test the method and to show its link to viewer interest. This study is realized in the scenario of a TV with a second screen interaction (tablet, smartphone), a behaviour that has become common for spectators.

Keywords: attention, head pose estimation, second screen interaction, eye tracking, Facelab, future TV, personalization.

1 Introduction

One of the goals of future TV is to offer new possibilities for personalization of content provided to users, including the implicit analysis of human behaviour.

To achieve the personalization goal several factors need to be taken into account: explicit interactions (pause, play, skip, click on a link, etc.), implicit interactions (looking to the TV or not) and context information (date, time, social networks, number of viewers, etc.). In this paper, we focus on implicit interaction and more specifically on a solution of head detection and pose estimation using a low-cost depth camera. This choice was made due to the democratization of this type of sensors and their arrival in the home through gaming platforms [17]. Moreover, TV manufacturers begin to integrate cameras into their new systems, regarding the sensors we can see the willingness of the makers to miniaturize sensors such as PrimeSense new camera "Capri" [21]. Thus, we can expect to see in the coming years 3D sensors directly integrated into televisions.

The next section provides information about the related work, section 3 details the implemented algorithm and two experiments. Section 4 relates the first results of the first experiment, while section 5 focuses on the second experiment. Section 6 provides some cues about the analysis of the results for media personalization and it is followed by the conclusion section.

M. Mancas et al. (Eds.): INTETAIN 2013, LNICST 124, pp. 55–64, 2013.

2 Related Work

Movement and orientation of the head are important non-verbal cues that can convey rich information about a person's behaviour and attention [24][12]. Until recently, the literature has mainly focused on the automatic estimation of the poses based on standard images or videos. One of the major issues that must be addressed to obtain a good estimator is to be invariant to variables such as: camera distortions, illumination, face shape and expressions or features (glasses, beard). Many techniques have been developed over the years such as appearance template methods, detector array methods, non linear array methods, manifold regression methods, flexible methods, geometric method, tracking method and hybrid methods. More information on these methods can be found in [18]. More recently, with the arrival of low cost depth sensor, more accurate solutions have emerged [6][8]. Based on the use of depth maps, those methods are able to overcome known problems on 2D images as illumination or low contrast backgrounds. In addition, they greatly simplify the spatial positioning of the head with a global coordinate system directly related to the metric of the analysed scene. Many of these techniques are based on a head tracking method which unfortunately often requires initialization and also undergoes a drift. Another approach, based on the frame to frame analysis as the method developed by [9], provides robust and impressive results. This method is well suited for a living room and TV scenario. It is robust to illumination conditions that can be very variable in this case (dim light, television only source of light, etc.) but is based on a 3D sensor like the Microsoft Kinect. The paper proposes a entire system of optimized head pose extraction.

3 Head Pose Estimation

3.1 Algorithm

The proposed system is based on the head detection and pose estimation on a depth map. Our goal is to achieve head tracking in real time and estimate the six degrees of freedom (6DOF) of the detected head (spatial coordinates, pitch, yaw and roll). The advantage of a 3D system is that it uses only geometric information on the point cloud and is independent of the illumination issues which can dramatically change in front of a device like a TV. The proposed system can even operate in the dark or in rapidly varying light conditions, which is not possible with face tracking systems working on RGB images. In addition, the use of 3D data provide more stable results than 2D data which can be mislead by projections of the 3D world on 2D images.

Figure 1 shows the global pipeline of the head pose estimation sub-system. First, the 3D point cloud is extracted from a Kinect sensor using the PCL library [20]. In a second step people face is detected and localized (the blue larger boxes in Figure 1). Those boxes are computed from the head of the skeleton extracted from the depth maps by using the OpenNI library [19]. The skeleton head provides the 3D coordinates of the area where a face might be located.

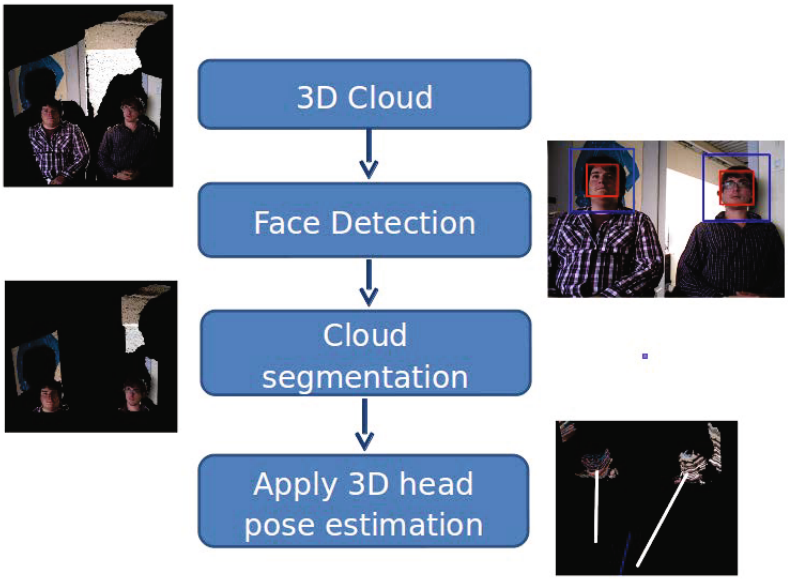


Fig. 1. Algorithm pipeline: 3D cloud extraction from the RGBD sensor, face localization and detection, 3D cloud segmentation and pose estimation

The smaller red boxes are 2D face detection which can be used for face analysis, but this issue is not in the focus of this paper. Once the 3D position of the head is extracted, the 3D cloud is segmented to optimize the last 3D head pose estimation step. The segmentation eliminates a lot of the points of the 3D clouds where the chances to find a face are very low and therefore boosts the computational efficiency of the method.

The 3D head pose estimation used here follows the development in [16] which is improved by 4 in terms of computation time due to the 3D point cloud segmentation. The 3D pose estimation algorithm is based on the approach in [7][10] and implemented in the PCL library [2]. This solution relies on the use of a random forest [3] extended by a regression step. This allows us to detect faces and their orientations on the depth map. The method consists of a training stage during which we build the random forest and an on-line detection stage where the patches extracted from the current frame are classified using the trained forest. The training process is done only once and it is not user-dependent. One initial training is enough to handle multiple users without any additional configuration or re-training. This is convenient in a setup where a wide variety of people can watch TV. The training stage is based on the BIWI dataset [10] containing over 15000 images of 20 people (6 females and 14 males). This dataset covers a large set of head poses (± 75 degrees yaw and ± 60 degrees pitch) and generalizes the detection step.

During the test step, a leaf of the trees composing the forest stores the ratio of face patches that arrived to it during training as well as two multi-variate Gaussian distributions voting for the location and orientation of the head. This step of the algorithm provides the head position and a rough head orientation on any new individual without the need of re-training. We then apply a final processing step which consists in registering a generic face cloud over the region corresponding to the estimated position of the head. This last step greatly stabilizes the final head position result.

3.2 Experiment

Our experimental setting consists of:

- a 46 inch HD TV,
- a sofa, located at 2.5m from the TV,
- a 3D camera positioned at 80 cm from the sofa and low enough to not obstruct the field of vision of the viewer,
- a 10 inches tablet that plays the role of a second screen.

These parameters allow us to calibrate our tracking system and reconstruct a simplified virtual 3D scene (Figure 4). The Kinect is located between the viewer and the TV which is not very convenient and it can be subject to viewer face occlusion when using second screen devices. Therefore the final setup will use the second generation Kinect which has a better resolution and should be capable to capture head motion when ideally located on top of the TV.

Within this setup, we performed two scenarios. The first one consists in detecting the head direction of a person watching TV in his living room. The idea is to discriminate between 1) watching TV, 2) watching the second screen (tablet, smartphone), 3) watching outside the TV, 4) watching out of the TV setup (no face detection but viewer detection). We asked participants to solve various puzzles on a tablet with increasing difficulty to keep them focused on the second screen like on the Fig. 2. The broadcast media is a zapping, a series of short clips of news, sports, politics, buzz, etc. In addition to this test, we also performed a second scenario. We used a commercial eye-tracking (Facelab 5 [23]) system which is able to measure both head direction and eye gaze direction. The eye-tracker was located at 1.80m from the TV screen and the viewer at 2.30m from the same TV as in experiment 1. The purpose of this second test was both to asses the 3D camera-based head detection, and also to have a first idea about the relationship between the head direction and eye direction.

4 Results of the First Scenario: Head from 3D Camera

Each frame can be processed up to 8 frames/sec on a Macbook Pro with an Intel Core i5 2.53GHz. This speed is enough to extract head direction and basic features like direction change and speed. In addition, the algorithm proposed here also works on a recorded 3D video (.oni format). In this case the processing

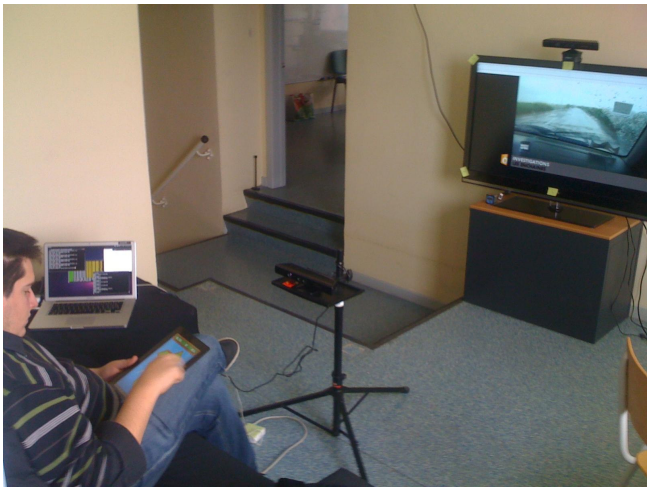


Fig. 2. Setup of the experiment with the user playing a puzzle game on the second screen (tablet) while a TV show is displayed on the main screen. The camera in the middle of the scene tracks head movements.

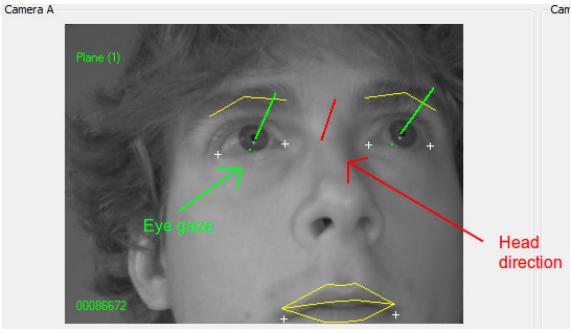


Fig. 3. Second experiment: Facelab interface showing head direction as a red vector between the eyes and eye gaze as two green vector located on the eyes

speed can be the same as the framerate (30 fps). Within the TV viewer profile personalization application, the use of pre-recorded video is possible as the head pose data is only sent when the context changes (viewers enter/leave, etc.) as explained in section 6.

To detect if a user watches TV or not, we reconstruct a virtual simplified model of the real scene (Figure 4). Therefore, knowing the 6DOF position of the face of the person detected, the camera position and the TV position it is possible to estimate the point of intersection between the TV and the orientation of the head. In this way, we can synchronize annotated media with the head tracker and estimate ($\pm$10 cm, on our 46" TV) where the user is looking.

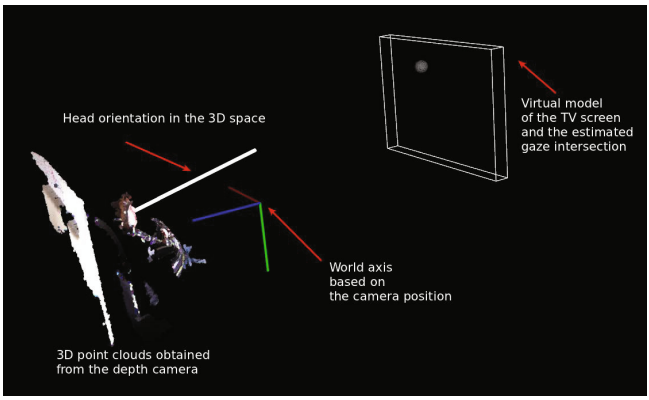


Fig. 4. 3D rendering of our system. On left: 3D point cloud from depth camera and head direction vector. On right: 3D model of the TV and intersection point between TV and viewer head position.

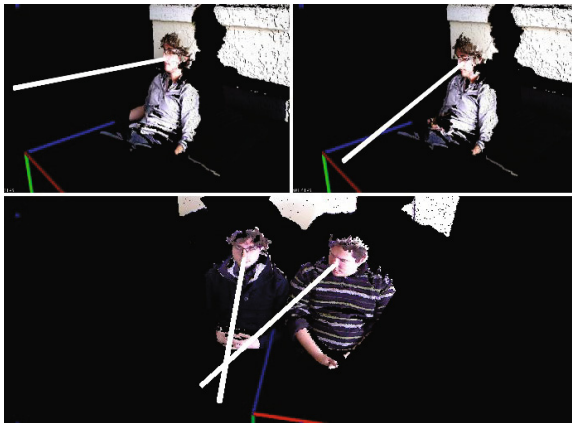


Fig. 5. Top-images: switch between main screen (left) and second screen (right). Bottom-image: Multiple head detection and orientation estimation.

Depending on the camera position, user head direction can be detected towards the main screen or the second screen (tablet) as in Figure 5, top images. To be able to achieve this measure, the 3D camera must see the viewer face 1) with no occlusions due to the second screen, 2) with a pitch angle which is small enough for the algorithm (± 60 degrees pitch).

Moreover, the algorithm can detect several users (as many as possibly detected in the camera field of view) and compute all users head directions as in Figure 5, bottom image. This feature allows us to check potential joint attention on the main TV screen.

5 Results of the Second Scenario: Attention from Head

The easier way to measure people overt [25] attention is to measure eye gaze or direction. Given the technical limitations of camera distance, it is not possible to access the viewer's eyes orientation. We than hypothesise that, at the TV setup distance (more than two meters from the main screen), the gaze of a person is considered to be close to the direction of his head. As stated in [18], "[...]*Head pose estimation is intrinsically linked with visual gaze estimation ... By itself, head pose provides a coarse indication of gaze that can be estimated in situations when the eyes of a person are not visible*[...]". Several studies rely and validate this hypothesis as shown in [1].

In the second experiment we firstly qualitatively compared the Facelab head direction detection with the proposed algorithm. The results are similar, and the proposed approach seems even to be more reactive to head movements, while the one of Facelab needs large head movements. However, a quantitative comparison is not simple due to the framerate difference between the two systems.

In a second step we compared the head direction with the eye gaze using again the Facelab system (Figure 6). The first results we obtained are consistent with the literature and show that there is a correlation between eye gaze and head direction. This correlation is higher when the gaze goes far from the image centre and for more dynamical content (fast moving videos). The head direction does not exactly follow the eye gaze which is much faster to attend events occurring on the TV screen, but the head direction accompany the gaze in a smoother way. The head and eye movements work together to both minimize their motion (effort) and maximizing the acquisition of interesting information in the scene. In this optimisation process, the head mechanics naturally act like a smoother while eye reactions can be much faster.

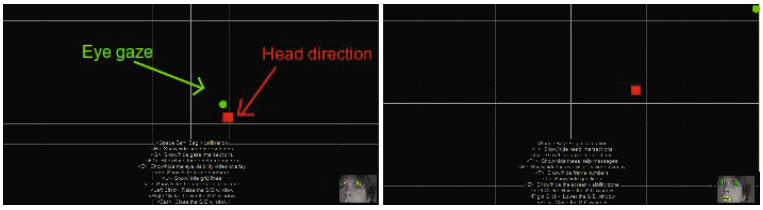


Fig. 6. Green circle: eye gaze, Red square: head direction. Left image: both are very close, Right image: the eye quickly shifted to the top-right corner and the head position followed in the same direction.

6 From Attention to Content Personalization

Based on our preliminary studies, we can say that head direction might be a rough approximation of eye gaze, thus of overt attention. Attention is a phenomenon based on two competing precesses: the top-down and bottom-up

attention [22]. Bottom-up attention is a generic approach also known as stimulus-driven or exogenous attention. Furthermore, it relies on the information innovation that the features extracted from the image can bring in a given spatial context. The top-down component of attention, which is also known as task-driven or endogenous attention, integrates specific knowledge that the viewer could have in specific situations (tasks, models of the kind of scene, recognized objects, etc.).

While bottom-up attention will be engaged each time that surprising images/motion or sounds arise, top-down attention shows that the viewer is specifically interested to the content displayed. Thus, for content personalization, the most important is to extract the viewer interest in terms of top-down attention.

To detect the top-down attention of the viewer, several features can be extracted.

- The classical case is that the user's attention is drawn from the second screen and stays focused on the main screen for a long time. This is a sign of sustained attention which shows that cognitive processes are engaged and it is not only a bottom-up attention due to surprising events [13]. A first feature is thus the time spent looking at the screen after a head position change.
- [15][4][5] show that it is possible, by classifying head trajectories based on their speed and amplitude, to distinguish attention switch due to bottom-up stimuli and those due to top-down information. The speed of the head direction change and the total angle of the head motion are additional features of the kind of attention which the viewer uses.
- In case of several users, joint attention (see Figure 5, bottom image) which is stable during enough time shows a discussion or a common subject of interest. Joint attention is an additional cue for top-down attention.

Based on the detection of focus on the main screen and the nature of the attention attracted by the media (sustained, bottom-up), it is possible to provide, for each media segment a weight of interest that the user implicitly expressed. This weight, mixed along with other contextual cues (time and date, number of viewers, the presence of children or not, etc.) and the explicit actions of the viewer (skip, play, stop, explore links, etc.) provides a good idea about the subsets of the media which are interesting for the viewer. This information let an ontology-based system propose to the viewer media which are close to the viewer interests depending on the context (viewing alone during the WE will most of the time be different from viewing in family during week days).

The data collected through the system is sent to a content personalization framework. At each change in context (new viewers entering, viewers leaving, kids, coming or leaving, etc.) the logs of the head focus for the viewers is sent (1: focus on the main screen, 2: focus on the second screen, 3: focus out of screens, 4: no head detected (the viewer is talking to another one or looking back ...)). In addition to the head focus, for the first two modes, the kind of attention (bottom-up or top-down) is also sent. These logs will be used to modify the

viewer profile in the given context. Some feature combinations will provide cues about a positive interest of the viewer (look to the main screen - mode 1 and top-down attention, look to second screen - 2 and top-down attention), others about a negative interest (look to the walls - mode 3) and others will provide a neutral result (not enough to know about the viewer interest, keep previous score like mode 4 or modes 1 and 2 with bottom-up attention).

To summarize, the system described in this paper, at the end of each user session (context change: when a user leaves the interaction zone, when a second user comes in), the logs containing the tracking data will be sent as REST [11] query to the remote personalization module called GAIN (General Analytics INterceptor) [14] which will use rule-based learning algorithms to change viewer profile accordingly.

7 Conclusions

In this paper, we presented a system architecture and two preliminary experiments on an implicit behaviour analysis system based on a 3D head tracker. This tool is optimized compared to previous publications and it is designed to feed a personalization framework capable of processing behavioural data to dynamically enhance a user profile. The preliminary results show that it is possible to extract implicit information and that head direction can provide cues about viewer interest which can be used in future TV personalization. In the future, 1) more extensive tests will be conducted to confirm the preliminary findings of our two experiments and 2) additional information will be provided concerning the kind of attention (bottom-up or top-down) which is crucial information to asses real viewer interest.

Acknowledgments. This work is supported by the Integrated Project LinkedTV (www.linkedtv.eu) funded by the European Commission through the 7th Framework Programme (FP7-287911).

References

1. Abe, K., Makikawa, M.: Spatial setting of visual attention and its appearance in head-movement. IFMBE Proceedings 25(4), 1063–1066 (2010)
2. Aldoma, A.: 3d face detection and pose estimation in pcl (September 2012)
3. Breiman, L.: Random forests. Machine Learning 45(1), 5–32 (2001)
4. Doshi, A., Trivedi, M.M.: Head and Gaze Dynamics in Visual Attention and Context Learning, pp. 77–84 (2009)
5. Doshi, A., Trivedi, M.M.: Head and eye gaze dynamics during visual attention shifts in complex environments 12, 1–16 (2012)
6. Fanelli, G., Dantone, M., Gall, J., Fossati, A., Gool, L.: Random Forests for Real Time 3D Face Analysis. International Journal of Computer Vision 101(3), 437–458 (2012)
7. Fanelli, G., Dantone, M., Gall, J., Fossati, A., Gool, L.: Random forests for real time 3d face analysis. International Journal of Computer Vision 101, 437–458 (2013)

8. Fanelli, G., Gall, J., Van Gool, L.: Real time head pose estimation with random regression forests. In: CVPR 2011, pp. 617–624 (June 2011)
9. Fanelli, G., Gall, J., Van Gool, L.: Real time 3d head pose estimation: Recent achievements and future challenges. In: 2012 5th International Symposium on Communications Control and Signal Processing (ISCCSP), pp. 1–4 (2012)
10. Fanelli, G., Weise, T., Gall, J., Van Gool, L.: Real time head pose estimation from consumer depth cameras. In: Mester, R., Felsberg, M. (eds.) DAGM 2011. LNCS, vol. 6835, pp. 101–110. Springer, Heidelberg (2011)
11. Fielding, R.T., Taylor, R.N.: Principled design of the modern web architecture. ACM Trans. Internet Technol. 2(2), 115–150 (2002)
12. Gaschler, A., Jentzsch, S., Giuliani, M., Huth, K., de Ruiter, J., Knoll, A.: Social behavior recognition using body posture and head pose for human-robot interaction. In: 2012 IEEE/RSJ International Conference on Intelligent Robots and Systems, pp. 2128–2133 (October 2012)
13. Henderson, J.: Regarding scenes. Current Directions in Psychological Science 16, 219–222 (2007)
14. Kliegr, T., Kuchar, J.: Gain: Analysis of implicit feedback on semantically annotated content. In: WIKT 2012, pp. 75–78 (2012)
15. Khan, A.Z., Blohm, G., McPeek, R.M., Lefèvre, P.: Differential influence of attention on gaze and head movements. Journal of Neurophysiology 101(1), 198–206 (2009)
16. Leroy, J., Rocca, F., Mancas, M., Gosselin, B.: Second screen interaction: An approach to infer tv watcher's interest using 3d head pose estimation. In: Proceedings of the 22nd International Conference on World Wide Web Companion, WWW 2013 Companion, pp. 465–468. International World Wide Web Conferences Steering Committee, Republic and Canton of Geneva (2013)
17. Microsoft. Kinect sensor, `http://www.xbox.com/kinect`
18. Murphy-Chutorian, E., Trivedi, M.M.: Head pose estimation in computer vision: A survey. IEEE Transactions on Pattern Analysis and Machine Intelligence 31(4), 607–626 (2009)
19. OpenNI. Open natural interfaces, `http://www.openni.org/`
20. PCL. Point cloud library `http://pointclouds.org/`
21. PrimeSense. Capri sensor, `http://www.primesense.com/news/primesense-unveils-capri`
22. Riche, N., Mancas, M., Duvinage, M., Gosselin, B., Dutoit, T.: Rare2012: A multi-scale rarity-based saliency detection with its comparative statistical analysis. In: Signal Processing: Image Communication (2013)
23. Seeingmachine. Facelab5, `http://www.seeingmachines.com/product/facelab/`
24. Vinciarelli, A., Pantic, M., Bourlard, H.: Social signal processing: Survey of an emerging domain. Image and Vision Computing 27(12), 1743–1759 (2009)
25. Wright, R.D., Ward, L.M.: Orienting of attention. Oxford University Press (2008)

Visualizing Rembrandt
An Artist's Data Visualization

Tamara Pinos Cisneros and Andrés Pardo Rodríguez

LIACS, Leiden University,
Niels Bohrweg 1, Leiden, The Netherlands
{tvpinos,apardor}@gmail.com

Abstract. Visualizing Rembrandt is a web application that helps users to view connections between Rembrandt and other artists with whom he had a professional relationship. These connections can be made by choosing from different criteria: teachers, pupils, influenced by, influenced on, human figure, landscape, drawing and paintings. The data for this project was provided by the RKD (Rijksbureau voor Kunsthistorische Documentatie) and an application built with Java and Javascript was used for its display. This application is an innovative tool that is helpful to display museums data in an efficient fashion, which can be a good support for visualizing and connecting data in museums and exhibitions (and can be used with different artists data).

Keywords: Rembrandt, data visualization, art, museum.

1 Introduction

Over the past few years museums professionals have strived to provide an interactive experience to visitors. This interest has been driven by the possibilities that new technologies can offer within an exhibition place, in order to enhance and expand the engagement and learnability that general public can gain. Interaction has become an important field within museums, no matter whether it is an art, science, natural or historical museum. But, what is this interaction all about? One of the most clear definition that can be find in this field is the one given by Joshua Noble in his book Programming Interactivity. He explains that interaction is an exchange of information between two or more active participants (Noble, 2012), where one party of this can be a computer and the other a human (the user).

Discussions regarding this exchange of information within museums must take into account the debates around them and their social role (Bell, 2002). These issues are addressed because they help to acknowledge the way museums share information, their intentions of doing so, and the target audience they seek to reach. The previous concerns mold the way technology can fit into an exhibition space, as new media and interaction can shift how a museum displays information in order to make it more accessible and appealing. Technology can then provide a breakthrough for paradigms in museums, as for instance the way data

M. Mancas et al. (Eds.): INTETAIN 2013, LNICST 124, pp. 65–70, 2013.

is displayed. Spectators usually read the information concerning an artwork or an object through a fact sheet hanged in the wall. The visitor has no interaction and does not receive any kind of feedback from it, hence there is no exchange of information.

Visualizing Rembrandt is a web application that displays Rembrandt's data through an interactive graph, where the user is in charge of choosing and filtering the type of data he/she would like to learn. This application is thought to fit exhibitions that include artworks from the dutch painter, and it aims to make the learnability process regarding this artist easier through a friendly user interface.In this specific case the interactivity between the user and the computer is based on the decisions the individuals may take and the data that is returned to fulfill the spectators needs. This interactivity is understood as Information Delivery, which refers to the approach of an information corpus or repository that is presented to the visitor, whether adaptively or by user selection (Wakkary, 2008, p.372). Visualizing Rembrandt is then a graphical mean to learn deeper about this particular painter, by allowing connections that would not be possible within a traditional approach towards art exhibitions.

1.1 Related Work

Interactivity in museums has been pursued from different approaches, that are not only centered in an exhibitions physical space. Museums professionals have put effort in developing interactivity in the museums websites in order to integrate it to the visitors experiences and expectations. Because of this, several museums offer to users the chance to build their own collections, where they can connect and collect information in a personal meaningful way (Marty, 2011). However, these web-based interactivity have had some problems, as users do not usually update, or even look again, the collections they have made. For this reasons studies about how visitors browse and interact with museums websites are still being made (Marty, 2007).

There are web-based examples of sites that contain artists information and artworks, but these still provide a static visualization of the involved data. The Google Art Project (http://www.googleartproject.com), for instance, allows users to browse and choose information from a broad array of artists. However, there is no possibility of visualizing connections between artists. The Google Art Projects works as a library, where the retrieval of images is efficient. One specific website that displays information about Rembrandt's artworks is The Rembrandt Database (http://www.rembrandtdatabase.org/Rembrandt), which is managed by the RKD (Rijksbureau voor Kunsthistorische Documentatie). This database is still in a beta version, and it is helpful for browsing through Rembrandt's paintings with more information that the one presented within the Google Art Project. However, in this version there is still a lack of connectivity between Rembrandt artists and their relationships.

Interactive museums guides are another trend that has grown over the past couple of years, and since the advent of audio- based museum guides, much research and development has been placed on increasing the technological capacity

for augmenting the museum visit experience. (Wakkary, 2008). Nowadays, museums guides are drifting from an audio setup where visitors had to type a specific number into special devices. It is common that museums provide ipods, or similar touch screen devices, to users in order to provide them with information. Visualizing Rembrandt is an application that can be embedded in this guides, in order to allow users to handle and manage the information.

2 Connecting Rembrandt to the Web

The objective of browsing through Visualizing Rembrandt is to enhance the museums visitors learning experience by providing more information about the artist and those who were somehow connected to him in as little space as possible and in a more dynamic way. For this reason the application can be displayed within a museum, but it can also be incorporated into the museums website. In the latter case, users can search in Visualizing Rembrandt from their computer or laptops at home. One of the many benefits of using information technologies as a tool for communication is that it allows to provide bigger amounts of information in little space and in different formats which, at the same time adapt to visitors communication preferences and reinforces the content transmission (Pujol-Tost, 2011).

2.1 Technical Setup

The interface has educational purposes, hence the search and result of information should be efficient, effective and satisfactory. Users do not need to have experience with advanced technology, however they should be familiar with basic web navigation. This application was built under the Java Enterprise Edition platform using PrimeFaces, d3.js and jquery for the front-end. And it is easily customized to be used with a different data set of artists.

2.2 Data Set

The data visualization graph presents information obtained from a data set provided by the Rijksbureau voor Kunsthistorische Documentatie at the Netherlands Institute for Art History and it consists of a list of 89 artists relevant to Rembrandt along with information about themselves, their artistic work (around 142 paintings in total) and details of the paintings when available. A previous research on the use of information visualization in museums show that data needs to suffice short and long-term explorations (Hinrichs 2008). Therefore the names of the artists are presented in a Hierarchical Bundle Graph (Fig.1) to allow an easy and simple overview of all the spectrum of connections, and invites the visitors to explore an artist more deeply if desired.

After the user has chosen the type of information he wants to see the information is displayed as the following: At the canvas a connection between Rembrandt and the other artists will be shown depending on the selection made by the user.

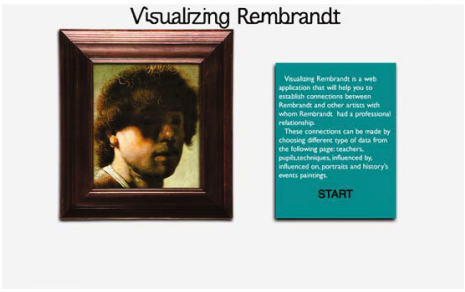


Fig. 1. Index

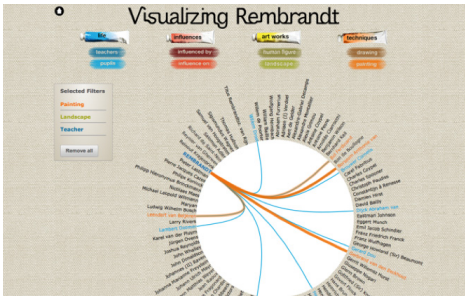


Fig. 2. Canvas

The connection is made with a line using the color of the selected sub-menu. For example, if a user wants to see who was a pupil of Rembrandt he/she should select the PUPIL submenu. This is light blue, a line starting from Rembrandt will end on the other artists that were his Pupils. The same principle is applied with the rest of the categories.

In order to ease the learnability of the system categories and colors are implemented. This is based on two concepts defined by David Benyon in his book Designing Interactive Systems: metaphors and blends. The categories are visualized with a painting metaphor with the use of oil tubes and paint strokes, as a Metaphor is generally seen as taking concepts from one domain and applying it to another(Benyon, 2010). However, metaphors are really blends in the interaction design, as explained by Benyon, because it takes input from at least two spaces, the characteristics of the domain described by the source and the characteristics of the target that we are applying it to (Benyon, 2010). The painting metaphor is blended with the graph in order to create a user friendly navigation. In addition of the graph and the visualization of data relationships, the user can learn more about specific artists with a pop up menu that shows after selecting a name (Fig. 3).

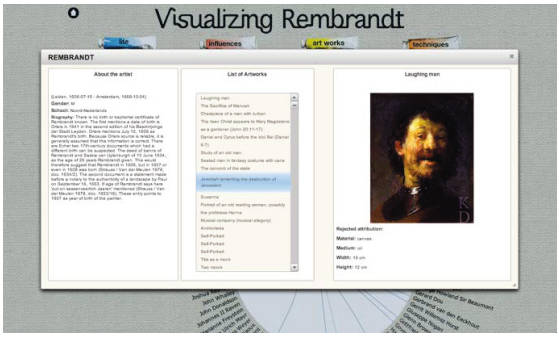


Fig. 3. Popup

For the sake of creating a satisfactory and effective experience the categories are easy to identify. The items that can be selected under each category have a clickable affordance and the action they trigger after being selected is instantly visible. The system was designed in order to make the tasks very easy, so there is a limited number of ways to complete tasks. It is intended to be a straightforward application.

2.3 Navigation

The interface has two main pages: one for the index (where instructions are explained - Fig. 1), and another for the canvas (where data is visualized -Fig. 2). The index resembles a painting hanging in a museum. Paintings of Rembrandt in a wooden frame change every time the page is open. Next to the frame there is a set of instructions of how to use the interface and a start button. After clicking the start button the user enters the navigation area which follows the painting metaphor. Here he finds a white canvas and four categories: LIFE (blue), INFLUENCES (red), ARTWORKS (yellow) and TECHNIQUES (orange). With their corresponding sub-menu items: LIFE : Teachers/ Pupils, INFLUENCES: Influenced by/ Influenced on, ARTWORKS: Human Figure / Landscape, TECHNIQUES: Drawing/ Painting

After the user clicks over the item the connections will be displayed at the canvas. A list of the selected items will be created at the left column, in order to remember the user the selected items. Beneath the list there is as well an erase all option.

3 Evaluation and Results

Three user evaluations were made during the developing process. A User Participation Evaluation and a SUS (System Usability Scale) was selected to measure the usability and experience. The first was made at a lo-Fi level, the second was a more functional prototype and the third a complete version of the application. From the lo-Fi prototype several technical suggestions were made to improve the prototype. The main goal of the second user evaluation was to evaluate the interface design, the navigability and the interactivity of Visualizing Rembrandt. Finally, the third evaluation was carried out with five users, the result showed a stronger positive feedback, as users comprehended better the application, its use and the role that it would have in a museum. The prototype at this stage incorporated the whole list of artists within the main circle, hence the new remarks made by the users for this were about the size of the circle and font size (in order to ease the readability). It is important to mention that in the last evaluation the users were more impressed with the application, for its novelty, its creativity and the way it can be implemented in museums in order to improve the visualization of data with the use of graphs. Users expressed their willingness to use the application when visiting an exhibition, as it would help them see and learn information by new means.

4 Conclusion and Discussion

One of the most important findings of the project was the acceptance of the idea to provide interactive means to display data at museums. Eventually this would make museums to be more aware of the data they own and the way they share it to the public. Initiatives like the Visualizing Rembrandt application can be of great importance in museums, because it will enable the spectators to approach the information in a different way. Users could make decisions of the type of information they would like to read. This will empower the museums visitors with decision making in an exhibition, where the common situation is that the spectator has no big participation.

5 Future Work

The process of developing Visualizing Rembrandt and its results showed that is a very appealing concept. Future developments of this project can be made for different exhibitions and museums. The project can be implemented in more exhibitions, and a new way of presenting data within museums could become popular. However, more tests need to be done, specifically in real contexts, to measure the way the application would be use in a real exhibitions. These would lead to make considerations about how the application can affect the visitors behaviours, or even if it would create a gap between the real objects and the digital information. On the other hand, improvements can be made in the system, by adapting it to other type of devices that would allow visitors to have an even more personal use, like with tablets and touch screens.

References

1. Benyon, D.: Designing Interactive Systems. A comprehensive guide to HCI and interactive design, 2nd edn. (2010)
2. Google Art Project, `http://www.googleartproject.com`
3. RKD Rijksbureau voor Kunsthistorische Documentatie, `http://website.rkd.nl`
4. The Rembrandt Database, `http://www.rembrandtdatabase.org/Rembrandt`
5. Laia Pujol-Tost: Integrating ICT in exhibitions. Museum Management and Curatorship 26(1), 63–79 (2011)
6. Hinrichs, U., Schmidt, H., Carpendale, S.: EMDialog: Bringing information visualization into the museum. IEEE Transactions on Visualization and Computer Graphics 14(6), 1181–1188 (2008)
7. Wakkary, R., et al.: Situating approaches to interactive museum guides. Museum Management and Curatorship 23(4), 367–383 (2008)
8. Marty, P.F.: My lost museum: User expectations and motivations for creating personal digital collections on museum websites. Library & Information Science Research 33(3), 211–219 (2011)
9. Marty, P.F.: Museum Websites and Museum Visitors: Before and After the Museum Visit. Museum Management and Curatorship 22(4), 337–360 (2007)
10. Bell, G.: Making Sense of Museums. The Museum as 'cultural ecology'. InteLabs (2012)

Stylistic Walk Synthesis Based on Fourier Decomposition

Joelle Tilmanne and Thierry Dutoit

Numediart Institute, University of Mons, Mons, Belgium
joelle.tilmanne@umons.ac.be
www.tcts.fpms.ac.be\~tilmanne

Abstract. We present a stylistic walk modeling and synthesis method based on frequency analysis of motion capture data. We observe that two peaks corresponding to the walk cycle fundamental frequency and its first harmonic can easily be found for most walk styles in the Fourier transform. Hence a second order Fourier series efficiently represents most styles, as assessed in the subjective user evaluation procedure, even though it results in a strong filtering of the original signals and hence a strong smoothing of the resulting motion sequences.

Keywords: motion capture, synthesis, Fourier transform.

1 Introduction

A broad field of applications can be found for human motion analysis and synthesis: entertainment (games, animation, etc.), medical applications, sports, artistic performances, etc. When considering humanlike motion synthesis, most methods aim at modeling and synthesizing motion sequences in the temporal domain, using a wide range of methods [1] such as keyframe animation [2], procedural models, motion graphs[3], Principal Component Analysis [4], Hidden Markov Models [5,6,7], etc. However, another way of understanding and analyzing motion data is to study it in the frequency domain. This temporal to frequency domain transformation can be obtained through Fourier analysis.

Troje [8] models motion by a continuous component, the walk fundamental frequency and its first harmonic. He models the Cartesian coordinates of 15 body markers and reports that keeping only these three contributions retains 99% of the variance of the input data. Bruderlin et al. [9] use multiresolution filtering for decomposing motion into several frequency bands whose amplitudes can be modified in order to change the motion style. Unuma et al. [10] apply Fourier transform on motion and modify its aspect by changing the weights of the Fourier coefficients.

In the present work, we applied a Fourier decomposition to walk motion sequences, either normal walks or walks presenting exaggerated styles. The Fourier decomposition was applied to the angular representation of the motion (i.e. angles applied to the joints of a skeleton with constant limb lengths) rather than to the 3D coordinates of the skeleton joints, and to walks presenting various styles, expressed with very diverse influences on the resulting walk motion.

M. Mancas et al. (Eds.): INTETAIN 2013, LNICST 124, pp. 71–79, 2013.

2 Method

2.1 Stylistic Walk Frequency Content Analysis

We used motion capture data from our previously recorded databases (Tilmanne et al. [11]) which contain walk sequences at several speeds (*eNTERFACE'08* database) and with different styles (*Mockey* database) for 18-articulations skeletons. The tridimensional angle data is represented through the exponential map parameterization, and hence by three values for each articulation or joint angle. Figure 1 displays the amplitudes of the Fourier coefficients for a discrete Fourier transform of the 54 exponential map values representing the 18 body joint angles during one normal walk sequence of the *eNTERFACE'08* database. The Fourier transform is calculated on each one of the 54 exponential map values, and the 54 resulting Fourier transforms are superimposed in the figure.

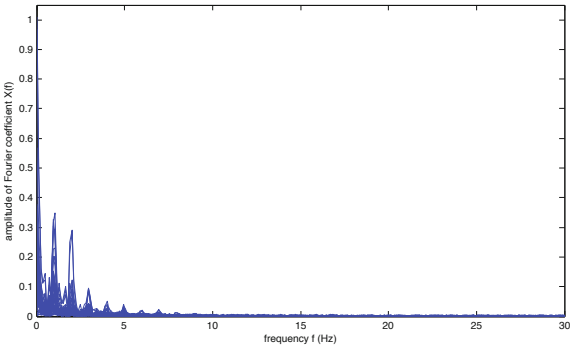


Fig. 1. Discrete Fourier transform of one normal walk sequence, for the 54 exponential map values of the 18 body joint angles

This discrete Fourier transformation clearly outlines the periodical nature of gait. Figure 1 displays a strong constant component for the frequency $f = 0$. This amplitude corresponds to the mean pose of the skeleton during the walk sequence. Outside from this first component, a noticeably higher peak appears at the fundamental frequency of the walk sequence, around 1Hz, which corresponds to the frequency of one complete walk cycle (= two steps) during the analyzed walk. The next highest peak is found at the first harmonic of the fundamental frequency, around 2 Hz, and corresponds to the frequency of one step.

The same profile of Fourier coefficient amplitudes was observed for all the speeds (slow, normal and fast) and 41 walkers of the *eNTERFACE'08* database. When analyzing more complex or elaborated walk styles, like the exaggerated walk styles present in the Mockey database, the profile of Fourier coefficients is sometimes more complex. Figure 2 illustrates the Fourier analysis of one continuous walk sequence for each one of the eleven Mockey styles. The red line corresponds to the mean walk cycle frequency of the analyzed sequence, calculated thanks to our automatic segmentation algorithm.

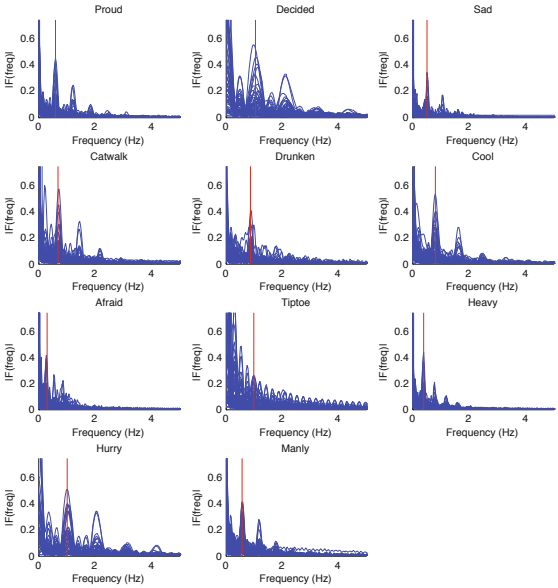


Fig. 2. Discrete Fourier transform of the eleven stylistic walk sequences from the Mockey database, for the 54 exponential map values of the 18 body joint angles. The red line represents the mean frequency of the walk sequence.

We observe that the two peaks corresponding to the walk cycle fundamental frequency and its first harmonic can easily be found for most styles. In general, however, the Fourier coefficient amplitudes display a more complicated profile over the eleven Mockey styles than for the normal walk of Figure 1. For some styles, like the *tiptoe* walk, the Fourier coefficients even display a completely different profile. Furthermore, the length of the available walk sequence can be an issue for an optimal evaluation of the Fourier coefficients and the Fourier coefficients cannot be calculated with the same precision for all the walk styles.

2.2 Fourier Based Stylistic Motion Modeling and Synthesis

Based on the Fourier analysis illustrated in Figure 1, it is obvious that the main components of a basic human walk can be modeled by taking into account only a few components of the Fourier transform. This approach has been introduced by Nikolaus Troje [12,8], who "linearizes" the periodic walk motion data by modeling it using only the first two components of the corresponding Fourier series. Each variable P_i of the motion data is represented by one continuous component and two cosines corresponding to the walk fundamental frequency and its first harmonic, as expressed in Equation 1:

$$P_i(t) = p_{i,0} + p_{i,1} \cos(\omega_0 t + \phi_{i,1}) + p_{i,2} \cos(2\omega_0 t + \phi_{i,2}) + err_i \qquad (1)$$

Each motion variable is hence represented by five time-independent parameters ($p_{i,0}, p_{i,1}, \phi_{i,1}, p_{i,2}, and\ \phi_{i,2}$), plus a sixth parameter independent of the considered motion variable (the fundamental frequency ω_0). The err_i term represents the difference between the original variable P_i and its modeling and is hence set to zero for the synthesis stage.

In his work, Troje applies the Fourier series decomposition to the Cartesian coordinates of the body joints. However, we used the angle data motion representation which we found more suitable for our synthesis purposes. In addition to avoiding synthesized motions implying varying limb lengths, this joint angle representation gets rid of one more non-kinematic factor influencing motion perception: skeleton proportions. Troje's work mainly focuses on gender recognition based on point light display walk sequences and shows that the male versus female walk recognition is based both on the static pose (ratio between shoulders and hip width for instance) and kinematics parameters. Since our final goal is to be able to control interactively the walk of a given virtual character, we do not want our style parameterization to be influenced by skeleton size related information. Basing our Fourier analysis on the angle data parameterization enables the synthesis model to rely on kinematic features only. As illustrated in Figure 1, the Fourier analysis of exponential map parameterization also displays peaks around the walk sequence fundamental frequency and first harmonic, as the Fourier analysis of cartesian coordinates used by Troje.

Let us consider a data matrix X of size 54 x N corresponding to one walk sequence of N frames parameterized by 54 exponential map values x_i ($i = 1, ..., 54$) at each frame k. The first step of our Fourier series decomposition consists in determining the $p_{i,0}$ coefficient of Equation 1. This coefficient is very easily calculated since it corresponds to the mean of each one of the 54 x_i:

$$p_{i,0} = \frac{1}{N} \sum_{k=1}^{N} x_{i,k} \qquad i = 1, ..., 54 \tag{2}$$

The next step consists in determining the fundamental frequency ω_0 of the walk sequence to be modeled. The peaks can be observed in the Fourier Transform coefficients, and extracted very easily for the eNTERFACE'08 database walks. Nonetheless, as illustrated in Figure 2, the highest peak does not always correspond to the walk fundamental frequency, especially when more complex walk styles are considered. However, if we want the Fourier analysis to linearize our data and enable us to compare our different styles, a common representation must be kept. We therefore decided to apply Troje's modeling procedure to all of our walk styles, even when it implies the loss of important frequency contributions, since these additional contributions corresponded to different frequencies for each style (in opposition to the fundamental frequency and its first harmonic which are present in every walk style).

In order to determine which peak of the Fourier coefficient amplitudes corresponds to the fundamental walk frequency, even when it was not the highest one, we used the step durations extracted thanks to an automatic step segmentation. The fundamental cycle mean period T_{mean} is obtained by multiplying this mean

step duration by two, and the inverse of the cycle mean period T_{mean} gives a mean cycle frequency $f_{mean} = \frac{1}{T_{mean}}$. As observed in Figure 2 (red lines), this f_{mean} frequency corresponds to the walk cycle fundamental frequency peak of the Fourier transform. We hence set $f_0 = f_{mean}$, and the fundamental angular frequency ω_0 is thus known: $\omega_0 = 2\pi f_0$.

Once the fundamental angular frequency ω_0 is known, the corresponding coefficients can easily be calculated. The complex Fourier coefficients $c_{i,1}$ and $c_{i,2}$ are then calculated by a projection of the data on the exponential axis corresponding to the fundamental frequency and to its first harmonic:

$$c_{i,1} = \frac{1}{N} \sum_{k=1}^{N} x_{i,k} * \exp^{-j\omega_0 k} \tag{3}$$

$$c_{i,2} = \frac{1}{N} \sum_{k=1}^{N} x_{i,k} * \exp^{-j2\omega_0 k} \tag{4}$$

The corresponding coefficients $p_{i,1}$, $\phi_{i,1}$, $p_{i,2}$ and $\phi_{i,2}$ are calculated thanks to $c_{i,1}$ and $c_{i,2}$:

$$p_{i,l} = 2 * abs(c_{i,l}) = 2 * \sqrt{real(c_{i,l})^2 + im(c_{i,l})^2}\ , \qquad i = 1, ..., 54,\ l = 1, 2 \tag{5}$$

$$\phi_{i,l} = phase(c_{i,l}) = \arctan \frac{im(c_{i,l})}{real(c_{i,l})}\ , \qquad i = 1, ..., 54,\ l = 1, 2 \tag{6}$$

Each of our 54 motion variables $P_i(t)$ being represented by five parameters plus one common parameter (fundamental frequency), one complete walk sequence of length N is represented by 271 values.

Using these 271 values and Equation 1, new walk sequences of any chosen length can be synthesized. This second order Fourier decomposition is illustrated in Figure 3 for 3 of the 54 original angle data values of the motion captured walk sequence. The original data used to calculate the Fourier coefficients is presented in the first subplot and its approximation through the second order Fourier series defined by Equation 1 is illustrated in the last subplot. The individual contributions of the continuous, fundamental frequency and first harmonic components are presented in subplots two to four.

However, even if new walk sequences of any chosen length can be synthesized, for a given style each step will remain exactly identical during the whole walk sequence. Furthermore, this decomposition process is equivalent to a very strong filtering since only the fundamental frequency and its first harmonic are kept. This approach hence results in a heavy smoothing of the synthesized data. However, new styles can be synthesized by taking new values in the 271-dimensional Fourier coefficient space and using them to synthesize walk sequence according to Equation 1 or by interpolation between the existing styles. If the new styles remain close to the space area determined by our original walks, they will lead to believable walk sequences.

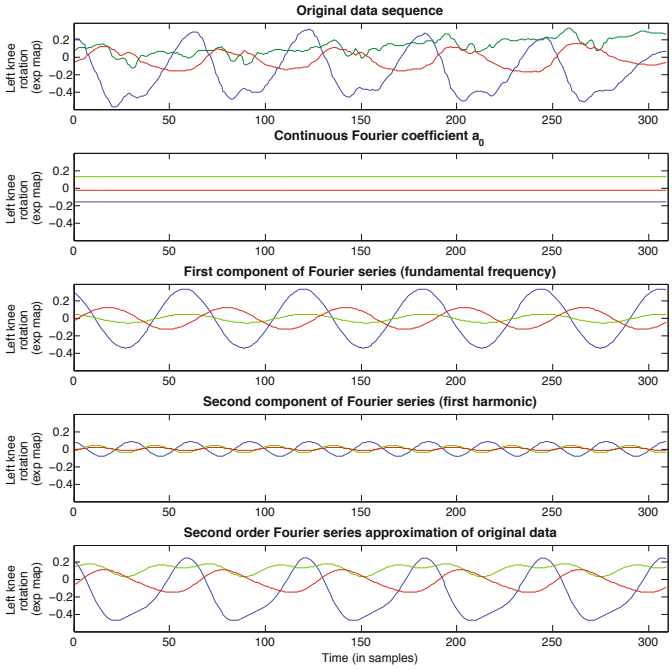


Fig. 3. Decomposition of a motion captured walk sequence into a second order Fourier series, illustrated for the three values of the rotation of the left knee, expressed in the exponential map parameterization (*expmap* 1, 2 and 3 illustrated in blue, green and red respectively). The original motion capture sequence is illustrated in the first subplot. The three next subplots represent respectively the continuous component of the Fourier decomposition, the fundamental frequency component and the first harmonic. The last subplot presents the approximation of the original data by the second order Fourier series.

3 Qualitative User Evaluation

In order to evaluate the quality of the Fourier synthesis process, a qualitative user evaluation was conducted. Thirty-seven subjects took part in the unsupervised web-based evaluation: 16 males and 21 females, with an age of mean 35.6 and standard deviation 12.6. The evaluation was divided into three series of 15 tests, with a maximum of 45 evaluations per participant.

In each test iteration, the participant was presented two videos at the same time. He could play each video as many times as he wanted. He was asked to position the synthesized stylistic walk on a scale from 0 to 4 compared to the original stylistic walk (0 being "not realistic" and 4 being "as natural as the original walk"). In the video, the motions to be assessed were applied on a simple blue stick-figure character.

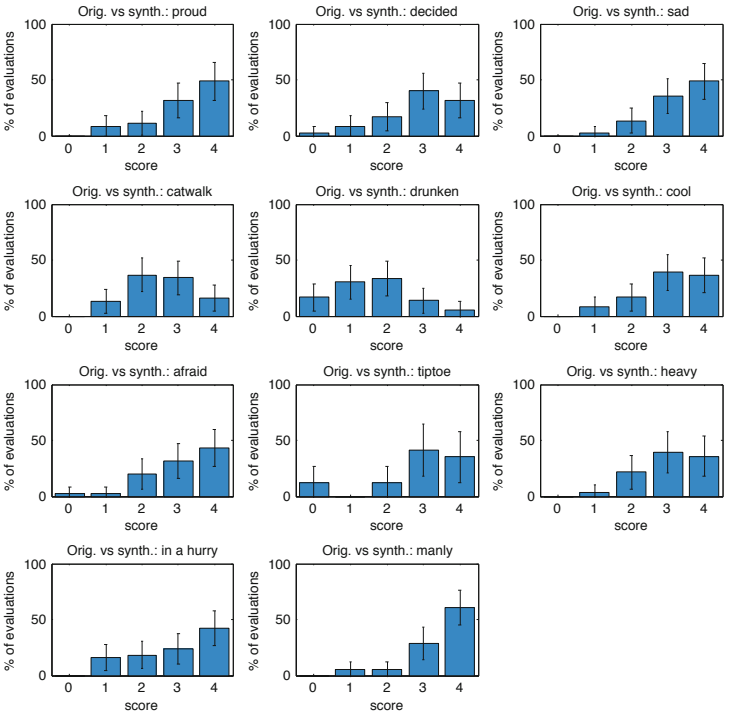


Fig. 4. Scores given by the evaluators to the synthesized walk sequences (Fourier-based approach) compared to original walk sequences, with 95% confidence intervals, for each on of the eleven styles of the Mockey database. A score of four corresponds to a synthesized walk as natural as the original walk, and a score of zero to a non realistic synthesized walk.

Two different method were assessed during this evaluation, and among the 67 tests of the complete evaluation set, 21 video pairs aimed at assessing the perceived naturalness of the Fourier series based synthesis. 390 evaluations were performed on these 21 walk pairs. The second method assessed was based on a principal component analysis (PCA) and is beyond the scope of the present paper.

Figure 4 displays the style by style results of the evaluation and Figure 5 the mean score per style. With a mean score of 2.91 but with wide variations between the extreme scores, as style specific mean scores range from 1.60 (drunken walk) to 3.45 (manly walk), the naturalness perception varies widely.

The overall score is quite good since evaluators were sensitive to the exaggerated nature of the original walk styles. Since the two cosines based recomposition smoothed part of the stylistic content of the walk, some evaluators orally reported that they found the synthesized sequences more believable than the original walk sequences. The results displayed in Figure 4 illustrate the fact that this method has some interest for periodical walk styles displaying a simple Fourier transform profile, but that it does not work in walks displaying important variations

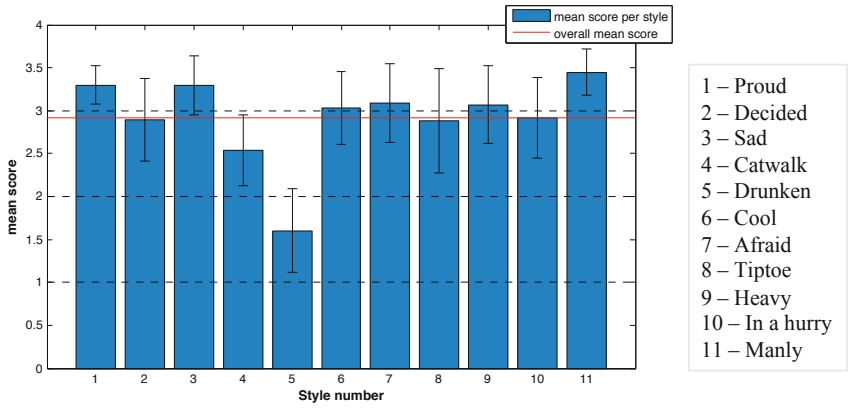


Fig. 5. Mean scores given by the evaluators to the synthesized walk sequences (Fourier-based approach), with 95% confidence interval, for each on of the eleven styles of the Mockey database. The overall mean score is displayed as a red line.

or more complex structures than periodical motions of the limbs aligned to the walk cycle frequency and its first harmonic.

4 Conclusion

Fourier analysis enables us to study the profile of different walk styles into the frequency domain. We apply Troje's approach to stylistic walks sequences, but adapt it and use it on exponential map angle data rather than Cartesian coordinates, so as to have a style representation independent of skeleton size and hence of the walker's morphology.

As can be observed in the Fourier analysis illustrated in Figures 1 and 2 and in the qualitative evaluation results, the walk modeling with a second order Fourier series applies better to normal walks than to stylistic walks which display more complex Fourier coefficients patterns and tend to display more than the two basic peaks observed in normal walks. This is especially the case for walk styles such as the *drunken* walk, where part of the style lies in the fact that two successive steps are very different from each other, which cannot be modeled with a periodic time series. However, we were able to apply the same procedure to all the walk styles and to analyze the resulting synthesized sequences. As can be foreseen, the synthesized sequences appear smoothed, since keeping only the contributions of two frequencies in addition to the constant component is equivalent to a strong filtering. With this approach, no cycle to cycle variation is possible since the walk sequence is described as a strictly periodic pattern, and even the small periodic variations that compose the motion are filtered out. It has to be noted that such a modeling technique is based on the periodical nature of gait, and hence cannot be adapted to non-periodical motions.

However, this Fourier series decomposition remains interesting to analyze the motion and its basic components. It is worth noticing that for most styles, a recognizable walk, even if not realistic, can be synthesized with as few as two cosines and one constant for each variable.

References

1. Forsyth, D.A., Arikan, O., Ikemoto, L., O'Brien, J., Ramanan, D.: Computational studies of human motion: Part 1, tracking and motion synthesis. Foundations and Trends in Computer Graphics and Vision 1(2-3), 77–254 (2005)
2. Williams, R.: Animator's Survival Kit: A Manual of Methods, Principles, and Formulas for Classical, Computer, Games, Stop Motion, and Internet Animators. Faber & Faber (2002)
3. Kovar, L., Gleicher, M., Pighin, F.: Motion graphs. In: ACM SIGGRAPH 2008 Classes, p. 51. ACM (2008)
4. Glardon, P., Boulic, R., Thalmann, D.: PCA-based walking engine using motion capture data. In: Computer Graphics International, pp. 292–298. IEEE (2004)
5. Brand, M., Hertzmann, A.: Style machines. In: Proceedings of the 27th Annual Conference on Computer Graphics and Interactive Techniques, pp. 183–192. ACM Press/Addison-Wesley Publishing Co. (2000)
6. Li, Y., Wang, T., Shum, H.Y.: Motion texture: A two-level statistical model for character motion synthesis. In: Proceedings of the 29th Annual Conference on Computer Graphics and Interactive Techniques, pp. 465–472. ACM (2002)
7. Tilmanne, J., Dutoit, T.: Continuous control of style and style transitions through linear interpolation in hidden markov model based walk synthesis. In: Gavrilova, M.L., Tan, C.J.K. (eds.) Transactions on Computational Science XVI. LNCS, vol. 7380, pp. 34–54. Springer, Heidelberg (2012)
8. Troje, N.F.: Retrieving information from human movement patterns. In: Shipley, T.F., Zacks, J.M. (eds.) Understanding Events: How Humans See, Represent, and Act on Events, vol. 1, pp. 308–334. Oxford University Press (2008)
9. Bruderlin, A., Williams, L.: Motion signal processing. In: SIGGRAPH 1995 Proceedings, pp. 97–104 (1995)
10. Unuma, M., Anjyo, K., Takeuchi, R.: Fourier principles for emotion-based human figure animation. In: SIGGRAPH 1995 Proceedings, pp. 91–96 (1995)
11. Tilmanne, J.: Joelle tilmanne's webpage, `http://tcts.fpms.ac.be/~tilmanne/`
12. Troje, N.: The little difference: Fourier based gender classification from biological motion. Dynamic Perception, 115–120 (2002)

Automatically Mapping Human Skeletons onto Virtual Character Armatures

Andrea Sanna, Fabrizio Lamberti, Gianluca Paravati, Gilles Carlevaris, and Paolo Montuschi

Dipartimento di Automatica e Informatica, Politecnico di Torino, Torino 10129, Italy
{andrea.sanna,fabrizio.lamberti,gianluca.paravati,
gilles.carlevaris,paolo.montuschi}@polito.it
http://www.polito.it

Abstract. Motion capture systems provide an efficient and interactive solution for extracting information related to a human skeleton, which is often exploited to animate virtual characters. When the character cannot be assimilated to an anthropometric shape, the task to map motion capture data onto the armature to be animated could be extremely challenging. This paper presents a novel methodology for the automatic mapping of a human skeleton onto virtual character armatures. By extending the concept of graph similarity, joints and bones of the tracked human skeleton are mapped onto an arbitrary shaped armature. A prototype implementation has been developed by using the Microsoft Kinect as body tracking device. Preliminary results show that the proposed solution can already be used to animate truly different characters such as a Pixar-like lamp, a fish or a dog.

Keywords: virtual character animation, automatic armature mapping, motion capture, graph similarity.

1 Introduction

The animation of virtual characters is an exciting and challenging task. The mesh describing the shape of a character is linked to a set of bones usually named armature (rigging). The manipulation of the armature allows the user to animate the character. The traditional approach uses forward and inverse kinematics techniques to fix a set of key frames (poses), which will be automatically interpolated by the animation program [1][2]. This approach has been often outperformed by motion capture solutions [3]. In this case, animators motions are tracked and can be recorded to a computer and then applied to the characters or directly used to make interactive animations.

Each method has its advantages and drawbacks. On the one side, keyframing can produce animations that would be difficult or impossible to act out. However, complex actions can be both very difficult and time consuming to reproduce. On the other side, motion capture can reproduce in a very accurate, fast and smooth way a variety of human (and animal) movements. Nonetheless, capture

M. Mancas et al. (Eds.): INTETAIN 2013, LNICST 124, pp. 80–89, 2013.

systems are, in general, very expensive and motion data could be hard to use for animating different kinds of characters (though a solution to partially cope with this latter limitation has been proposed in [4]).

The present paper addresses the problem to match the human skeleton of the animators (which act out the scene as if they were the characters to be animated) on a generic armature in an automatic and efficient way. In a number of previous works, such as [5] [6] and [7], the above association was implemented manually. Unfortunately, manual association can be a time consuming and difficult task, since only skilled animators are generally able to immediately identify the *best* match.

The proposed solution aims to find an efficient mapping of the human skeleton onto the virtual character armature, by exploiting an extended graph similarity criterion. In short, the matching algorithm analyzes and transforms a skeleton into a graph describing its constituting parts and the connections among them. Several parameters are taken into account, namely armature topology, general user preferences, symmetry and motion constraints. The above parameters are used to compute a similarity matrix, which is then exploited to associate each bone in the considered armature to the *most similar* bone in the human skeleton. Users can either accept the association that has been automatically found or modify the proposed bone mapping according to their own needs.

The current implementation uses armatures defined in Blender [8] and the Microsoft Kinect sensor [9] as motion capture device. Nonetheless, the proposed methodology is general and it could be easily extended to any other tracking system. A mapping between the human skeleton tracked by the Kinect and the Blender armature allows the devised method to translate the local movements of a human skeleton bone into translations and rotations of the related armature bones. In this way, a markerless motion capture system for digital puppetry with generic characters is actually implemented.

The paper is organized as follows: Section 2 briefly reviews other approaches that have been designed to map a human skeleton onto a virtual character armature. Section 3 presents the proposed solution and, in particular, shows how the similarity scores in the mapping matrix are computed. Section 4 proposes some applications of the matching algorithm to non-anthropometric virtual characters.

2 Background

One of the first attempts to interactively control anthropometric limbs by inverse kinematics is proposed in [10]. However, this method is only meant for controlling sub-parts of the skeleton independently. Hence, it is not always able to cope with constraints that require the whole body to be animated. The issue of controlling all the body parts is tackled in [11]. Here, the goal is to map the movements made by a performer onto an animated character by only considering constraints on the end-effectors. An extension targeted to the control of non-anthropometric characters is proposed in [12]. In this latter work, an intermediate skeleton with less degrees of freedom is used and the remaining degrees of freedom are computed analytically.

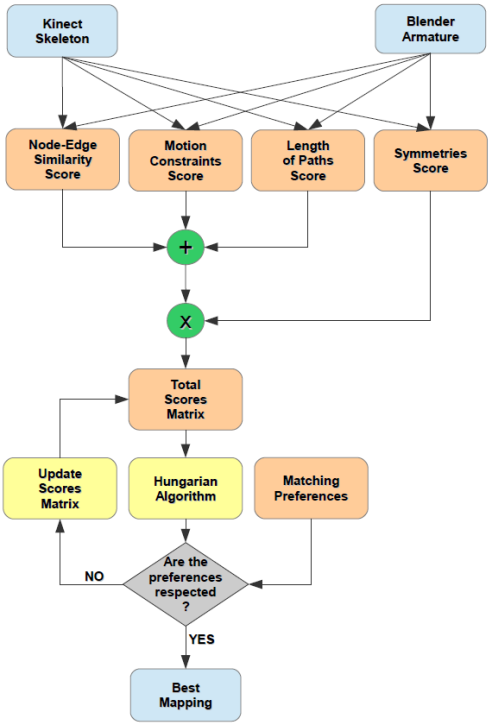


Fig. 1. Flow-chart of the mapping algorithm

The above works basically relies upon specific representations of motion (e.g., through simplified skeletons). A comparable approach is also used in [13], where a data structure especially dedicated to motion adaptation is proposed. Moreover, in all the works considered, the focus is mainly on human-like animation.

One of the most recent and impressive approaches known in the literature to animate generic-shape characters by the skeleton of the animator is reported

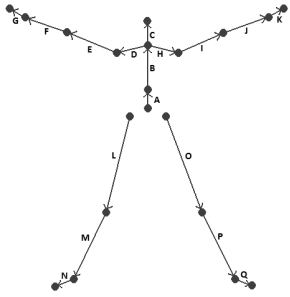


Fig. 2. Human skeleton extracted by the tracking application via the Microsoft Kinect

in [14]. The solution proposed allows the animator to directly manipulate the mapping of the skeleton onto the mesh of the object to be controlled. The character mesh is segmented from the background. Then, the body of the animator is *embedded* to position vacated by the object. By a vocal command, limbs of the animator are attached to the parts of the mesh the body overlaps. These attachments serve as constraints for the deformation model that is inspired by the Embedded Deformation method proposed in [15].

3 The Proposed Solution: Skeleton Mapping

Figure 1 shows the flow-chart of the mapping algorithm, which will be described in details in the following sections. From the chart, it can be easily noticed how different blocks actually contribute to determine the score matrix that is used to map the human skeleton onto the Blender armature. Such blocks consider armature topology details (e.g., node-edge similarity), motion constraints, length of kinematic chains and symmetries. Moreover, user preferences can be exploited to *force* some mappings, thus possibly overriding other criteria.

3.1 Graph Representation

The first step of the matching algorithm analyzes both the skeleton produced by the tracking device and the Blender armature of the virtual character to be animated.

As mentioned in the Introduction, the Microsoft Kinect is used as capture device in this work. Tracking data contain information about the center of mass and the position of each of the twenty joints of the captured skeleton, along with the status of each joint. Status information indicates whether the joint position is being tracked or inferred (which happens when the Microsoft Kinect cannot see this point and tries to accurately *guess* it based on information from previous frames and neighboring joints). Joints tracked for the skeleton are split into three main sections:

1. the central area, containing the head, the neck, the spine and the hip center;
2. the arms, containing for each arm the shoulder, the elbow, the wrist and the hand;
3. the legs, containing for each leg the hip, the knee, the ankle and the foot.

Figure 2 shows the skeleton extracted by the Kinect application. A socket connection is created between the Kinect application and a Blender Python script controlling the execution of the program inside the Blender Game Engine (BGE). The script constantly receives user's skeleton data from the Kinect application, computes the necessary transformations required and applies them to the armature to be controlled (for more details about the software architecture see [5]).

Blender armature is explored starting from the root bone and a mathematical description is generated for it. In particular, bones are associated to graph nodes/vertices and relations between nodes are mapped to graph arcs/edges.

The graph can be represented by an adjacency matrix [16]; given a graph G_A, $G_A = G(V_A, E_A)$ where V_A are the vertices and E_A are the edges, if the cardinality of V_A is n_a, then the adjacency matrix A of this graph is a $n_a \times n_a$ matrix in which entry $[A]_{ij}$ is equal to 1 if and only if $(i, j) \in E_A$, 0 otherwise. The adjacency matrix of an undirected graph will always be symmetric. Another useful graph representation is obtained by means of pair of matrices called edge-source matrix A_s and edge-terminus matrix A_t. This representation allows self-loops to be considered in the graph [16]. Let $s_A(i)$ denote the source of edge i, and let $t_A(i)$ denote the terminus of edge i. Then A_s and A_t can be defined as follows:

$$[A_s]_{ij} = \begin{cases} 1 & \text{if } s_A(j) = i \\ 0 & \text{else} \end{cases}$$

$$[A_t]_{ij} = \begin{cases} 1 & \text{if } t_A(j) = i \\ 0 & \text{else} \end{cases}$$

The graph representation given by A_s and A_t has the following properties:

- the adjacency matrix A is equal to $A_s A_t^T$;
- $A_s A_s^T$ is equal to a diagonal matrix D_{A_s} with the out-degree (i.e., the number of outgoing edges) of node i in the i-th diagonal position;
- $A_t A_t^T$ is equal to a diagonal matrix D_{A_t} with the in-degree (i.e., the number of incoming edges) of each node in the corresponding diagonal entry.

3.2 Node-Edge Similarity Scores

The approach presented above uses an iterative procedure to assign a similarity score between pairs of nodes belonging to two different graphs, thus allowing a match between the two bone configurations (i.e., the skeleton extracted by the Kinect application and the Blender armature). In the similar way as [16], the matching strategy is based on the coupled node-edge method. This method returns similarity scores considering not only the node similarity scores, but also edge similarity.

Given two graphs G_A and G_B, a simple way to give a definition of an edge score is: an edge in G_B is like an edge in G_A if their source and terminal nodes are similar, respectively. A is the adjacency matrix of G_A and B the adjacency matrix of G_B. D_{A_s}, D_{A_t} and D_{B_s}, D_{B_t} are the diagonal matrices containing the out-degree and the in-degree values of every node in G_A and G_B, respectively.

By iterating a certain number of times (usually, a satisfactory convergence is obtained with 11 iterations [16]) equation (1), a $n \times m$ scores matrix X is obtained, where n is the total number of bones in the Blender armature and m is the number of bones in the Kinect skeleton.

$$x_k \longleftarrow (A \otimes B + A^T \otimes B^T + D_{A_s} \otimes D_{B_s} + D_{A_t} \otimes D_{B_t}) x_{k-1}. \tag{1}$$

The symbol $\otimes$ represents the Kronecker's matrix product, k the k-th iteration and x_k a column vector obtained by concatenating the columns of the scores

matrix X. The iteration method presented well recognizes nodes that are very similar and provides good results if one of the two graphs is a subgraph of the other one.

Nonetheless, the above score is not sufficient to denote the similarity between the two armatures. Hence, other parameters (beyond the graph topologies) need to be taken into account in order to propose the user an efficient and accurate bone mapping.

3.3 Motion Constraints Scores

Another parameter to be considered in the mapping process is represented by the motion constraints related to each bone: two bones (one of the human skeleton and one of the Blender armature) exhibiting similar degrees of freedom should be preferred by the mapping technique. For example, it is easier to control a virtual segment exhibiting a high degree of freedom by means of a hand rather than with the spine. These constraints are taken into account while updating the scores matrix X by assigning a *penalty* to bones with degrees of freedom that are different. In particular, a value proportional to the existing difference is applied. If two bones have completely different movement types, their score is set in such a way that they cannot be matched. Motion constraints scores are added to node-edge similarity scores (see Figure 1).

3.4 Length of Paths Scores

Another criteria considered to update the matching scores is represented by the length of kinematic chains: a long path of connected bones should be mapped onto a similarly long path. A sort of *bonus* is added (see Figure 1) to the score of all those bone pairs that share the same position in a chain starting from the bone root. The bonus enhances the probability that a bone in the Blender armature, placed in a certain position, will be mapped onto a bone placed in the same position in the human skeleton. Updating the matrix of scores according to this criterion brings another advantage: the probability of mapping sequential bones in a chain by respecting the natural rank order is implicitly increased.

3.5 Symmetries Scores

Armatures may exhibit one or more symmetries. This behavior is easy to verify in models like animals, where some parts of the body (like the legs) are symmetric and they could be subdivided into small groups. A main symmetry in the human skeleton can be obtained by splitting left parts from right parts. This kind of symmetry can be present also in Blender armatures. Usually, a Blender animator marks a bone in the left or right parts of the skeleton by adding the suffix ".L" or ".R" to the end of the bone's name, respectively. By searching for the bones containing ".L" or ".R" in their names, it is possible to force a mapping of these bones onto the corresponding parts of the human body. To this purpose, the

Kinect skeleton is split in five groups: body, left arm, right arm, left leg and right leg. Depending on the bone type in Blender, the search space is reduced to a few groups that compose the Kinect skeleton. This approach allows the proposed technique to always map the left part of a model onto the user's left arm or onto the user's left leg, and vice versa. Scores can assume the following values: 0 (to avoid mapping a bone in the left part onto a bone in the right part and vice versa), 1 (to leave unchanged the score of bones not related to symmetries, like the spine bones) and 2 (to force left part bones to be mapped onto left part bones and vice versa). Symmetries scores multiply the previously obtained scores recorded in the matrix.

3.6 Evaluation Component: Hungarian Algorithm

In order to identify the best matching between two graphs, the node pairs with the highest scores in the matrix, according to a certain evaluation criterion, have to be found. This problem is known as the maximum weight bipartite graph matching problem. A common algorithm for identifying such a maximum weight is the Hungarian algorithm, described in [17]. In the proposed technique, the Hungarian algorithm is applied to the scores matrix to obtain a matching between the nodes of the graph representing the Blender armature and the nodes of the graph representing the Kinect skeleton, which maximizes the sum of squared matched scores. Thus, for each bone in the Blender armature, the Kinect bone it will be mapped onto is obtained.

3.7 Preferences

After having performed several tests by exploiting all the previous scores, it was realized that, in many cases, the approach proposed correctly chose some parts of the Kinect skeleton like the shoulders, but these segments were indeed difficult to use for controlling and animating characters. This issue has been tackled by adding user's preferences to the overall mapping strategy.

If some armature bones are mapped onto the arms, forearms or hands are preferred to the shoulders. Of course, any other preferences could be coded in the

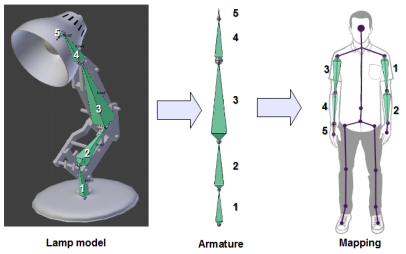


Fig. 3. Mapping of the human skeleton onto the armature used to animate a Pixar-like lamp (a video is available at `http://130.192.5.7/intetain2013/lamp.avi`)

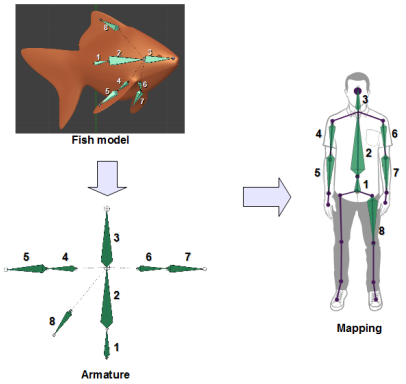


Fig. 4. Mapping of the human skeleton onto the armature used to animate a fish (a video is available at `http://130.192.5.7/intetain2013/fish.avi`)

matrix. Furthermore, the user can choose to avoid the mapping of the possible unused parts belonging to a mapped arm. The same approach is applied for the legs. After the application of the preferences, the Hungarian algorithm is used again to assign all the other unused bones.

4 Experimental Results

This section presents the application of the devised mapping approach to three non-anthropometric characters (a Pixar-like lamp, a fish and a dog), whose Blender armatures have been mapped onto the human skeleton extracted by the Kinect application. Results are shown in Figures 3, 4 and 5, respectively. The kinematics chain of the lamp is mapped part onto two bones of the left arm (bones 1 and 2) and part onto three bones of the right arm (bones 3, 4 and 5). Arms are selected here, since their preferences score is the highest one (see Section 3.7). A completely different mapping has been obtained for the fish. In this case, it can be noticed how the symmetry (see Section 3.5) related to the fins is taken into account for mapping bones 4 and 5 on the right arm and bones 6 and 7 on the left arm. Since bones 1, 2 and 3 are topologically similar to the spine of the human skeleton, they are mapped onto it. The computation of mapping scores is detailed, for this character, in the Appendix that is available at `http://130.192.5.7/intetain2013/appendix.pdf`. The dog armature is the most complex one among the three considered. Again, because of topological similarity, the spine of the dog is mapped onto the spine of the human skeleton (bones 1, 2 and 3). The symmetry is used to map the left parts of the skeleton onto the left parts of the armature and vice versa. Moreover, still because of topological similarity, the arms are mapped onto the front paws (the tail makes the similarity score for the back paws, with respect the skeleton arms, lower than the front paws). In this case, the user can choose to obtain a mapping also for the head and the tail or, as shown in the Figure 5, to leave these two bones unmapped as

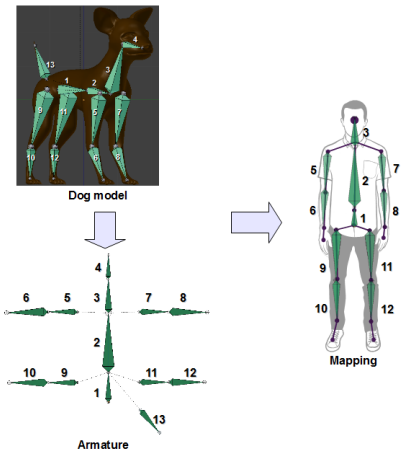


Fig. 5. Mapping of the human skeleton onto the armature used to animate a dog (a video is available at `http://130.192.5.7/intetain2013/dog.avi`)

the five main kinematics chains of the skeleton have been already (even if not completely) used.

5 Conclusion and Future Works

This paper presents a novel automatic procedure to map the human skeleton captured by a tracking system (the Microsoft Kinect, in the proposed implementation) onto the armature of a virtual character to be animated. The proposed solution is able to efficiently tackle issues related to the control of non-anthropometric characters by taking into account topological similarity scores as well as other parameters such as motion constraints, kinematic chains length, user preferences and so forth. The mapping between the armatures is static, meaning that the proposed algorithm assigns direct relationships among bones based on the analysis of the topologies of the two armatures. Both skilled animators and, above all, an audience with little or no experience in computer animation could take advantage of the devised approach. The step of mapping the animator's skeleton onto the armature of the character to be animated is efficiently automated, thus reducing time losses and frustrating attempts.

Future works will be mainly aimed to gather a larger number of user feedbacks (currently, the system is being tested by a few students of the Computer Animation course of the Master of Science degree in Computer Science at Politecnico di Torino), in order to evaluate how the proposed mapping is close/far to/from the statistically best solution. The approach could/should be improved and extended in order to let it cope with armatures including a number of bones that is larger than the one provided by the tracking device. In this case, forward kinematics could not be used for the mapping, unless the target armature is reduced by collapsing pairs of bones.

Currently, the proposed method favours the static relations between bones by exploiting morphological features of the armature to be animated; indeed, less importance is assigned to the kinetic model (degree of freedom of the bones, range of movements). Thus, future works will also consider and study in depth the kinetic mapping between the armatures to improve the translation between human movements into those of the target model to be animated.

Acknowledgments. The authors would like to thank the anonymous reviewers for their valuable comments and suggestions to improve the quality of the manuscript.

References

1. Burtnyk, N., Wein, M.: Computer generated key frame animation. Journal of the Society of Motion Picture and Television Engineers 8(3), 149–153 (1971)
2. Burtnyk, N., Wein, M.: Interactive skeleton techniques for enhancing motion dynamics in key frame animation. Communication of the ACM 19(10), 564–569 (1976)
3. Menache, A.: Understanding motion capture for computer animation and video games. Morgan Kaufmann, New York (2000)
4. Gleicher, M.: Retargeting motion to new characters. In: Proceedings of the ACM Siggraph 1998, pp. 33–42 (1998)
5. Sanna, A., Lamberti, F., Paravati, G., Domingues Rocha, F.: A Kinect-based Interface to Animate Virtual Characters. International Journal of Multimodal User Interfaces, doi:10.1007/s12193-012-0113-9
6. The Bloop project, `http://dm.tzi.de/research/hci/bloop`
7. The Brekelmans Jasper web site, `http://www.brekel.com`
8. The Blender project, `http://www.blender.org`
9. The Kinect web site, `http://www.xbox.com/kinect/`
10. Tak, S., Young Song, O., Ko, H.S.: Spacetime sweeping: An interactive dynamic constraints solver. In: Proceedings of the Computer Animation, p. 261. IEEE Computer Society (2002)
11. Shin, H.J., Lee, J., Shin, S.Y., Gleicher, M.: Computer puppetry: An importance-based approach. ACM Trans. Graph. 20(2), 67–94 (2001)
12. Monzani, J.S., Baerlocher, P., Boulic, R., Thalmann, D.: Using an intermediate skeleton and inverse kinematics for motion retargeting. Computer Graphics Forum 19(3) (2000)
13. Kulpa, R., Multon, F., Arnaldi, B.: Morphology-independent representation of motions for interactive human-like animation. Computer Graphics Forum 24, 343–352 (2005)
14. Chen, J., Izadi, S., Fitzgibbon, A.: KinÊtre: Animating the world with the human body. In: ACM SIGGRAPH 2012 Talks, pp. 39–144 (2012), doi:10.1145/2343045.2343098
15. Sumner, R.W., Schmid, J., Pauly, M.: Embedded deformation for shape manipulation. In: Proceedings of the ACM Siggraph 2007 (2007)
16. Zager, L.: Graph Similarity and Matching. Department of Electrical Engineering and Computer Science, Massachussets Institute of Technology (2005)
17. Andràs, F.: On Khun's Hungarian method - a tribute from Hungary. Egervàry research Group on Combinatorial Optimization Technical report, TR-2004-14 (2004), `http://www.cs.elte.hu/egres/tr/egres-04-14.pdf`

KinectBalls: An Interactive Tool for Ball Throwing Games

Jonathan Schoreels, Romuald Deshayes, and Tom Mens

Software Engineering Lab, NUMEDIART Research Insititute
University of Mons – UMONS, Belgium
firstname.name@umons.ac.be

Abstract. We present a tool that was developed in the context of the first author's masters project. The tool implements an interactive computer game combining the real and the virtual world in a seamless way. The player interacts with the game by throwing balls towards a wall on which a virtual 3D scene is projected. Using the *Kinect* 3D sensor, we compute and predict the trajectory, speed and position of the ball. Upon impact with the screen, a virtual ball continues its trajectory in the virtual scene, and interacts with the objects around it using a physical and a graphical 3D engine *Bullet*, and *Ogre3D*. The prototype game has been successfully tested on a large number of people of varying ages.

Keywords: Kinect, HCI, virtual reality, object tracking.

1 Introduction

Creating new games and entertainment applications using affordable state-of-the-art devices has gained a lot of recent interest thanks to the emergence of new HCI techniques and a trend towards the use of natural interaction. The first major step that revolutionized the gaming industry was Nintendo's Wii console, allowing humans to play games with body gestures. Microsoft responded with the Kinect sensor capable of seeing and reacting to the world in 3D. Since its release, an important number of applications using this sensor have been published on the internet [5,2,13,1].

While Kinect's main strength is its ability to track a user's body, it can be used in other ways to serve different goals. We exploited the raw information provided by Kinect's 3D sensor to track a moving ball. We integrated this in a prototype interactive game *KinectBalls* that bridges the gap between the real and virtual world. The aim of the game is to bring down a pile of virtual boxes by throwing a real ball towards them (see Figure 1). Videos of a live tool demonstration carried out with an audience of high school students and small children during a science fair at the University of Mons, can be found at `http://youtu.be/v02BcA-EPrI`. Others have recently developed a similar tool to simulate a pétanque game, using two high-speed PSEye camera's for ball tracking, and a webcam for face detection and tracking [3].

M. Mancas et al. (Eds.): INTETAIN 2013, LNICST 124, pp. 90–95, 2013.

Fig. 1. Left: Live demonstration of the *KinectBalls* prototype. **Right:** The projection matrix (frustum) of the 3D virtual world.

The concept of the application is fairly simple. A beamer projects a virtual 3D world on a screen or wall. The player then throws a ball towards the projected scene. Using the Kinect, the tool tracks the ball's trajectory and speed and predicts the point and time of impact with the wall. At the predicted time and position of impact, a virtual ball is created with the same parameters as the real ball, and it continues its trajectory in the virtual world to interact with other virtual objects. Although there are some technical limitations in the current prototype implementation (low framerate, low resolution) they can easily be addressed by using other input devices than the Kinect (see e.g. [3]). A wide variety of ball throwing games could be implemented in a similar way (such as basket ball, penalty shots, bowling, petanque...) These games can for instance be used in small rooms to train the motor skills of young children by aiming accurately at objects in a virtual scene.

To implement our tool, we deliberately constrained ourselves to an as affordable solution as possible using inexpensive yet state-of-the-art devices capable of 3D vision, together with open source libraries for physical 3D rendering.

2 Object Tracking

One of the biggest problems with traditional ball tracking algorithms is their inability to take into account 3D information. To address this problem, stereo vision has often been used to retrieve information about the 3D position of the ball. The main limitation of this technique is that it requires a relatively complex setup. For example, to track a tennis ball, up to 6 cameras have to be placed and calibrated together before being able to accurately track a moving ball [12].

The problem we are addressing is more simple than tennis ball tracking because the area where the ball needs to be tracked is much smaller and the hand-thrown ball to be tracked moves at a much slower speed (a few m/s). On the other hand, the system has to respond instantly, that is to say, even before the ball hits the wall. This requires us to predict the ball's trajectory and its impact with the wall. We will see in Section 3.2 how this can be achieved.

Over the last decade, a lot of effort has been put in finding robust algorithms to track a moving ball in realtime (e.g., for soccer games, tennis games, golf etc.)

[15]. Using 2D cameras, the main challenge resides in being able to differentiate between the object to be tracked and the rest of the scene. To do so, many techniques exploit chromatic and morphological [7] features of the object to be tracked. Other techniques for object tracking have been developed since. The most widely used one is the so-called *fiducial*-based tracking. This technique uses visible markers placed on the tracked object, thus improving the speed, robustness and accuracy of the tracking algorithm [9,14]. The major drawback of such techniques is that they are invasive as the tracked objects have to carry markers. When objects with strong contours have to be tracked on non-cluttered scenes, a RAPID-like method can be used [8,10]. The main advantage is that this method is quite simple and was also one of the first methods to run in real-time. Many enhancements for this method have been proposed to make it more robust, such as the use of a more complex least-squares curve fitting method [6]. A more detailed overview of the different kinds of tracking techniques can be found in the excellent survey proposed by [11] that discusses most popular model-based 3D tracking methods.

3 Architecture of KinectBalls

We have developed our tool in a modular way, in order to facilitate changes to (1) the characteristics of the moving object, (2) the game application and (3) its 3D rendering. Figure 2 shows the 4 modules of our tool: data acquisition, object detection, trajectory prediction and graphical rendering.

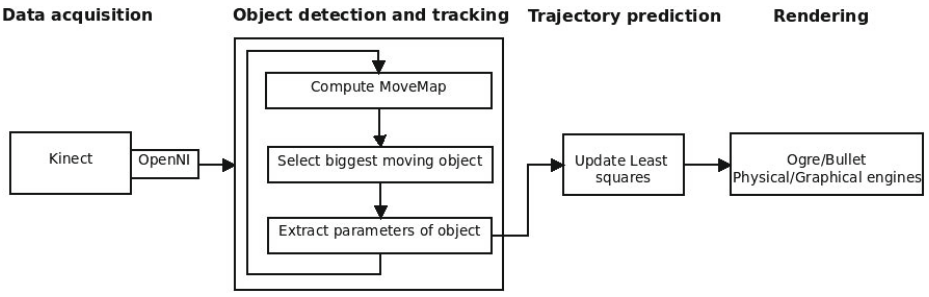


Fig. 2. Architecture of the prototype framework

3.1 Object Detection and Tracking

The most widely used affordable 3D sensor today is undoubtedly the Kinect. Its infrared projector and sensor allow to analyze and create a complete depth map of the observed scene in real-time at a framerate of 30Hz. We used this sensor to ease the detection and tracking of moving objects.

An important challenge is to differentiate between the background and the moving object. The infrared camera provides a set of successive frames representing snapshots of the observed scene and constructs a *depth map*, i.e., a 3D image where each pixel has three (x,y,z) values representing the exact position in metric space w.r.t. the camera. By comparing a frame F_n with the previous frame F_{n-1}, we compute the difference in depth (z-axis) for each pixel. If this difference exceeds a certain threshold T (allowing us to filter out noise) and if a sufficient number of adjacent pixels have undergone a similar difference in depth, we conclude that something has moved. We apply this idea to create a matrix

$$MoveMap(i,j) = \begin{cases} 1 & if\ |F_n(i,j) - F_{n-1}(i,j)| > T \\ 0 & otherwise \end{cases}$$

Every 1 in $MoveMap$ corresponds to a moving pixel in terms of depth. The 0's are considered as being part of the background. Using this matrix we can easily track moving objects using the following setup. The camera faces the wall on which a virtual scene is projected to which the ball will be thrown, thus the only moving object that will be detected is the ball (since all other objects will not change position). However, due to imprecision of the 3D sensor, the edges of some of the objects composing the scene might still be detected as moving. Therefore, to improve the robustness of the tracking algorithm, we look for the biggest square of 1's in $MoveMap$. We can assume with a fairly high confidence that the biggest moving thing in the scene is the ball.

The next step of the algorithm is to detect the shape of the moving object. With the technique of the biggest square, we only get an approximation of the moving object's shape. By refining our algorithm, we consider all the adjacent 1's, compute the centroid of this new shape and use this point as the position of the ball to approximate the trajectory.

3.2 Trajectory Prediction

To predict the moving ball's trajectory, we store the centroid computed on each frame F_n. As soon as we have at least 2 positions of the ball (i.e., two frames in which a sufficiently big square was detected), we use a least squares regression model to approximate the trajectory of the ball with 3 second-degree polynomials (1 polynomial for each axis). At each new frame where a moving ball is detected, we update the regression model by taking into account the newly detected position of the ball.

Knowing the exact 3D position of the wall on which the virtual scene is projected, we use the computed regression model to predict the position and the time at which the ball will hit the wall. The speed of the ball is also computed using the derivative of the position. The closer the ball gets to the wall (and the more data points are used), the more precise the trajectory prediction will be. At the predicted time of impact, a virtual ball is created at the predicted position. The virtual ball will continue its route using the regression model parameters provided by the trajectory approximation algorithm.

To transform a position in the real world to a position in the virtual world, we convert the coordinate system of the real world (given by the Kinect) into the coordinate system of the virtual world (computed by the projector's parameters and its position relative to the screen and the Kinect). Thus, we calibrated [4] these two devices by calculating their intrinsic parameters. The extrinsic parameters are estimated using a matrix $M = (R, T)$ containing the rotation R and the translation T to be applied.

To create an immersive virtual world, we modified the projection matrix defined by a perspective frustum (i.e., a pyramid lying between two parallel planes cutting it, see right of figure 1) of the 3D rendering engine to match the projector's intrinsic parameters. We measured the distance between the projector and the screen, the vertical and horizontal size of the screen to set the frustum parameters. This way, we can use the same scale in both worlds, and we can easily create an impression of a virtual box (increasing the level of realism when the ball continues through the virtual world).

4 Lessons Learned and Future Work

We tested our setup with an Intel i5 computer with 4Gb RAM and an ATI 7850 graphical processor during a full day in front of a live audience. Calibration was a challenge, since it depends on the angle of width and position of both the projector and the Kinect relative to the screen. We did not take into account the distortion of the camera and projector because the precision of the Kinect was not sufficiently high to gain any important benefit. For higher resolution devices, distortion should be taken into consideration.

Kids of 5 years and older interacted with the game very enthusiastically and without requiring any explanation. With a supple throw, between 4 to 10 successive positions of the ball were detected. With this amount of points, the precision of the predicted impact of the ball varied between 1 and 5 centimeters. Only when the ball was not long enough in the field of view, or when it was thrown too fast or too straight, the resolution and framerate of the Kinect did not allow to compute the trajectory correctly. Doubling the framerate to 60Hz would mostly solve this problem.

Some adult players reported a lack of immersion, because they had difficulties to interpret the 3D virtual world, as it was only projected in 2D on the screen. Using a stereoscopic 3D projector could address this problem. Another way to make the game more immersive is by tracking the position of the player w.r.t. the projector using a second Kinect device. The rendered virtual scene can then be adapted to match the user's relative position to the screen.

The game could be extended into a multiplayer version with multiple balls thrown simultaneously. This would require to identify different balls (e.g. based on their color), and detecting possible collisions between them.

5 Conclusion

We presented *KinectBalls*, an interactive game capable of tracking a moving ball thrown towards a projected virtual scene. The technique used for achieving this requires only one very affordable 3D sensor to track the ball and predict its trajectory and impact on the wall. The developed algorithms are fast enough to run in real time on a standard computer. After calibration, the solution worked fine in all tested indoor situations, but various improvements can be made to increase the level of immersion.

References

1. Bailly, G., Walter, R., Müller, J., Ning, T., Lecolinet, E.: Comparing free hand menu techniques for distant displays using linear, marking and finger-count menus. In: Campos, P., Graham, N., Jorge, J., Nunes, N., Palanque, P., Winckler, M. (eds.) INTERACT 2011, Part II. LNCS, vol. 6947, pp. 248–262. Springer, Heidelberg (2011)
2. Boulos, M.K.N., Blanchard, B., Walker, J.M.C., Tripathy, R.G.-O.A.: Web GIS in practice X: A Microsoft Kinect natural user interface for Google Earth navigation. Int'l J. Health Geographics 10, 14 (2011)
3. Dalpé, S., Monat-rodier, J., Riendeau, G., Voutsinas, P.: Poly-pétanque. In: Int'l Meeting on Virtual Reality and Converging Technologies, LAVAL VIRTUAL (2013), `http://youtu.be/0Z2VDdaS3rs`
4. Deshayes, R.: Reconstruction algorithmique d'objets 3D. Master's thesis, Faculty of Sciences, University of Mons, Belgium (June 2011)
5. Deshayes, R., Jacquet, C., Hardebolle, C., Boulanger, F., Mens, T.: Heterogeneous modeling of gesture-based 3D applications. In: MoDELS Workshops (2012)
6. Drummond, T., Cipolla, R.: Real-time visual tracking of complex structures. IEEE Trans. Pattern Anal. Mach. Intell. 24(7), 932–946 (2002)
7. Gong, Y., Sin, L.T., Chuan, C.H., Zhang, H., Sakauchi, M.: Automatic parsing of TV soccer programs. In: IEEE Int'l Conf. Multimedia Computing and Systems (ICMCS), pp. 167–174 (1995)
8. Harris, C.: Tracking with rigid objects. MIT Press (1992)
9. Hoff, W.A., Nguyen, K., Lyon, T.: Computer vision-based registration techniques for augmented reality. In: Intelligent Robots and Computer Vision XV, pp. 538–548 (1996)
10. Klein, G., Drummond, T.: Robust visual tracking for non-instrumented augmented reality. In: IEEE/ACM Int'l Symp. Mixed and Augmented Reality (ISMAR), pp. 113–122. IEEE Computer Society (2003)
11. Lepetit, V., Fua, P.: Monocular model-based 3D tracking of rigid objects: A survey. Foundations and Trends in Computer Graphics and Vision 1(1), 91 (2005)
12. Pingali, G.S., Opalach, A., Jean, Y.: Ball tracking and virtual replays for innovative tennis broadcasts. In: Int'l Conf. Pattern Recognition, pp. 4152–4156 (2000)
13. Ren, Z., Meng, J., Yuan, J., Zhang, Z.: Robust hand gesture recognition with Kinect sensor. In: ACM Int'l Conf. Multimedia, pp. 759–760 (2011)
14. State, A., Hirota, G., Chen, D.T., Garrett, W.F., Livingston, M.A.: Superior augmented reality registration by integrating landmark tracking and magnetic tracking. In: SIGGRAPH, pp. 429–438 (1996)
15. Tong, X.-F., Lu, H.-Q., Liu, Q.-S.: An effective and fast soccer ball detection and tracking method. In: Int'l Conf. Pattern Recognition, pp. 795–798 (2004)

Medianeum: Gesture-Based Ergonomic Interaction

François Zajéga[1], Cécile Picard-Limpens[1], Julie René[2], Antonin Puleo[1], Justine Decuypere[1], Christian Frisson[1], Thierry Ravet[1], and Matei Mancaş[1]

[1] Numediart Institute, University of Mons, Bd. Dolez 31, Mons, Belgium
francois.zajega@umons.ac.be
[2] ISEN, Ecole d'Ingénieurs de Lille, France

Abstract. The proposed MEDIANEUM system consists in an interactive installation allowing general audiences to explore a timeline and access informational multimedia data such as texts, images and video.

Through a Microsoft Kinect depth sensor, users' skeletons are captured and their gestures are tracked to interact with the data presented on a screen in an ergonomic way.

The graphical user interface is built upon *ProcesSwing*, our version of the *Processing* IDE embedded into a standard Swing Java GUI widget toolkit application, and the TimelineJS library from Vérité.co/Northwestern University, allowing to create online, personalized and interactive timelines that mash up historical events, sorted in definable categories.

Keywords: timeline, Kinect, ergonomics, gestures, interface, interaction.

1 Introduction

The Medianeum project is an attempt to use gesture-based interfaces in museums. The prototype described in this paper is used at the Mundaneum [1], the archive centre of the French Community of Wallonia-Brussels and Temporary Exhibition Space. The application of the proposed gesture interface is the interaction of visitors with a timeline containing the life of Henri La Fontaine, one of the founder of the *Mundaneum*, and the related historical events. The proposed interface is able to handle the general audience which visit a museum. People use the interface to interact with the timeline in order to explore the different historical events presented/supported as video, images and texts and shown on a screen.

In section 2 we describe the proposed captured system. Section 3 describes the novel approach introduced in this paper for an optimized interaction. Sections 4 and 5 detail the use of events by the system and the graphic of the interface providing feedback tu the user. Finally we conclude in section 6.

[1] Mundaneum: http://www.mundaneum.org

M. Mancas et al. (Eds.): INTETAIN 2013, LNICST 124, pp. 96–103, 2013.

2 Capture System

To accelerate development , *Processing*[2] [2] has been chosen as a basis. It proposes a wide variety of libraries that are ready to use and easy to deploy. However, *Processing* was too limited to develop the whole application: we missed GUI elements such as sliders or buttons, as well as the possibility to integrate an HTML rendering engine. The core of *Processing* has been embedded into a standard Swing Java GUI widget toolkit application. We named this swing compliant version of *Processing ProcesSwing* [3]. It is available for download on the numediart website[3].

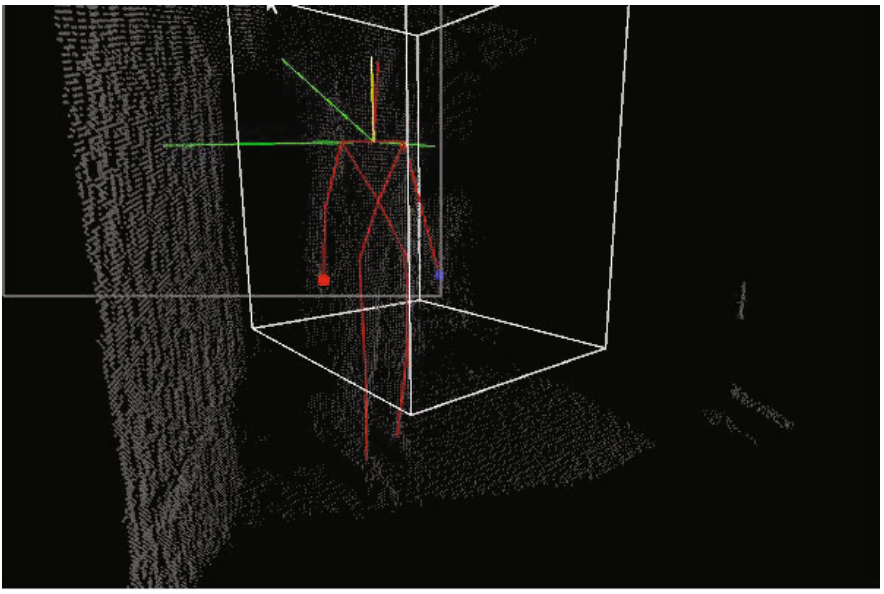


Fig. 1. We define an active area where the user can freely move its upper-body. The active area is represented by a parallelepiped in the virtual world.

Medianeum uses the Microsoft Kinect sensor which provides hands-free control capacities. For this purpose we use NITE, the open source drivers along with motion tracking middleware from PRIMESENSE, in association with OpenNI [4] [1] which allows us to retrieve the skeleton of users (in red in Figure 1).

We define *active* areas where the users must be located to be able to interact with the system. The area is defined by a parallelepiped in the virtual world and is marked as a dot on the ground or a ray of light in the physical world (see Figure 1). When the user, represented as its skeleton in the virtual world,

[2] Processing: `http://processing.org`
[3] numediart tools: `http://www.numediart.org/tools`
[4] OpenNI: `http://openni.org`

penetrates this parallelepiped, the system checks that both shoulders are in the active area. As soon as this condition is fulfilled, the user is considered as active.

3 Interaction Design

3.1 Early Approach: Planar Representation of the Gesture Movements

Most of the existing software providing similar interaction as KinVi [4] use a planar representation of gestures (Figure 2).

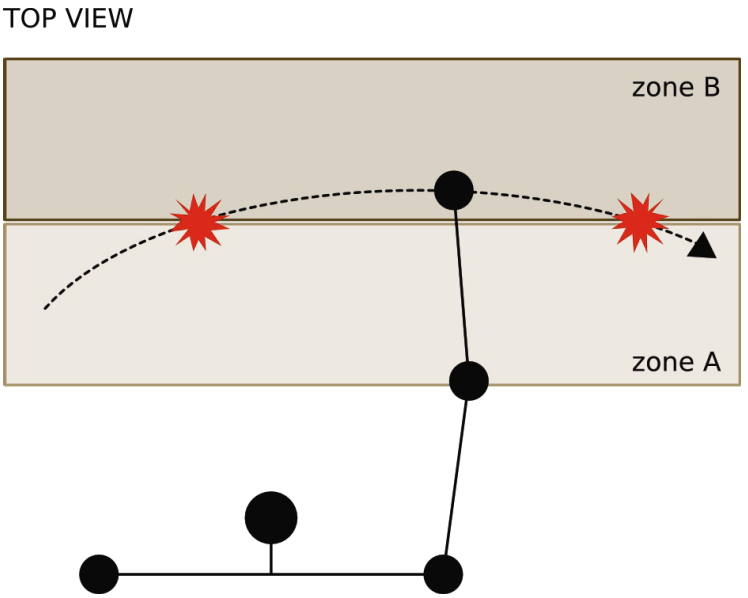


Fig. 2. In the KinVi approach, the relative depth of the hand is varying during displacement in the layout

However, when we try to imitate Kinvi-like interactions (Figure 3) we rapidly realize the limits of such an approach. In this case, a pop-up window showing the boxes and hands' positions in those boxes must be displayed for the user to understand the setup (Figure 2). In particular, the method has two main drawbacks:

1. Size of the pop-up window: to be understandable, the pop-up screen has to be relatively big. As a consequence, the screen area available to the content exploration is reduced.

   In addition, this pop-up window has been shown to distract the user from the important information.

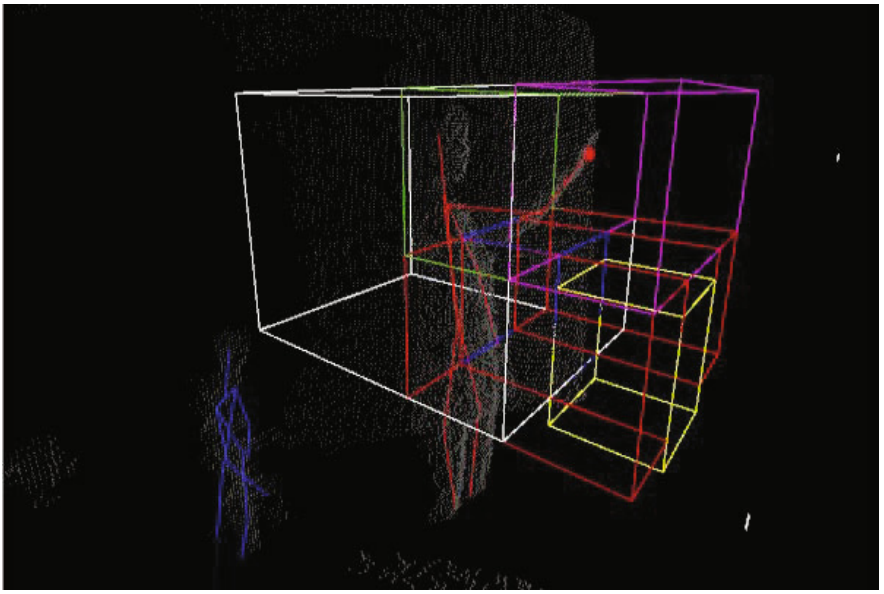


Fig. 3. First tests with a similar approach than KinVi interaction design. We place a big virtual parallelepiped, called *layout*, in front of the active area and divide it in two layers. The first layer is equivalent to a *mouse over* or *highlight* in a standard GUI environment. The second layer allows the user to click. Buttons are small parallelepiped, some having the two stages, some having only the highlight stage.

2. Ergonomics of human-system interaction: buttons are difficult to locate and very sensitive to shaking. In this case, a large layout is easier to use (wider areas) but forces the user to bend in order to reach the bottom buttons.During the displacement in the layout, the relative depth of the hand is varying (see Figure 2). This is not compatible with the parallelepiped shape of the buttons. To avoid that issue, we have to deepen all the buttons. This is, once again, inducing wider movements to reach the click area which can be painful in long interactions.

3.2 Final Approach: Spherical Representation of the Gesture Movements

In this case we developed a method which is intuitive enough to avoid displaying any pop-up window with user avatar and virtual controllers. The hand is simply represented as a pointer on the screen, leaving a larger space to the content to be explored.

However, this is not solving the ergonomic problem of parallelepiped shapes. A 2D dimensional projection of the hand on a virtual plane involves losing the possibility to use the depth of the pointer, and we want to keep this opportunity.

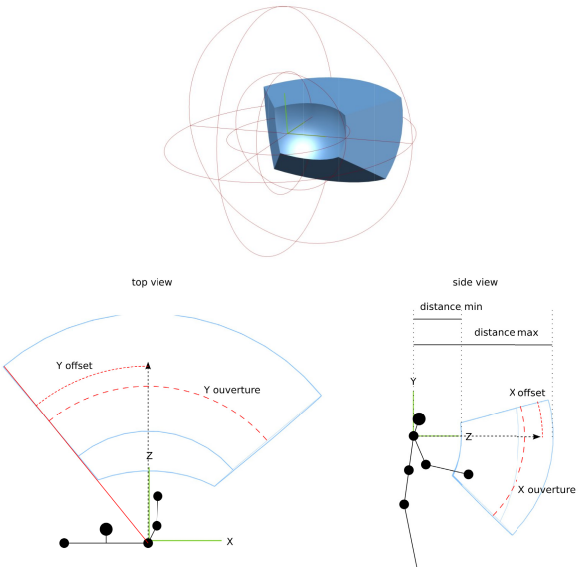


Fig. 4. Left: the parallelepiped layout that represents the active area is twisted to turn it into a portion of a sphere. Right: the position of the hand is calculated as a spherical position and the layout size is defined by four angles and two length.

We observe that when a user moves his arm from left to right, he doesn't follow straight lines. In our case, the user is not placed close to a table, a desk nor a wall. Thus, having no direct physical contact with a planar surface, the arm is moved following the natural morphology of the body. Based on this observation, we twist the virtual parallelepiped, called *layout*, to turn it into a portion of a sphere (see Figure 4, Left).

We attach the sphere to the shoulder linked to the active hand. The position of the hand is now calculated in spherical coordinates and no more in Cartesian ones.

X and Y-coordinates of the hand pointer are processed from the angles of the hand/shoulder vector in the XZ-plan and YZ-plan respectively. The Z-coordinate is calculated as the distance between the hand and the shoulder compared to the minimum and maximum sphere radius. The comparison with minimal and maximal values allows us to normalize all the axis between 0 and 1. This way of retrieving a 3 dimensional position increases theoretically the precision of the movement (see Figure 5). A planar projection tends to reduce the amplitude when the movement approaches the line perpendicular to the plane.

This layout is user-centric. By gluing the centre of the sphere to the shoulder, the interaction area stays at the right location, even if the user moves. The sphere is adapted according to the morphology of the user. The radius of the sphere is calculated proportionally to the length of the torso/neck distance, these two points being the most stable in the skeleton.

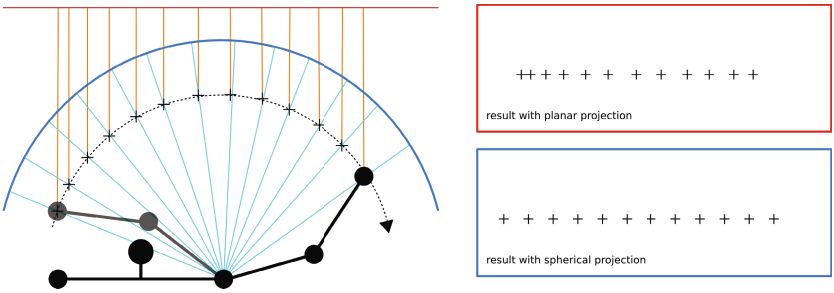


Fig. 5. A planar projection tends to reduce the gesture amplitude when the movement goes further from the barycentre of the skeleton, which is not the case for a spherical projection

In this configuration, we are able to detect click-like movements when the distance between the hand and the shoulder is quickly increasing.

4 Events Management

Skeleton movements of the active user is analysed by the system for appropriate feedback. We define events for interaction and two main categories of events are fired by the system.

1. *State changed* events
   - **User entering the active area.** This event is triggered when the two shoulders of the user are inside the parallelepiped defined as the the active area. If any is set as active yet, the user is flagged as active, preventing any other skeleton to be considered as active.
   - **User leaving the active area.** This event is triggered when the two shoulders of the user are outside the parallelepiped defined as the the active area. If the leaving user was active, the system will accept any other user in the active area.
2. *Interaction* events (the interaction events are obviously emitted when a user is set active.)
   - **Active user's hand entering the layout for the first time.** A specific event is triggered when the hand of the user enters the layout for the first time.
   - **Active user's hand inside the layout.** At each iteration,the position of the hand's user is broadcasted if if the hand is inside the layout.
   - **Active user's hand leaving the layout.** A specific event is triggered when the hand of the user leaves the layout.

5 Graphical Interface

The graphical interface is built on web technologies. Chromium is used as rendering engine and is embedded as a viewport using the `http://www.eclipse.org/swt/`SWT library[5] [5]. This allows us flexibility: since HTML is not compiled, it can be modified via a simple text editor without influencing the core of the application.

We use TimelineJS[6] library from Vérité.co/Northwestern University to display multimedia content. TimelineJS also proposes several ways to load the content: via the Google Doc API, via plain HTML, or via JavaScript Object Notation (JSON). TimelineJS is developed to be used with a mouse-like device. Due to the relatively poor precision of the pointer compared to a standard mouse, we implement a upper layer for control. All the buttons are set bigger and the drag method that allows vertical scrolling in the timeline is made easier (see Figure 6). To achieve this, a set of Javascript methods detect collisions with invisible areas placed above the interface. The active areas are also managed via Javascript to track interaction with the different elements.

Fig. 6. The graphical interface showing the timeline and the controllers for content navigation

Before launching the timeline, a series of screens welcome the user, allow him to select his language and give him a very short tutorial (see Figure 7). A screensaver is shown when no user is in the active area.

[5] `http://www.eclipse.org/swt/`
[6] `http://timeline.verite.co`

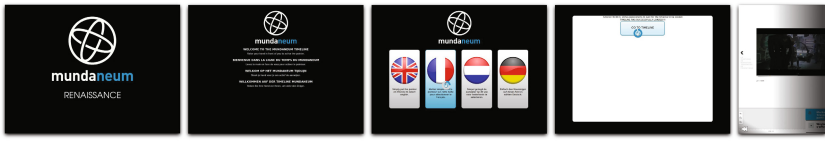


Fig. 7. The graphical interface showing the timeline and the controllers for content navigation

6 Conclusions

We built gesture-based system for complex scenarios such as museums where the users are heterogeneous which is:

1. Reliable from the capture point of view:
   An open space, unmanaged, is typically the worst place to install a precise motion detection.
2. Intuitive in terms of interface:
   This exhibition will be accessible to a broad audience and we can not suppose that the audience is already trained to computer interfaces similar to multiple touch screen (smart-phones, tablets, etc) and data exploration.
3. Robust in terms of installation:
   The exhibition lasts seven months, the Mundaneum being open seven days a week.
4. Accessible for fast Content Management:
   Content could be updated by the Mundaneum staff without the need of nay technical person.

We will further investigate user reactions in museums but also in other applications like TV control while sitting or standing, alone or with other users, etc.

3D gestures recognition will also be integrated to send specific events to any interface providing linked data to the users.

Acknowledgments. Numediart is a long-term research program centered on Digital Media Arts, funded by Région Wallonne, Belgium (grant N°716631). This work has been partly supported by Communauté Wallonie-Bruxelles under the Research Action ARC-OLIMP (grant N°AUWB-2008-12-FPMs11).

References

1. OpenNI library, `http://www.openni.org`
2. Processing: A java-based programming tool, `http://www.processing.org`
3. Numediart Institute/ UMONS, ProcesSwing, `http://www.numediart.org/tools`
4. KinVi 3D: A Kinect-Enabled Virtual Interface Gadget for Windows Control, `http://www.kinvi3d.net/`
5. Eclipse, SWT: The Standard Widget Toolkit, `http://www.eclipse.org/swt/`

About Experience and Emergence - A Framework for Decentralized Interactive Play Environments

Pepijn Rijnbout[1], Linda de Valk[1], Arnold Vermeeren[2], Tilde Bekker[1], Mark de Graaf[1], Ben Schouten[1], and Berry Eggen[1]

[1] Department of Industrial Design, Eindhoven University of Technology
P.O. Box 513, 5600 MB Eindhoven, The Netherlands
{p.rijnbout,l.c.t.d.valk,m.m.bekker,m.j.d.graaf,
b.a.m.schouten,j.h.eggen}@tue.nl
[2] Dept. of Industrial Design Engineering, TU Delft, P.O. Box, 2600 AA Delft, The Netherlands
a.p.o.s.vermeeren@tudelft.nl

Abstract. Play is an unpredictable and fascinating activity. Its qualities can serve as an inspiration for design. In designing for play, we focus on play environments with players and multiple interactive objects. The current understanding of how to design these objects and interaction opportunities to create meaningful interactions and engaging user experiences is limited. In this paper we introduce a framework focusing on the development of decentralized interactive play environments for emergent play. This framework combines knowledge from different fields including play, user experience, emergent behavior and interactions. Two case studies demonstrate its use as a tool for analysis.

Keywords: Framework, open-ended play, emergence, user experience, interactions.

1 Introduction

Imagine a playground with interactive objects that can be touched, crawled into or climbed on. Children can run around and use these objects in their play. For example, they can follow a bright light that jumps from one object to another. If they catch the light, it changes color. The children are challenged and feel competition: who catches the light first? Such a playground offers freedom to children to create their own play and provides triggers to renew play.

Play is an intrinsically motivated activity situated outside of everyday life and with no direct benefit or goal [9]. Play is unpredictable [2] and unstable [9]; it can constantly be changed or disturbed. Since long, people have been designing for play. Toys have been developed as objects to play with (e.g. buildings kits, dolls) and playgrounds as environments to play in (e.g. with swings, seesaws). Lately, these designs have become much more interactive, i.e. integrating interactive technology like sensors and actuators. Our research is part of the Intelligent Play Environments (I-PE) project which focuses on the development of interactive playgrounds that

M. Mancas et al. (Eds.): INTETAIN 2013, LNICST 124, pp. 104–113, 2013.

playfully persuade people to be more physically and socially active. In our view, interactive playgrounds can serve as an addition to more traditional playgrounds and they can exist next to each other.

A particular direction within designing for interactive play is open-ended play. In open-ended play, play objects offer interaction possibilities instead of ready games. Children can attach meaning to these possibilities and create their own games with them [1]. Designing for open-ended play is challenging as, in contradiction to games with rules, the emergent play behavior is hard to imagine beforehand. Environments for emergent play have the potential to lead to long-term engaging experiences. To support this emergent play, the environment has to be open, flexible and robust.

In this paper we present a framework which can serve for analyzing decentralized interactive play environments (DIPE). We define DIPE as a collection of communicating interactive elements, or agents, each with their own interaction rule set; in short, a decentralized system. These agents are able to communicate with other agents and to decide on actions based on locally available information. Decentralized systems have the ability to self-organize, to adjust to a wide variety of situations including many that were not foreseen in the design stage. Furthermore, they sometimes have emergent properties. Other benefits of decentralized systems are its scalability – the self-organizing mechanisms work even at large numbers of agents – and robustness – even when a substantial number of agents would be removed the overall system still self-organizes, still keeps going [5]. These properties fit the purpose of emergent play very well. On a higher level, DIPE and its players also form a decentralized system. The emergent play that occurs in this higher-level system is what the I-PE research project aims for.

The framework presented in this paper combines our various insights from previous work. We have looked at relations between certain design decisions and the supported playful user experiences throughout the total experience of interaction [24]. Simultaneously, the framework developed by Rozendaal et al. [18] has already shown us the bigger picture, illustrating the relations between interactive systems on one side and a design aim (behavioral change) on the other side. Yet, a more systematic overview of important elements and their relations in DIPE is needed in order to better understand the complexity of environments for emergent play. The framework presented in this paper combines the three focus areas of play, interactions and emergence. Moreover, it supports the understanding of relationships between different elements within play environments for emergent play. The framework can help explain and understand design decisions. The framework illustrates the context of play in which the design is used, the designed (Micro) level and the emergent (Macro) level [5].

The remainder of the paper is structured as follows. First, we will give an overview of related work on play, interactions and emergence. Then we introduce our three-leveled framework. Next, we present two case studies and analyze them using the framework. This paper ends with a discussion and conclusion of the framework.

2 Related Work

Our research connects knowledge from various fields together, including play and games, interaction design and emergence. In this section, these fields will be discussed in more detail.

2.1 Play and Games

Previous research on designing for play covers a wide gamut. One specific direction within designing for play is open-ended or emergent play; play that is not pre-defined but actually developed during use [1]. Examples of open-ended play designs are, among others, ColorFlares [1], Interactive Pathway [20] and Morel [11]. An example of an interactive playground designed for open-ended play is described in [23].

In order for open-ended play to be successful, the design should leave room for interpretation. This process can be supported by ambiguity of interaction [6, 19]. It creates an opportunity for people to establish a personal engagement with a system as they can interpret the interactions for themselves. Play is then a result of the dialogue between players and the design. This is closely related to the theory of situated action [22] which assumes that, in contrary to Norman's action cycle [14], players do not structure their activity beforehand but that the activity develops during interaction in the context of use.

Closely related to our work is the MDA model by Hunicke et al. [10], which focuses on designing for digital games and presents three components for this: Mechanics, Dynamics and Aesthetics. Mechanics concerns the components of the game, e.g. the chess pawns and board and the official rules of chess. Dynamics refers to the behavior that comes forward during play, e.g. strategies and adaptation of the rules. Aesthetics describes the experiences of the players, e.g. expressing themselves through play or wanting to win the competition. Instead of digital games, we design for environments for emergent play. In our design approach the linearity of the MDA model is less applicable, yet we do recognize the same components. We will refer to the MDA model in the description of our framework. The component of Aesthetics is related to the playful experiences framework by Korhonen et al. [12], who identified twenty playful user experiences. In our own work, we have already built upon this work by considering the importance of time in designing interactive play objects. Three stages of play were defined as part of the overall experience of interaction [24]. These stages are: invitation, exploration and immersion. In the invitation stage potential players are attracted towards the design. Once they start exploring the opportunities for interaction, players move to the exploration stage. The immersion stage concerns the actual play experience when players decide upon their own rules and goals. In their work on interactive playgrounds, Tetteroo et al. [23] defined a design taxonomy existing of three layers: play classes, dimensions of play and playground interactions. The three layers can be used to structure the design process on how to design interactions for interactive playgrounds. In our approach the interaction opportunities are embedded in tangible objects in the playground, which implies more emphasis on how to design objects and interaction opportunities to support open ended-play in a playground with multiple objects.

2.2 Interaction Design

In the previous section, ambiguity of interaction has been mentioned as a potential design quality for open-ended play. Ambiguity of interaction has already largely been explored in the arts. An example of this is the Senster designed by Edward Ihnatowicz (see [25]). The Senster was a robotic sculpture that reacted on sounds, but was frightened by loud sounds or if someone would try to touch it. From observing people interacting with the Senster, Ihnatowicz realized that people saw a form of animal-like intelligence in it. We believe the ambiguous nature of decentralized systems embedded in play objects might provide a similar or even richer experience for play.

2.3 Emergence

The field of emergent behavior has been widely studied in natural phenomena like the flocking behavior of birds [16] or the organizational structure of ants [7]. Resnick [15] investigated how phenomena like traffic jams can be understood by analyzing them as decentralized systems with emergent properties. Van Essen et al [3] propose a new approach in using decentralized systems in interactive designs. Fromm [4] refers to emergent properties as "a property of a system is emergent if it is not a property of any fundamental element". He describes a typical difficulty encountered when designing systems with emergent properties: emergence is the 'unexpected' macro behavior of local interaction rules of elements on micro level [5]. He proposes a design strategy combining top-down and bottom-up approaches in several iterations in order to link the micro level and the macro level of an emergent system [5]. The goal is to design the macro level, yet only the local rules of the elements can be changed.

3 Framework

In this section we describe our framework. The presented framework aims at providing a structured overview of focus areas that are important for developing engaging play opportunities that have the potential to lead to different types of play experiences. It illustrates the link between designed objects and emergent events. It can support designers in explaining and understanding DIPE.

The framework is structured around three levels (see Figure 1). These levels are not mutually exclusive and can influence each other. The levels are: *Context of Play, Micro* and *Macro*. Below, all three levels will be discussed in more detail. For each level, we will give an introduction and explain their content. After that we will discuss how the different levels are related.

Context of Play. This level focuses on the context of use and the overall design aim. Understanding the context of play is important as it can provide both possibilities and restrictions for (the use of) open-ended designs. Firstly, the physical environment may already determine what kind of behavior is appropriate.

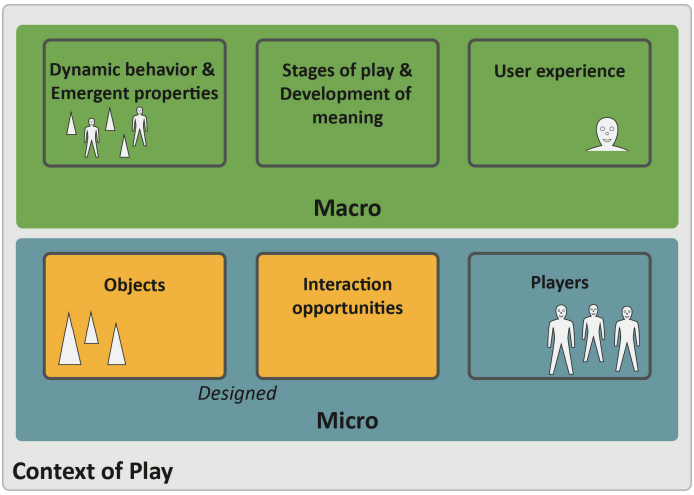


Fig. 1. Framework with three levels: Context of Play, Micro and Macro

An outdoor playground offers more freedom to move than a small inside room. Secondly, the social context largely influences play behavior. For example, whether people know each other influences if and how they interact with each other. Thirdly, the context is also shaped by the design aim. Defining this aim provides information to focus the concept development and to be able to validate the design. It concerns desired attitudes, behaviors and motivations of the users involved in interacting with the design. It illustrates which goal(s) the designer wanted to achieve and what the intended effect of use [13] is.

Micro. The Micro level describes the basic elements in a DIPE. From a design point of view, this level describes the elements of the system that are actually designed and can be directly influenced: Objects and Interaction Opportunities. It refers to the Mechanics in the MDA model [10]. If we approach a DIPE as a decentralized system, elements in this system include both the objects and the *Player*s, as interactions between objects and players lead to dynamic behavior.

Objects concern the designed parts of the system. Several aspects of the objects are relevant including the physical design, the interaction rules and the system states. The physical design refers to the form of the object, its size and the materials used. The interaction rules are the rules that describe how objects react on input of the players or of other objects, and what output they create. *Interaction Opportunities* define the possible actions that are supported by the objects at a certain time. It should be clear to players that they can explore these opportunities [14]. For instance, the affordances of the objects (e.g. a ball triggers rolling, a button triggers pressing on it), the motivating feedback (e.g. a sound when an object is shaken) and feed forward (e.g. a tile that lights up to attract attention) that the objects provide.

Macro. This level indicates which factors of the design emerge from interaction and can only be influenced indirectly. It includes Dynamic Behavior & Emergent Properties and Stages of Play & Development of Meaning. This refers to the Dynamics in the MDA model [10]. The Macro level also includes the User Experience, which refers to the Aesthetics in the MDA model [10].

Dynamic Behavior & Emergent Properties focuses on the decentralized system (of both objects and people) that changes over time. The collection of objects and people together create a dynamic system. The nature of the dynamic behavior needs to be considered when analyzing or designing DIPE. For example, a system with moving lights that speeds up when players start to interact (dynamic behavior) will most likely challenge players to speed up and be physically active. *User Experience* refers to the experiences of the people interacting with the design. Experience is part of the Macro level as it emerges from the interactions of the players in the Micro level. Generally, people tend to strive for experiences that fulfill some kind of psychological need [18, 21]. Korhonen et al. provide an extensive list of examples of playful experiences [12]. *Stages of Play* [24] refers to the dynamics of play: the total experience of interaction during play that changes over time. Another dynamic process in play is *Development of Meaning*. Players create their own rules by attaching meaning to the interaction possibilities and use these to support their current game play.

Relations between Levels. The different levels in the framework are closely related. The properties of the Micro level influence what happens at the Macro level. Emergence in the system is the result of interaction rules of the objects, and the local behavior of the players. Events on the Macro level influence the behavior of the elements in the Micro level. Therefore the two levels provide different perspectives on players.

At the Micro level, players can be considered elements of the system during play. They form a decentralized system together with the designed objects. Players are shaped by their personal characteristics (e.g. personality, mood). At the Macro level players experience interacting with the system. We cannot directly influence those experiences, as described by Hassenzahl in his work on user experience: one cannot design the experience itself, one can only design *for* an experience, e.g. increase the likelihood for an experience to happen when interacting with the product [8]. For example, the decisions of players on how to use interaction opportunities can influence the experiences that arise from interacting with the design. This is where players form a link between Micro and Macro: there is a strong two directional relation between the actual behavior of players and the experience of players. In the same way emergence at the Macro level is supported by local interaction rules of objects at the Micro level.

As both objects and people are part of the decentralized system, this makes it a hybrid system: they both influence the overall system behavior. Thus, the emergent properties are formed and influenced by people and interactive objects. When analyzing DIPE one can distinguish three different types of communication in such a hybrid system: between objects, between players and objects and between players.

4 Case Studies

In this section we present two case studies and describe them using the framework. The first case study is a design developed by the authors themselves. The aim of this case study is to further clarify the framework and how its components can be recognized in the design. The second case study covers a design developed by other researchers who are not familiar with the framework. In this case study, we focus on applying the framework as an analytic tool, highlighting how the framework can be used to gain insights on potential changes or additions to the current design.

4.1 Case Study 1: FlowSteps

The design FlowSteps [17, 24] is developed as part of the I-PE project (see Introduction). FlowSteps consists of multiple, interactive mats that support open-ended play. The mats provide two colors of light output that react differently on the actions of the children. When no-one is playing with the mats, one mat randomly lights up in either red or blue. If a player steps on red, the mats provide options for a next move, while stepping on a blue mat lets players choose their own next move. Players can attach meaning to the interaction possibilities and position of the mats and create rules and games together. A prototype of FlowSteps was built, consisting of six interactive mats, and evaluated with twenty children (see also [24]).

Analysis. The three levels of the framework are represented in the FlowSteps in the following way. In terms of the Context of Play level, FlowSteps has the intention to stimulate physical movement and playing together. It is designed to support open-ended play. The potential target group consists of children aged 6-8 years old. At the Micro level we recognize the designed elements of the FlowSteps which are the *objects*: the six mats, and the *interaction opportunities*: a pressure sensor as input and the two colors of light, red and blue, as output. Interaction rules programmed in the mats determine which lights are active and how the mat responds to pressure or signals of the other mats. The Macro level includes the *dynamic behavior* and *emergent properties*. For the FlowSteps, these components are not fully incorporated. The emergent behavior that arises during play mainly involves the players. Concerning *development of meaning*, FlowSteps leaves room for players to interpret the various interaction opportunities that are part of the Micro level and attach their own meaning to them. Furthermore, the design is developed to support the total experience of interaction through the three *stages of play*: invitation, exploration and immersion. For instance, FlowSteps incorporates an active state in the invitation stage, lighting up one mat in either red or blue to attract players to start interacting with the mats. Moreover, design decisions such as the flexibility to move the mats around support the exploration stage, while the two different colors lead to different playful *experiences* in the immersion stage as challenge and competition.

When looking at the relations between the levels and its elements of the framework, the open-endedness of the FlowSteps at the Micro level leads to diverse forms of game play at the Macro level. Design decisions made at the Micro level

clearly influences player's behaviors and experiences at the Macro level. For instance, the scarcity of the blue light inspires some children to wait for the blue light to appear. The observations also show relations between experience and development of meaning. Different play intentions result in different meanings of the interaction opportunities. Some children focus on physically active games mostly related to competition: trying to move as fast as possible in order to catch the light and to beat the other player. Other children enjoy slower, tactic game play with the intention to discover how the objects exactly work. They consider the lights as interesting actions that need further investigation.

4.2 Case Study 2: Morel

Morel is a play object designed to "facilitate the emergence of new forms of outdoor physical play" [11]. Kenji Iguchi of Keio University in Japan developed the Morel. The Morels are cylindrical shaped objects approximately the size of a football that can sense the presence of another Morel by wireless communication. If two Morels are in range, sound feedback is given to the player. If players squeeze their own Morel, another Morel in range is 'charged'. Emission of a rising tone will provide feedback about the charge. If the charge is at maximum level, the Morel will launch itself.

Analysis. First, let's take a look at how the components of the framework can be recognized in the current design. Concerning the Context of Play, Morel is designed for outdoor physical play. Its aim is to create new and enriched play experiences by providing open forms of interactions. In this way people can define their own set of rules using the Morel. The Micro level includes the *objects* themselves. Besides that, the *interaction opportunities* are formed by the foam-like appearance of the Morel which makes it shock proof and squeezable. Also, the Morel provides sound feedback when it is in range of another Morel and can be launched by squeezing another Morel. In the Macro level, a collection of Morels alone does not show *dynamic behavior* or *emergent properties.* The player has to interact with the Morel to activate it. The Morels create an opening to define new communication lines between players using them, in this way creating opportunities for play. *Development of meaning* is an important factor in this. People playing with the Morel should incorporate it in their play by giving the provided interaction opportunities a meaning in play.

Secondly, by analyzing this case with the framework, we thought about several potential changes for the Morel and how this would affect the resulting play. These changes may not be relevant for the current design intention but can improve the design for other intentions. In the current design, the communication between objects is limited and will not lead to dynamic behavior without interaction with players. Implementing different interaction rules in the design, with more communication between the objects, can lead to behavior that arises from only the collection of Morels. For example, a larger collection of Morels can start making sounds, as if the system is excited. In this way the collection of Morels starts challenging players. Another option is incorporating adaptive behavior to support for instance the stages of play in order to support the experience of interaction over a longer period of time.

5 Discussion and Conclusion

The discussions of the case studies made us aware of the importance of apparent and less apparent links between levels and their content. Furthermore, we explored how the framework can be used to evaluate changes in the design. The framework combines multiple elements concerning play, interactions, experience and emergence. It illustrates these elements and their relationships. In this way it differs from, for example, the framework presented by [12] which focuses merely on the playful experiences, or the MDA model by [10], which is described in a rather linear setting. When developing DIPE the two models above need to be extended. With the presented framework we made a first step in analyzing relations between different elements involved in both system and play, and in the two levels, Micro and Macro.

In this paper we have presented a framework for decentralized interactive play environments. We have demonstrated the potential of the framework as a descriptive tool for analysis. In the first case study we noticed the elements of the framework helped us to understand and explain how the observed play emerged from the designed interactive objects. Furthermore, from both case studies we noticed the framework helped us to find opportunities for improvement. This is a first step in validation of the framework. The proposed framework can serve as a tool to analyze elements and their relations. In this way it serves as a contribution to other design researchers in this field. Moreover, it may also be relevant for other designers who want to design for emergence and experience. We will continue our work on this framework in future research by applying it to more cases.

By analyzing the case studies, we reflected mostly on the effect of design decisions and the relations between the different levels. But the framework also supported us in gaining first insights into the design process: how does designing a DIPE occur? It became clear that there is not one way of performing such a design process. It can be approached both top-down (from an experience) and bottom-up (from objects in a system). We believe developing DIPE means all elements pass view, to come to meaningful solutions. This design process will certainly be part of our future research.

Acknowledgements. This research is part of the Creative Industry Scientific Programme (CRISP), which is funded by Dutch government FES funding.

References

1. Bekker, T., Sturm, J., Eggen, B.: Designing playful interactions for social interaction and physical play. Personal and Ubiquitous Computing 14(5), 385–396 (2010)
2. Deen, M., Schouten, B.A.M.: Let's start playing games! How games can become more about playing and less about complying. In: Fun & Games (2010)
3. van Essen, H., Rijnbout, P., de Graaf, M.: A design approach to decentralized interactive environments. In: Nijholt, A., Reidsma, D., Hondorp, H. (eds.) INTETAIN 2009. LNICST, vol. 9, pp. 56–67. Springer, Heidelberg (2009)
4. Fromm, J.: Types and forms of emergence, Kassel University (2005), http://arxiv.org/abs/nlin.AO/0506028
5. Fromm, J.: On engineering and emergence, Kassel University (2006), http://arxiv.org/abs/nlin.AO/0601002
6. Gaver, W., Beaver, J., Benford, S.: Ambiguity as a resource for design. In: SIGCHI Conference on Human Factors in Computing Systems, pp. 233–240. ACM Press (2003)

7. Gordon, D.M.: The organization of work in social insect colonies. Nature 380, 121–124 (1996)
8. Hassenzahl, M.: User Experience and Experience Design. In: Soegaard, M., Dam, R.F. (eds.) Encyclopedia of Human-Computer Interaction. The Interaction Design Foundation, Aarhus (2011), http://www.interaction-design.org/encyclopedia/user_experience_and_experience_design.html
9. Huizinga, J.: Homo Ludens: A Study of the Play Element in Culture. Beacon Press, Boston (1955)
10. Hunicke, R., LeBlanc, M., Zubek, R.: MDA: A formal approach to game design and game research. In: AAAI Workshop on Challenges in Game. AAAI Press (2004)
11. Iguchi, K., Inakage, M.: Morel: Remotely launchable outdoor playthings. In: SIGCHI International Conference on Advances in Computer Entertainment technology. ACM Press (2006)
12. Korhonen, H., Montola, M., Arrasvuori, J.: Understanding playful experiences through digital games. In: 4th International Conference on Designing Pleasurable Products and Interfaces, pp. 274–285 (2009)
13. Lockton, D., Harrison, D., Stanton, N.: Design with intent: Persuasive technology in a wider context. In: Oinas-Kukkonen, H., Hasle, P., Harjumaa, M., Segerståhl, K., Øhrstrøm, P. (eds.) PERSUASIVE 2008. LNCS, vol. 5033, pp. 274–278. Springer, Heidelberg (2008)
14. Norman, D.: The Design of Everyday Things. Basic Books, New York (1990)
15. Resnick, M.: Turtles, Termites and Traffic Jams: Explorations in Massively Parallel Microworlds. MIT Press, Cambridge (1994)
16. Reynolds, C.W.: Flocks, herds, and schools: a distributed behavioral model. In: 14th Annual Conference on Computer Graphics and Interactive Techniques, pp. 25–34. ACM Press (1987)
17. Rijnbout, P., de Valk, L., de Graaf, M., Bekker, T., Schouten, B., Eggen, B.: i-PE: A decentralized approach for designing adaptive and persuasive intelligent play environments. In: Wichert, R., Van Laerhoven, K., Gelissen, J. (eds.) AmI 2011. CCIS, vol. 277, pp. 238–244. Springer, Heidelberg (2012)
18. Rozendaal, M., Vermeeren, A., Bekker, T., de Ridder, H.: A research framework for playful persuasion based on psychological needs and bodily interaction. In: Salah, A.A., Lepri, B. (eds.) HBU 2011. LNCS, vol. 7065, pp. 116–123. Springer, Heidelberg (2011)
19. Sengers, P., Gaver, B.: Staying open to interpretation: engaging multiple meanings in design and evaluation. In: 6th Conference on Designing Interactive Systems, pp. 99–108. ACM Press (2006)
20. Seitinger, S., Sylvan, E., Zuckerman, O., Popovic, M., Zuckerman, O.: A new playground experience: Going digital? In: Ext. Abstracts CHI 2006, pp. 303–308. ACM Press (2006)
21. Sheldon, K.M., Kasser, T., Elliot, A.J., Kim, Y.: What is satisfying about satisfying events? Testing 10 candidate psychological needs. Journal of Personality and Social Psychology 80(2), 325–339 (2001)
22. Suchman, L.: Plans and situated actions. Cambridge University Press, Cambridge (1987)
23. Tetteroo, D., Reidsma, D., van Dijk, B., Nijholt, A.: Design of an interactive playground based on traditional children's play. In: Camurri, A., Costa, C. (eds.) INTETAIN 2011. LNICST, vol. 78, pp. 129–138. Springer, Heidelberg (2012)
24. de Valk, L., Rijnbout, P., Bekker, T., Eggen, B., de Graaf, M., Schouten, B.: Designing for playful experiences in open-ended intelligent play environments. In: IADIS International Conference Games and Entertainment Technologies, pp. 3–10 (2012)
25. Zivanovic, A.: The development of a cybernetic sculptor: Edward Ihnatowicz and the Senster. In: 5th Conference on Creativity & Cognition, pp. 102–108. ACM Press (2005)

MashtaCycle: On-Stage Improvised Audio Collage by Content-Based Similarity and Gesture Recognition

Christian Frisson[2,⋆], Gauthier Keyaerts[1], Fabien Grisard[2,3], Stéphane Dupont[2], Thierry Ravet[2], François Zajéga[2], Laura Colmenares Guerra[2], Todor Todoroff[2], and Thierry Dutoit[2]

[1] aka Very Mash'ta and the Aktivist, artist residing in Brussels, Belgium
http://www.mashtacycle.be
[2] University of Mons (UMONS), numediart Institute
Boulevard Dolez 31 B-7000 Mons Belgium
http://www.numediart.org
[3] ACROE / Institut Phelma, Grenoble, France
christian.frisson@umons.ac.be

Abstract. In this paper we present the outline of a performance in-progress. It brings together the skilled musical practices from Belgian audio collagist Gauthier Keyaerts aka *Very Mash'ta*; and the realtime, content-based audio browsing capabilities of the *AudioCycle* and *LoopJam* applications developed by the remaining authors. The tool derived from *AudioCycle* named *MashtaCycle* aids the preparation of collections of stem audio loops before performances by extracting content-based features (for instance timbre) used for the positioning of these sounds on a 2D visual map. The tool becomes an embodied on-stage instrument, based on a user interface which uses a depth-sensing camera, and augmented with the public projection of the 2D map. The camera tracks the position of the artist within the sensing area to trigger sounds similarly to the *LoopJam* installation. It also senses gestures from the performer interpreted with the *Full Body Interaction (FUBI)* framework, allowing to apply sound effects based on bodily movements. *MashtaCycle* blurs the boundary between performance and preparation, navigation and improvisation, installations and concerts.

Keywords: Human-music interaction, audio collage, content-based similarity, gesture recognition, depth cameras, digital audio effects.

1 Introduction

Since the advent of affordable signal sensing and processing, ubiquitous and social networks, massive crowd-sourced multimedia datasets are being enriched everyday. These technologies allow audio artists to easily create their sounds or source these elsewhere, digitally, from the "ocean of sounds" [19].

⋆ Corresponding author.

M. Mancas et al. (Eds.): INTETAIN 2013, LNICST 124, pp. 114–123, 2013.

Fig. 1. Picture of an early version of stage setup at the numediart Institute where Gauthier Keyaerts interacts with *MashtaCycle* by his position tracked by a Kinect sensor. In the top left corner of the screen his segmented body as sensed by the Kinect through *OpenNI/NiTE* appears in the *FUBI* view. In the down left corner, visuals designed by François Zajéga are played back, acting as visual score. In the right column, a collection of sounds is visualized and rendered with the *MediaCycle* framework. Picture courtesy of Laura Colmenares Guerra (http://www.ulara.org).

Western digital music often relies on scores to transcribe and describe musical pieces. Here we consider sound samples as vocabulary, as do musical genres such as hip-hop, DJ'ing, electro-acoustic music. The map of sounds visualized in the *MashtaCycle* instrument becomes the score.

However, the emphasized musical expression in terms of sound generation offered by computer music suffers from an important drawback: the control of the sound generation. The NIME and ICMC conference communities, for instance, have been focusing on addressing this issue since decades [1]. William Brent proposes a fully opensource pipeline for content-based audio browsing through free-form gestural control [4] that seems suitable for prototyping, while Viglensoni digs deeper into the technologies offered for gestural control and sound synthesis [12]. We aim at offering a musical instrument that can complement the advantages offered by tangible and free-form interfaces [7].

In section 2, we describe the architecture of *MashtaCycle*. In section 3, we provide an overview of our accomplishments through this project and open with perspectives towards improvements, not without acknowledging (in section 3) all the people that made this project feasible.

2 Architecture

Figure 2 illustrates the architecture of the *MashtaCycle* system: sound files are first imported to be organized by content-based similarity as described in subsection 2.1, so as to be navigable in a 2D-space. To make the audio browsing more performative, simple gesture recognition techniques are made use of as explained in subsection 2.2. Mappings are drawn from gestures not only to sound rendering cues and effects (subsection 2.3), but also to generative visuals (subsection 2.4). The prototyping method that helped to refine the mappings up to an hardcoded system is discussed in subsection 2.5. Subsection 2.6 closes the architecture description by an opening: *MashtaCycle* can be hybridized from different settings, as an installation and as a performative tool.

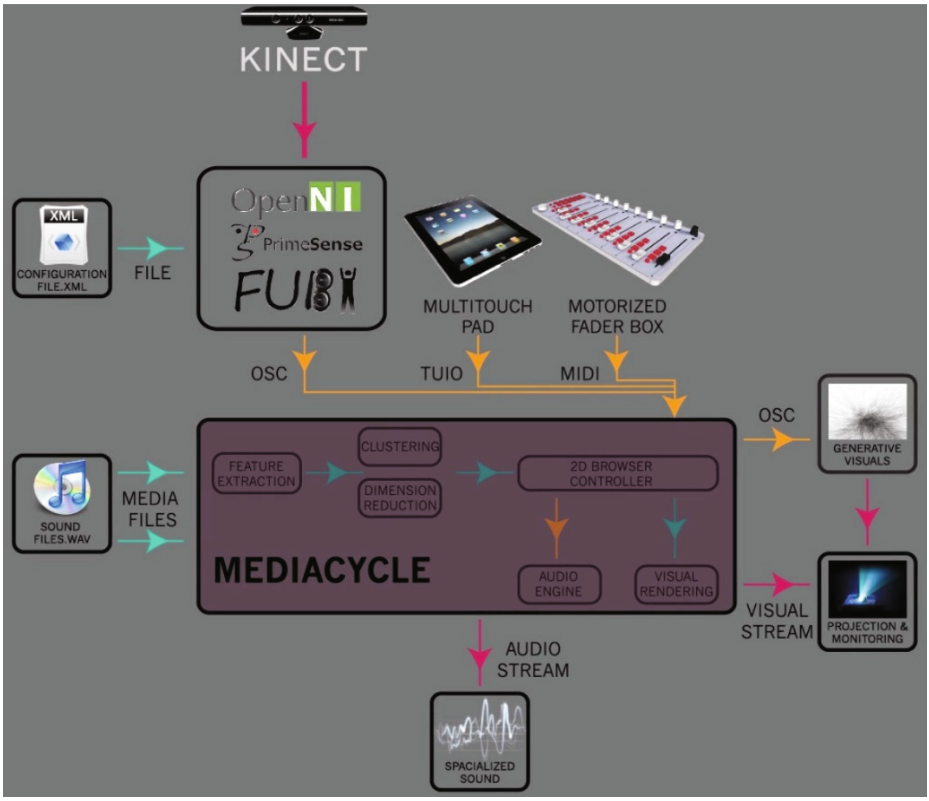


Fig. 2. Architecture of the current version of *MashtaCycle*

2.1 Content-Based Visual Audio Browsing

MashtaCycle is the next iteration of the *AudioCycle* series of applications, with the difference that it is tailor-made for an artist. *AudioCycle* is an application for organizing sounds by content-based similarity developed since 2008 in the numediart Institute of the University of Mons, described in [5], [9] then [8].

The notable difference since the last entertainment uses of *AudioCycle* is the choice of the algorithm that reduces the dimension of the audio feature space down to the 2D dimension of the visual space of the audio browser.

We would before use a simple algorithm named the "*Propeller*" that would, after a K-Means clustering step with a user-defined number k of clusters, create a visualization featuring a central symmetry, with k propels evenly-distributed angularly by the position of the centroid of their cluster, each node of each cluster positioned accordingly to its Euclidean distance to the centroid in the feature space. At the time it allowed to distribute sounds from the same instrument in a collection constituted of monophonic instrument loops in easily distinguishable areas, similarly to musicians in a rock band or a symphonic orchestra, as elaborated in [8]. This visualization was affected by several flaws:

- sounds weren't always properly clustered by instrument (or timbre), what is highly dependent on the number of clusters desired by the user versus the number of actual classes in the sound database;
- sounds were often occluding each others;
- an artifact inherent to its algorithm often made the visualization look more like the head of a string trimmer (gardening tool) than a propeller, sounds from the same cluster escaping the centroid in a curved line instead of being grouped in a propel.

After an evaluation of content-based visualizations by some of the authors [6], we changed our default visualization algorithm in favor of the Student-t distributed Stochastic Neighbor Embedding (t-SNE) algorithm and refined the choice of audio features used for the clustering and visualization, in short Mel-Frequency Cepstral Coefficients (MFCC) and their first- and second-order derivatives plus Spectral Flatness, using *Yaafe* [13]. For further explanation we direct the readers to the paper describing the evaluation [6]. It dramatically improves all the aforementioned issues (up to the accuracy unsupervised content-based organization can offer), notably by providing a layout where visual neighbors are separated by a repeatable distance, reducing overlap and easing the navigation though less closest node jitter.

In [11], the authors of the *CataRT* Max/MSP based environment for concatenative synthesis discuss other strategies for visual mapping.

2.2 Gesture Recognition

LoopJam [8], an installation made with *AudioCycle* controlled by a Kinect camera, allowed visitors to trigger sounds by their position in front of the projected map of sounds. At the time, it used Daniel Roggen's *QtKinectWrapper*

(`https://code.google.com/p/qtkinectwrapper/`), that we modified to send the 2D position in the plane of the floor of all users through OpenSoundControl (OSC). We aim at turning this installation into a musical instrument for a single user. To do so, we needed to offer more control on the sound rendering than just looped playback activation of the closest node, still through gestures that would be sensed with a Kinect camera.

To our knowledge, not many "plug and play" solutions are available for gesture recognition using the Kinect. The *Kinect SDK* from Microsoft may provide gesture recognition methods, but since it is available only on Microsoft platforms, it was instantly discarded for our use. To name a few, *Kinectar* (`http://ethnotekh.com/software/kinectar/`) provides heuristics-based gesture recognition (variation of distances between joints), it has the advantage of being designed by an artist, Chris Vik, for himself, what goes beyond the growing Kinect hacks, but it relies on many dependencies including closed-source ones. *XKin* was nominated for the Open Source Software Competition of ACM Multimedia 2012 [14]. It uses Hidden Markov Models (HMM) for hand poses recognition. While it offers a fully opensource pipeline with *libfreenect* instead of PrimeSense's *OpenNI/NiTE* combination that prevents a "drag-and-drop" installation and distribution since *NiTE* is closed-source, it is for now restricted to hand gestures, rather than full body gestures.

We chose to fork the *Full Body Interaction Framework (FUBI)* [10] (`http://www.hcm-lab.de/fubi.html`), opensource, from the University of Augsburg, which supports four gesture categories: 1) static postures, 2) gestures with linear movement, 3) combination of postures and linear movement and 4) complex gestures. In addition, the framework enables to detect the number of fingers the users are showing in front of the sensor, but it requires users to keep a fixed distance to the sensor. We added an *OpenSoundControl (OSC)* bridge to *FUBI* so as to directly communicate with our *MediaCycle* application. For that purpose we used the lightweight *OscPkt* library (`http://gruntthepeon.free.fr/oscpkt/`).

In the last chapter of his book [15], Dan Saffer catalogs free-form gestures and movements for the design of gestural interfaces and provides insight on references diving further into the topic, in the field of Human-Computer Interaction. Much research have been performed on musical gestures [21], borderline between art, science and technology.

Sound painting, originated by Walter Thompson in 1974 [17] and later popularized by John Zorn (notably his *Cobra* series), allows a conductor to have an orchestra of musicians improvise music by communicating through a sign language featuring hundreds to thousands of signs, in categories such as "who" (instrument or musicians), "what" (musical content and processing), "how" and "when". Figure 3 illustrates how sound painting inspired us through the design process.

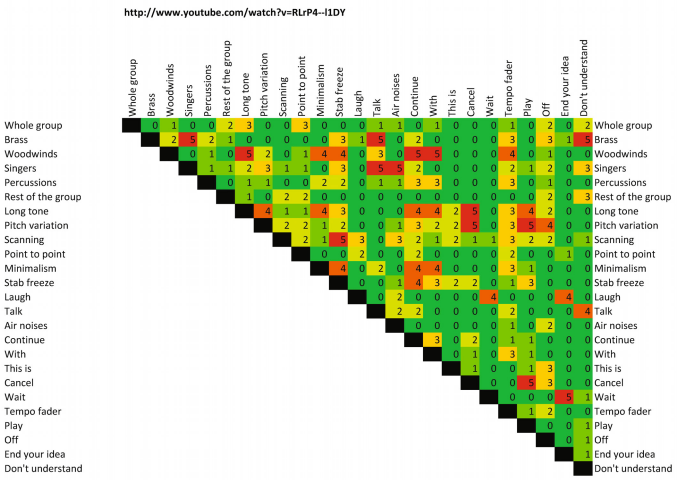


Fig. 3. Subjective evaluation by one of the authors of a subset of the Sound Painting gestures explained in Thomas Claus' video (http://www.youtube.com/watch?v=RLrP4--l1DY). A first selection step keeps the gestures that seemed to be detectable using a Kinect sensor. A second step rates these on a $[0; 5]$ integer Likert-like scale that estimates the probability of crosstalk between each possible couple, presented as a confusion matrix, with cue times of gestures in the video.

2.3 Sound Playback, Synthesis and Effects

A lot of research has been performed on the control of digital audio effects, either by direct gestural control or by adaptive control based on the content-based sound features [20]. Rather than adding support for usual audio effects plugin software development toolkits such as *VST* or *AudioUnits* to iterate over the *LoopJam* [8] installation, we created a new audio engine based on the *Synthesis Toolkit (STK)* [16] (`https://ccrma.stanford.edu/software/stk/`) itself built upon the *RtAudio* backend and providing classes for common audio effects (chorus, delay, reverb, etc...), since this solution is suitable for fast prototyping and avoids potential issues generated by third party plugins.

2.4 Visual Rendering

The monitoring views offered by the *MediaCycle* browser and the *FUBI* gesture recognition tool are mandatory for using the system as a musical instrument, but these don't help the system to qualify properly as a piece of interactive arts which would offer a more poetic visualization. A sound-dependent visual rendering is being prototyped by Belgian artist François Zajéga (`http://www.frankiezafe.org`), In Figure 1 extracted from the demo video of the *MashtaCycle* project, visuals created by François Zajéga were played back

in a loop as a movie file so as to influence the audio collage improvisation by Gauthier Keyaerts, similarly to a graphical score.

A visual rendering that reacts to the sound content improvised by Gauthier Keyaerts is in progress. To do so, audio features will be extracted from the sound rendering after the effects processing. While the *MediaCycle* framework doesn't yet support realtime stream analysis, an intermediate prototyping solution is offered by William Brent's *timbreID* objects for *PureData* [4].

2.5 Mapping, Prototyping

Building upon earlier works with *MediaCycle* [9], we decided to apply the same fast user interface prototyping method using the *PureData* environment. Lots of mapping-related steps affect the architecture of *MashtaCycle*: 1) audio features to 2D space, 2), gestures to audio effects, 3) post-processed audio rendering to visuals. For now one-to-one mappings are being used, except for step 1). Figure 4 illustrates an example of mapping.

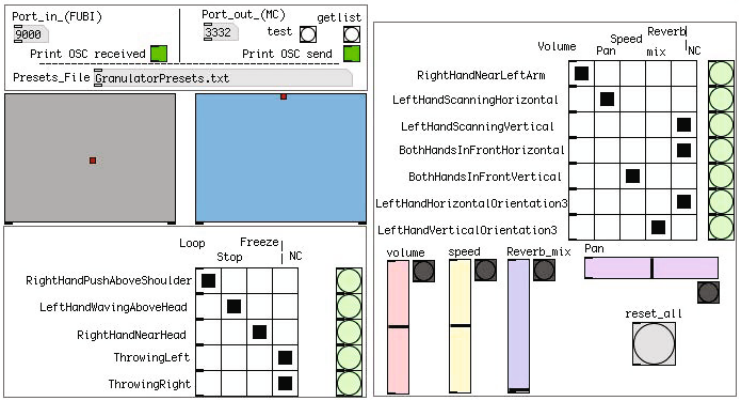


Fig. 4. *PureData* patch for the 1-to-1 mapping of gestures to sound effects

After the prototyping phase, mappings have been hard-coded in C++ temporarily inside the *FUBI* fork so as to reduce the number of applications to launch and monitor (what we do with *Lingon* (`http://sf.net/p/lingon/`), a Mac OSX GUI for *launchd*), and the overhead of the *PureData* application.

2.6 Hardware and Setup Requirements

The architecture of this project has been sketched over the last subsections.

Since the depth sensing camera emits a grid of infrared beams and receives its reflections from the scene, stage lights have to be dimmed properly or equipped with filters (infrared, or blue colored) so as not to interfere with the sensing.

Many loudspeaker configurations may be used. For a good sound immersion we would use at least a cost-effective quadraphony (two loudspeakers surrounding the stage, another two behind the audience).

Multiple screen configurations are possible, along what the artist and the audience may see, with the following views activated or not: 1) *FUBI* gesture recognition, 2) *MediaCycle* visual audio browsing, 3) (audio-dependent) visuals. Different setups may be presented so as to suit venues, for instance:

1. demo: the actual setup illustrated in Figure 1
2. concert: the audience only sees the audio-dependent visual rendering on a large screen projection, the artist monitors the *FUBI* gesture recognition and the *MediaCycle* audio browsing views from a second projected screen located in one side of the stage

An innovative feature of this work is that after a concert, the musical instrument can be turned into an installation that the audience can visit to better understand it while chilling out and discussing with the crew of this project.

3 Conclusion, Perspectives

We designed a first prototype of a musical instrument tailored for an artist wanting to create improvised audio collages, using gesture recognition through a depth sensing camera and content-based visual audio browsing.

Our fork of the *Full Body Interaction Framework (FUBI)* with *OpenSoundControl (OSC)* output and new gestures is available on `http://github.com/ChristianFrisson/FUBIOSC`, we hope most of it will be integrated back to the main distribution of *FUBI*, and we plan to integrate it as a *flext* object for *PureData* and *Max/MSP* into the *DeviceCycle* [9] distribution for fast user interface prototyping.

While this work satisfies the artist it is designed for, it still needs a quantitative evaluation. We tweaked the current gesture recognition technique up to the point that the artist using the tool feels that gestures are properly recognized most of the times. We may measure the repeatability in detecting a certain number of same instances of a given gesture, but we plan first to change the gesture recognition method in favor of another one from the field of machine learning such as Hidden Markov Models (HMM) [3,18], Dynamic Time Warping (DTW) [2]. The current method is not morphologically-independent and requires fiddling with numbers on an XML file, what we believe is not suitable for most potential users of this project, first and foremost the artist for whom it is designed, when new gestures are requested. Gesture design through recording is in progress.

Some of the authors, which are musicians of various levels of training, would prefer contact-based and/or tangible interaction over the free-form interaction offered in this project [9]. Gestural interfaces with haptic feedback are being prototyped, similarly to [22].

Acknowledgements. The majority of the authors are funded by a grant from the Walloon Region of Belgium (N°716631) that allowed to found the numediart Institute.

Gauthier Keyaerts obtained a grant from the *Digital Arts section* of *Fédération Wallonie-Bruxelles* (`http://www.arts-numeriques.culture.be`) that complements the financial coverage of the design phase of this project. We would like to thank the committee for having nominated our project and Anne Huybrechts for having advised us carefully while submitting our proposal.

Fabien Grisard obtained an ExploRA'sup grant from the French Region of Rhône-Alpes that covers parts of his visiting at the numediart Institute for his MSc graduation internship, focusing on the gesture recognition/design.

We are redeemable to Felix Kistler, from the University of Augsburg, Germany, for creating *FUBI*, releasing it open source (`http://sf.net/p/fubi`), and answering our related questions spot-on and almost in realtime.

We thank our colleagues Nicolas d'Alessandro and Joëlle Tilmanne for their expert insight on more solid gesture recognition methods, we hope to work with them towards an improved version of this project.

We thank the artistic associations and individuals that have been curating iterations of prototypes this project builds itself upon: Philippe Franck and the team of *Transcultures* (`http://www.transcultures.be`) plus the team of movie theater *Galeries* (`http://www.galeries.be`) hosting *LoopJam* for the *Citysonic* 2012 festival in Brussels (`http://www.citysonic.be`); the organizers of the Kikk festival (`http://www.kikk.be`), who welcomed us for the 2012 edition, Stephan Dunkelman, Marianne Binard and the team of *Halolalune* (`http://www.halolalune.be`) plus Wim Verhulst and the team of the *Musical Instruments Museum (MIM)* of Brussels (`http://www.mim.be`) for hosting *LoopJam* through the annual week of *La Semaine du Son* in late January 2013 (`http://www.lasemaineduson.be`).

References

1. Bekkedal, E.: Music kinection: Musical sound and motion in interactive systems. Master's thesis, Department of Musicology, University of Oslo (2012)
2. Bettens, F., Todoroff, T.: Real-time DTW-based gesture recognition external object for max/msp and puredata. In: Proceedings of the 6th Sound and Music Computing Conference, Porto, Portugal, July 23-25 (2009)
3. Bevilacqua, F., Zamborlin, B., Sypniewski, A., Schnell, N., Guédy, F., Rasamimanana, N.: Continuous realtime gesture following and recognition. In: Kopp, S., Wachsmuth, I. (eds.) GW 2009. LNCS, vol. 5934, pp. 73–84. Springer, Heidelberg (2010)
4. Brent, W.: Physical navigation of virtual timbre spaces with timbreID and DILib. In: Proceedings of the 18th International Conference on Auditory Display, Atlanta, GA, USA, June 18-21 (2012)
5. Dupont, S., Frisson, C., Siebert, X., Tardieu, D.: Browsing sound and music libraries by similarity. In: 128th Audio Engineering Society (AES) Convention, London, UK, May 22-25 (2010)

6. Dupont, S., Ravet, T., Picard-Limpens, C., Frisson, C.: Nonlinear dimensionality reduction approaches applied to music and textural sounds. In: IEEE International Conference on Multimedia and Expo (ICME), San Jose, USA, July 15-19 (2013)
7. Frisson, C.: Designing tangible/free-form applications for navigation in audio/visual collections (by content-based similarity). In: Graduate Student Consortium of the ACM Tangible, Embedded and Embodied Interaction Conference (TEI 2013), Barcelona, Spain, February 10-13 (2013)
8. Frisson, C., Dupont, S., Leroy, J., Moinet, A., Ravet, T., Siebert, X., Dutoit, T.: LoopJam: Turning the dance floor into a collaborative instrumental map. In: Proceedings of the New Interfaces for Musical Expression (NIME), Ann Arbor, Michigan, USA, May 21-23 (2012)
9. Frisson, C., Dupont, S., Siebert, X., Tardieu, D., Dutoit, T., Macq, B.: DeviceCycle: Rapid and reusable prototyping of gestural interfaces, applied to audio browsing by similarity. In: Proceedings of the New Interfaces for Musical Expression++ (NIME++), Sydney, Australia, June 15-18 (2010)
10. Kistler, F., Sollfrank, D., Bee, N., André, E.: Full body gestures enhancing a game book for interactive story telling. In: International Conference on Interactive Digital Storytelling, Proceedings of ICIDS 2011 (2011)
11. Lallemand, I., Schwartz, D.: Interaction-optimized sound database representation. In: Proceedings of the 14th International Conference on Digital Audio Effects (DAFx 2011), Paris, France, September 19-23 (2011)
12. Martin, A.G.V.: Touchless gestural control of concatenative sound synthesis. Master's thesis, McGill University, Montreal, Canada (2011)
13. Mathieu, B., Essid, S., Fillon, T., Prado, J., Richard, G.: Yaafe, an easy to use and efficient audio feature extraction software. In: Proceedings of the 11th ISMIR Conference, Utrecht, Netherlands (2010)
14. Pedersoli, F., Adami, N., Benini, S., Leonardi, R.: XKin: eXtendable hand pose and gesture recognition library for kinect. In: Proceedings of the 20th ACM International Conference on Multimedia (MM 2012), pp. 1465–1468. ACM, New York (2012)
15. Saffer, D.: Designing Gestural Interfaces. O'Reilly Media, Inc. (2009)
16. Scavone, G.P., Cook, P.R.: RtMidi, RtAudio, and a Synthesis Toolkit (STK) update. In: of the International Computer Music Conference (2005)
17. Thompson, W.: Soundpainting: The Art of Live Composition. In: Walter Thompson, 2006. with instructional DVD (2006)
18. Tilmanne, J.: Data-driven Stylistic Humanlike Walk Synthesis. PhD thesis, University of Mons (2013)
19. Toop, D.: Ocean Of Sound: Aether Talk, Ambient Sounds and Imaginary Worlds. Serpent's tAIL (1995)
20. Verfaille, V., Wanderley, M.M., Depalle, P.: Mapping strategies for gestural and adaptive control of digital audio effects. Journal of New Music Research 35(1), 71–93 (2006)
21. Wanderley, M., Battier, M. (eds.): Trends In Gestural Control of Music. Ircam - Centre Pompidou (2000)
22. Zadel, M.: Graphical Performance Software in Contexts: Explorations with Different Strokes. PhD thesis, McGill University, Montreal, Quebec, Canada (2012)

DanSync: A Platform to Study Entrainment and Joint-Action during Spontaneous Dance in the Context of a Social Music Game

Michiel Demey, Chris Muller, and Marc Leman

IPEM, Dept. of Musicology
Ghent University, Belgium
{michiel.demey,chris.muller,marc.leman}@ugent.be

Abstract. This paper presents a social music game, named DanSync, as a platform to study joint-action. This game context proves to be an effective manner to study spontaneous dance of players in a laboratory setting. Because of the gameplay participants are engaged in dancing to music with a strong motivation. Performance of dance synchronization to music is studied throughout the gameplay. Joint-action in a dyad is quantified in terms of correlation and phase-locking. Furthermore, entrainment and social bonding in small groups is studied by introducing perturbations in the music stimulus.

Keywords: Entrainment, gaming, music.

1 Introduction

A large body or research has explored entrainment, interpersonal synchronization and joint action in relation to each other (a number of comprehensive reviews include e.g. [1] for a dynamical systems approach; [2] for a review from an action simulation and motor resonance point of view; [3] for a review focused on joint action and social connection). A lot of this work has covered the perceptual and motoric basis for both interlimb and interpersonal coordination (see also [4]).

A number of main findings in previous work have been that people can (and often will unintentionally) synchronize their movements to movements that they see or hear others doing and the emerging social and physical factors that modulate the likelihood of interpersonal coordination.

There appears to be a clear link between the amount of (coordination dynamics-) information shared and frequency detuning; the larger the difference in eigenfrequencies between two systems, the stronger the informational link has to be to enable successful synchronization (e.g. [5]) as described in [1].

Most of this work, however, is done in controlled laboratory settings, and as such does not necessarily correspond to real-world behavior. Furthermore, most work thus far studies only behavior in dyads or solo, and does not make a comparison of solo, dyad and group in the same environment. Also, most of this work focuses on one typical movement, not on whole body movement.

M. Mancas et al. (Eds.): INTETAIN 2013, LNICST 124, pp. 124–135, 2013.

Several studies exist on the nature of social entrainment in cases where a group of people perform a common task and music is used to organize and coordinate the effortful activities [6]. There is evidence which suggests that action coordination is greatly dependent on whether the actors are trying to achieve a common goal; interacting in a competitive task is less likely to result in temporal coordination than interacting in a cooperative task [7]. Furthermore, social competence can predict coordination success in dyads [8].

Most studies are done with controlled temporal stimuli; metronome clicks etc, even though music provides a rich tool for studying these phenomena, its more ecological, its engaging, its multi-sensory (see also [9] for a multi-sensory experiment), and music can be competitive or cooperative. Also, music tightly connects these phenomena as it involves people synchronizing behavior to one another, people becoming entrained to both each others actions as well as to indirect results of actions (e.g. sounds in music), and it involves people coordinating their actions to perform together. In the past years, research in this field has focused on both controlled situations such as people tapping their fingers to an isochronous stimulus [10] and sporadically on more ecologically valid situations in which people for instance dance together in a room (e.g. [11]). All studies focus on one very specific movement, be it rocking motion [12,13], finger tapping [10], pendulum swinging [14]. Most of these behaviors are rather isolated, and do not always represent a real human interaction through full body movement, like e.g. dancing together or music production [15].

There have not been many studies that investigate the role of joint and individual goals in a social musical context. In this paper we present a novel way to empirically explore synchronization and entrainment aspects of dance in a social context. We developed a game in which 10 participants are motivated to dance in a spontaneous manner and to actively interact with other participants. This game is played in a setting resembling a club (including the lighting and a sound-system) but nevertheless is situated in our lab enabling a motion capture recording of all players present. In this study we explore the feasibility of the setup and analysis tools to provide further insights to the following aspects of joint-action and entrainment:

(1) Are people motivated in a social game context and dance in a spontaneous manner?

(2) Do people improve their synchronization over time and under of the influence of others?

(3) Is it possible to quantify social bonding and to optimize the conditions to maximize this effect?

This paper presents first the context of the experiment elaborating on the technical setup and the gameplay. In a following section the data analysis is explained followed by the results from this analysis. We conclude with a discussion and outlook.

2 Setup

The results presented in this paper are based on a social music game called DanSync which is played by a group of 10 participants. This game is developed through various iterations at our institute [16,17,18,19]. In what follows the technical setup is outlined followed by a description of the gameplay.

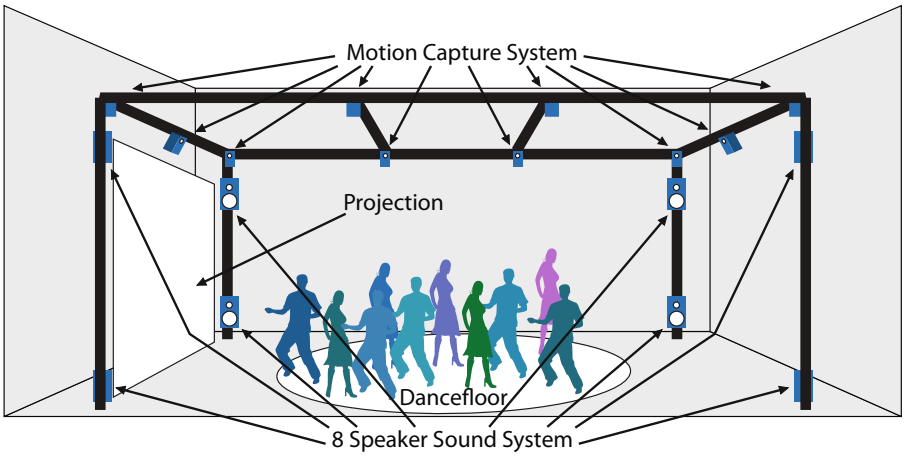


Fig. 1. A schematic overview of the different components of the setup used

2.1 Technical Setup

Various technologies are used in DanSync which are setup in a Truss structure of 5 m×5 m and 3 m of height as shown in figure 1. The dance movement of the players is measured with a Motion Capture (MoCap) system of Optitrack, Natural Point. With this MoCap system the 3D position of IR-reflective markers on a helmet worn on the head of 10 players is captured in real-time at a sampling frequency of 100 Hz. This data is sent via UDP to the central game computer running the game logics developed in Max/MSP and Jitter.

During gameplay the BPM of the movement of each participant is determined in real-time through the use of an FFT algorithm as described in detail in section 3.1. This BPM is compared to the BPM of the song played and a score is derived. All participants wear an iPod touch around their forearm in such a way that it is comfortable to wear and is easily visible during gameplay. This iPod touch receives real-time feedback from the central game computer via WiFi. In this way each participant has an individual feedback on their dance performance during the game. The central game computer plays back the music over an octafonic sound system and also projects visuals on the front wall of the dance arena. These visuals are used to instruct the participants on the gameplay before the game starts and presents an overview of the scores for all players after each round.

2.2 Gameplay

The DanSync game consists of 5 rounds schematically presented in figure 2.

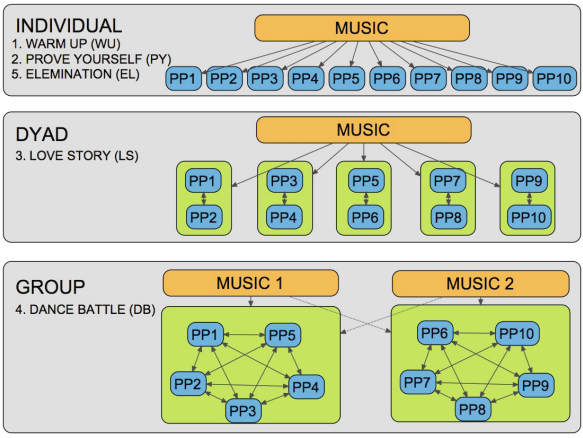


Fig. 2. A schematic overview of the different interaction models applied to the gameplay

1. WARM UP (WU): the first round of the game after receiving instructions on the gameplay. The goal is to synchronize dance movement to the tempo of the music but the scores obtained do not account for in the final score.
2. PROVE YOURSELF (PY): the task is the same as in the WU round, however, the obtained score is taken into account for the final score of the game.
3. LOVE STORY (LS): couples are formed by the game algorithm and the participants have to find their team mate based on the interpersonal distance visualized in real-time on their iPod. The task is to dance as close as possible as a couple to the tempo of the music.
4. DANCE BATTLE (DB): two teams of 5 players each are formed by the game algorithm and visualized as a blue and a red team on the projection on the front wall of the dance arena. 2 songs corresponding to each team are heard. On the iPod the team color is indicated together with a volume slider representing the loudness of the song at that time. Every 25 s one song is faded out together with a fade in of the song of the other team. The task is to dance to the music of your team and to keep on doing so even when the music of the other team is played.
5. ELIMINATION (EL): in this final round bad players are excluded one by one resulting in a single player at the end of the round who wins DanSync.

Each round takes approximately 3 minutes to play resulting in a 15 minute duration of the complete game. The songs played during the rounds were:
WU Don't Stop - The Subs (House, 128 BPM), PY Lonely Boy - The Black Keys (Rock, 84 BPM), LS Jolie Coquine - Caravan Palace (Charleston, 125 BPM), DB1 Acid Phase - Emmanuel Top (Techno, 134 BPM), DB2 Amphetamine - Drax LTD II (Techno, 142 BPM), EL Yeah 3x - Chris Brown (Pop, 130 BPM).

3 Data Analysis

This section presents concepts to quantify entrainment in relation to the music stimulus and to the other players. The input signal is the vertical position of the head of each participant at a sampling rate of 100 Hz. From this vertical displacement the tempo is determined on the one hand and the correlation and phase-lock is quantified on the other. These two analysis paths are described in detail in the following sections.

3.1 Synchronization to the Music

To study the amount of synchronization of the dance movement of the participants to the tempo of the music the vertical position of the head is used as an input signal. This signal is filtered by a bandpass filter in the range of 0.5 - 4 Hz corresponding to 30 BPM to 240 BPM respectively. From this signal the BPM value can be calculated using an FFT where 4 s of data is analyzed with an overlap of 2 s. The data contained within the 4 s under inspection corresponds to 400 data samples. This data is zero-padded to a total length of 6000 samples to obtain a resolution of 1 BPM in the frequency spectrum. In this spectrum the highest peak is located and the corresponding BPM value is determined.

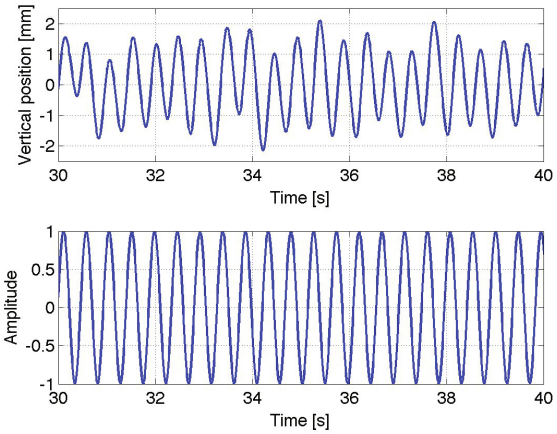


Fig. 3. Top: The vertical position of the head of a player after filtering. Bottom: A sinusoidal signal representing the tempo of the music.

To study the synchronization of the movement to the tempo of the music in more detail we make use of the cross-correlation of the filtered movement signal with a sinusoidal signal with the same frequency as the tempo of the music. Both signals are shown in figure 3.

The cross-correlation uses datawindows of 3 s (300 samples) with an overlap of 25% and lag values of 0.01 s (1 sample) in the range of ± 2 s. In this way

one can obtain a representation as shown in figure 4 where the x-axis represents time, the y-axis the lag and the color code the correlation value. The periodic structure in the vertical direction is due to the repetitive movement signal.

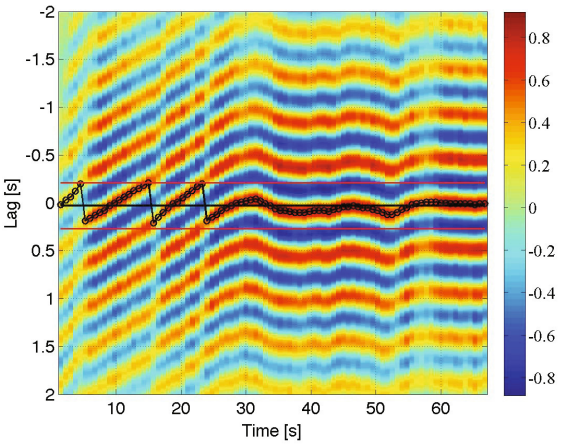


Fig. 4. The cross-correlation of headmovement with sinusoidal signal representing the tempo of the music. The points of maximal correlation are indicated by circles.

To further quantify this cross-correlation we make use of a peak-picking algorithm to locate the points of maximal correlation and their corresponding lag values (similar as in [20]). In order to obtain a more continuous path the maxima are located in an iterative process where the lag range is initiated at ± 2 s. Next the median of the lag values corresponding to the maximal correlation values is determined and a window of ± 0.5 times the BPM of the music is centered around this median defining the lag range in which the maxima are to be located in the next pass. The process is aborted when the lag range is unchanged since the last iteration and therefore an optimal result is obtained.

The peak-picking algorithm results in a set of maximal correlation values and their corresponding lags which are indicated with a circle in figure 4. The median is indicated with a horizontal black line and the window of ± 0.5 times the BPM of the music is indicated with the two horizontal red lines.

3.2 Interpersonal Coordination

The synchronization of individual players with other participants can be studied in the same way as described in the previous section. Here, the cross-correlation is calculated between the movement signals of both subjects under study. The peak-picking algorithm results in the maximal interpersonal correlation values and their corresponding lags.

4 Results

This section first presents the study of synchronization of dance movements to the music played and the interpersonal synchronization, followed by the effects of joint-action in a dyad and ends with an example of how the disturbance of an entrained system can be used to quantify social bonding.

All results presented are obtained in 4 test sessions where each time 10 players played the game DanSync. The last group of people played DanSync two times consecutively with a short break in between. The group of participants consisted of 10 male and 30 female subjects with an average age of $26,1 \pm 9,7$ years old.

4.1 Improvement of Synchronization to Music

To quantify the amount of synchronization of the players to the tempo of the music played we make use of the points of maximal correlation between the movement of all players and a sinusoidal signal representing the tempo of the music. When considering the 3 individual rounds of the game namely WU, PY and EL we can obtain 3 distributions for each time the game was played (5 times in total). These distributions are obtained by omitting the first 5 s to account for the time people need to start to synchronize. The data in the EL round is truncated to 45 s since this is the time when the first player is eliminated and the group decreases in number of players.

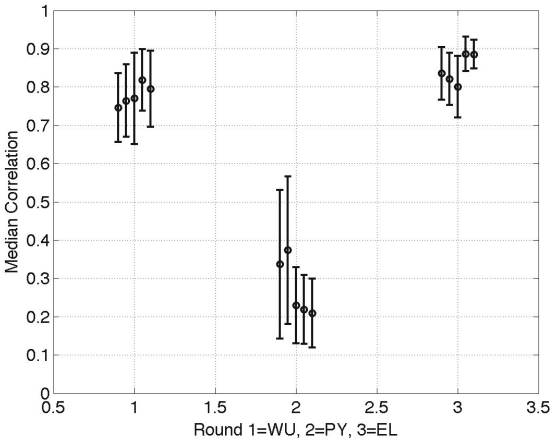


Fig. 5. The median of the correlation values for each group of participants for the 3 individual rounds WU, PY and EL. The errorbars represent the median absolute deviation.

Since the Lilliefors test rejects normality, we use the median instead of the mean value. As an error estimate we make use of the median absolute deviation or mad. The result is presented in figure 5.

In this figure we can clearly see similar correlation values for each game round for the different groups. The systematic difference between conditions is most likely due to the nature of the song. The Wilcoxon rank sum test rejects the hypothesis that the medians are the same between the first and last round making the high values in the last round significant. One can wonder if the high correlation values in the last song are due to the song or due to the motivating factor of the gameplay.

4.2 Improvement of Interpersonal Synchronization

To study the interpersonal coordination the correlations between the movements of the different subjects is calculated. For the individual rounds the correlation for each participant with all other players is calculated. In this way we can obtain 45 crosscorrelations for each unique pair within the 10 participants.

From the maximal correlation values obtained in each of the 45 crosscorrelations we take the median and represent them in a distribution. The normality was tested using a Lilliefors test and was accepted in all distributions. Next, the difference between distributions was tested using an ANOVA which resulted in a significant difference between all 3 distributions and a significant difference between the first and the last distribution. An example is shown in figure 6 for the first testsession.

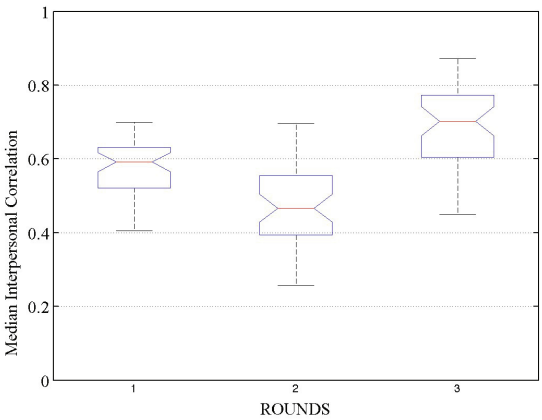


Fig. 6. ANOVA result of the distributions of the median of the maximal correlation values in each of the 45 possible crosscorrelations

From this analysis we can conclude that people dance in a more correlated manner with each other in the last round of the game. This is also the round where the highest correlation with the music was observed. One can raise the question wether an increased synchronization to a music stimulus induce social bonding or vice versa.

4.3 Joint-Action in a Pair (Dyad)

To study the joint-action in a dyad we make use of the data obtained in the LS round. The first 20 s of the data is omitted to account for the time the participants need to find their corresponding partner as assigned by the game logics. Using the median of the interpersonal correlation values and the mad of the corresponding lag values obtained through the peak-picking picking algorithm we can see a clear correlation as presented in figure 7. There is a clear correlation between the median correlation between players within a couple and the variation of their relative phase. In other words, the couples who have a clear phase-lock also correlate well or vice versa.

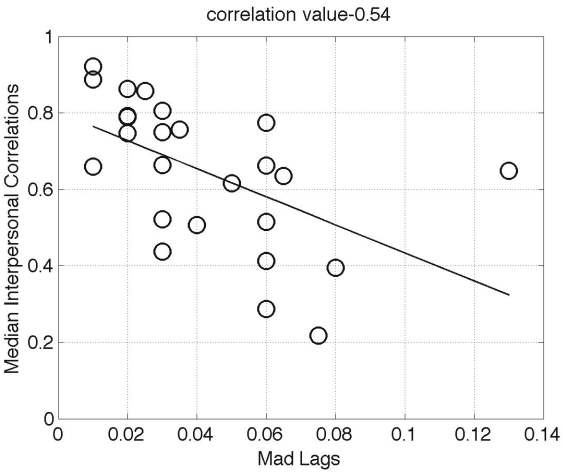


Fig. 7. Correlation between the median absolute deviation of the lags vs the median of the interpersonal correlation values for each dyad during the LS round

4.4 Joint-Action in a Group

To study joint-action in small groups of 5 participants the data in the DB round is analyzed. The BPM value of each participant is calculated for a timeframe of 4 s of data with an overlap of 2 s as described in section 3.1. The median and median absolute deviation of these BPM values is calculated for each timeframe under study and presented in figure 8 for a single group of participants. The red and blue datapoints correspond to the two teams. The horizontal straight lines correspond to the BPM value of the two songs. At the bottom of the figure the loudness of the songs played is shown where the lower value corresponds to silence and the top value is the nominal loudness.

Here we can see that the song of the blue team starts to play at the beginning of the round and both teams synchronize to that BPM. After 25 s the loudness of the song of the red team is at maximum and the song of the blue team is completely faded out. At this time both teams synchronize to the music of the

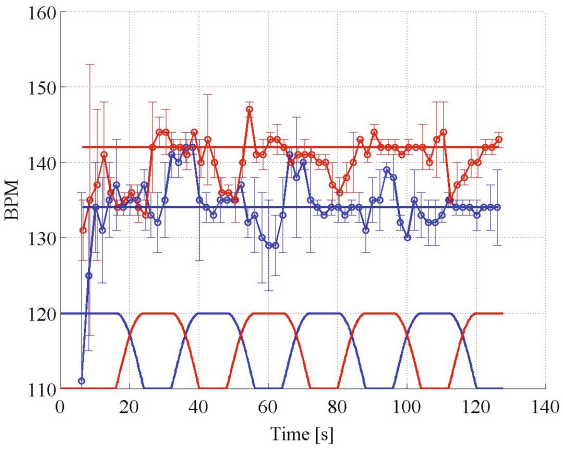


Fig. 8. An example of the performance during the dance battle round. A detailed explanation of the presented graphs is provided in the text.

red team while the goal of the game is to stay synchronized with the song of your own team. The same dynamics occur at 40 s. However, at the next cross-fade at 55 s the blue team manages to keep synchronized with its own song while the song of the red team sounds. This behavior is maintained throughout the remainder of the round with some perturbations in the synchronization.

The median BPM value of the group represents the synchronization to the musical tempo while the median absolute deviation resembles the amount of social bonding.

5 User Evaluation

After the game was played the 40 users where presented with a questionnaire probing their experience and their evaluation of the technology and gameplay. All participants found the game to be pleasant (82%) or very pleasant (18%). The majority of players found that dancing together was (very, 13%) motivating (78%) and the feedback presented on the projection and the iPod was experienced as (very, 25%) motivating (68%). Those players who rated their own performance as very good had high final score $(80,24 \pm 9,82)$, those who rated their own performance as bad or very bad resulted with a low final score $(65,00 \pm 11,43)$. A quarter of the participants found the motion capture helmet and the iPod as disturbing (25%) or very disturbing (0%). None of the participants found the projection disturbing. Most participants had danced better (88%) or much better (3%) due to the feedback on the iPod and the majority of the players found the added value of the iPod high (75%) or very high (10%). Almost all players found the feedback on the iPods sufficient (95%) and 35 participants found the feedback resembling their performance well (68%) or very

well (20%). All players found the explanation of the game clear (60%) or very clear (40%) and the goal of each round was also clear (72%) or very clear (28%) for all participants. All participants would like to play the game again.

6 Discussion and Outlook

In this paper we presented a novel way to study entrainment to a music stimulus in a social game context. The game context provides an ecologically valid setting where participants can dance to music in a spontaneous manner with full body movements. The objective quantification of the users dance movements is made possible through the analysis of the vertical displacement of the head. The gameplay was experienced by the users as motivating and fun and the technologies used provided a clear added value to the dance performances of the players.

Our preliminary data shows a clear link between narrow phase distributions in dance tempos and interpersonal correlation in terms of median dance tempos as well as phase locking. This is in line with earlier findings on oscillator variability and entrainment described in e.g. [1].

The setup and the analysis tools presented in this paper allow us to present the data on a comprehensible level of interpretation on various aspects of social entrainment namely: synchronization of dance movements to music with a focus on the effect of the song and the effect of the social and game context, interpersonal synchronization and adaptation, joint-action in a dyad and in a group condition.

Related to the questions raised in the introduction we can state that:
(1) Based on the questionnaires conducted people are motivated in a social game context and dance in a spontaneous manner.
(2) Playing our game appears to increase the level of synchronization over the course of the different levels; further studies can clarify the roles of social bonding and interpersonal synchronization.
(3) It is possible to quantify social bonding using the tempo of the dance movements and the associated median and absolute median deviation.

We are convinced that future studies can focus on each of these aspects separately using the tools developed in this paper to research social entrainment in more detail.

References

1. Schmidt, R., Richardson, M.: Dynamics of interpersonal coordination. In: Coordination: Neural, Behavioral and Social Dynamics, pp. 281–308 (2008)
2. Phillips-Silver, J., Keller, P.: Searching for roots of entrainment and joint action in early musical interactions. Frontiers in Human Neuroscience 6 (2012)
3. Marsh, K., Richardson, M., Schmidt, R.: Social connection through joint action and interpersonal coordination. Topics in Cognitive Science 1(2), 320–339 (2009)
4. Schmidt, R., Fitzpatrick, P., Caron, R., Mergeche, J.: Understanding social motor coordination. Human Movement Science 30(5), 834–845 (2011)

5. Kelso, J., Del Colle, J., Schöner, G.: Action-perception as a pattern formation process. In: Attention and Performance 13: Motor Representation and Control, pp. 139–169 (1990)
6. McNeill, W., McNeill, W.: Keeping together in time: Dance and drill in human history. Harvard University Press (1997)
7. Johnston, L., Miles, L., Richardson, M., Schmidt, R., Marsh, K., Yabar, Y.: Implicit mimicry and synchronization: Impact of and on liking. In: North American Meeting of the International Society for Ecological Psychology, University of Cincinnati, Cincinnati, OH (2006)
8. Schmidt, R., Christianson, N., Carello, C., Baron, R.: Effects of social and physical variables on between-person visual coordination. Ecological Psychology 6(3), 159–183 (1994)
9. Hove, M., Iversen, J., Zhang, A., Repp, B.: Synchronization with competing visual and auditory rhythms: Bouncing ball meets metronome. Psychological Research, 1–11 (2012)
10. Repp, B.: Sensorimotor synchronization: A review of the tapping literature. Psychonomic Bulletin & Review 12(6), 969–992 (2005)
11. Woolhouse, M., Tidhar, D.: Group dancing leads to increased person-perception. In: Proceedings of the 11th International Conference on Music Perception and Cognition (ICMPC 2011), Seattle, USA, pp. 605–608 (2010)
12. Richardson, M., Marsh, K., Isenhower, R., Goodman, J., Schmidt, R.: Rocking together: Dynamics of intentional and unintentional interpersonal coordination. Human Movement Science 26(6), 867–891 (2007)
13. Demos, A., Chaffin, R., Begosh, K., Daniels, J., Marsh, K.: Rocking to the beat: Effects of music and partner's movements on spontaneous interpersonal coordination. Journal of Experimental Psychology: General 141(1), 49–53 (2012)
14. Schmidt, R., O'Brien, B.: Evaluating the dynamics of unintended interpersonal coordination. Ecological Psychology 9(3), 189–206 (1997)
15. Clayton, M.: Observing entrainment in music performance: Video-based observational analysis of indian musicians' tanpura playing and beat marking. Musicae Scientiae 11(1), 27–59 (2007)
16. Demey, M., Leman, M., Bossuyt, F., Vanfleteren, J.: The musical synchrotron: Using wireless motion sensors to study how social interaction affects synchronization with musical tempo. In: Proceedings of the 8th International Conference on New Interfaces for Musical Expression (2008)
17. Deweppe, A., Leman, M., Lesaffre, M.: Establishing usability for interactive music applications that use embodied mediation technology. In: Proceedings of the 2009 European Society for the Cognitive Sciences of Music Conference (ESCOM) (2009)
18. Leman, M., Demey, M., Lesaffre, M., Van Noorden, L., Moelants, D.: Concepts, technology, and assessment of the social music game sync-in-team. In: Proceedings of the 2009 International Conference on Computational Science and Engineering, vol. 4, pp. 837–842. IEEE Computer Society (2009)
19. De Nies, T., Van de Walle, R., Vervust, T., Vanfleteren, J., Demey, M., Leman, M.: Beatled - the social gaming partyshirt. In: Proceedings of the SMC 2011 - 8th Sound and Music Computing Conference, Padova - Italy, July 06-09, pp. 2011–2018 (2011)
20. Boker, S., Rotondo, J., Xu, M., King, K.: Windowed cross-correlation and peak picking for the analysis of variability in the association between behavioral time series. Psychological Methods 7(3), 338 (2002)

Graphical Spatialization Program with Real Time Interactions (GASPR)

Thierry Dilger

Freelance sound designer, Paris, France
thierry.dilger@gmail.com
www.sonabilis.com

Abstract. There is a dominant paradigm that links playback of audio files in a multi-channel sound system. It consists of a "top view" (or 3D) representation of the listening room with speakers set virtually in this space. The main drawback of this paradigm is the lack of order and harmony of trajectories representation leading to highly complex systems. In addition, it is very difficult to have an overview of a sound piece whole spatialization process. GASPR software gives the composer a new graphical representation of trajectories in space and time. It is based on a programmable behavioral video game engine. It is also possible to use any kind of sensors to control it live. GASPR relies on the RGB (red, green blue) color coding working on three axes: time (x), sound setup (y) and intensity of each sound (z). This paradigm opens up new doors for interactive surround sound composition.

Keywords: surround sound, spatialization, interaction, trajectories, video game engine, sound behavior, graphical representation, color mapping, sound installation, sound art.

1 Introduction

GASPR software creation started in September 2011. At this time, it was a necessity to create a new tool in order to achieve the realization of one big street sound installation during the festival “La fête des lumières” in Lyon.

The project is located in the heart of the city (place Gailleton). The space to be occupied is a large rectangle of 40 meters long and 10 meters wide where Shadow_Collectif [1] chose to work on the jungle theme: “canopy” project.

In charge of the sound creation, I decided to give life to this public space with the simple idea to have sound behaviors. For example, a sound from a monkey (one sound object) can scare other monkeys and will go the opposite direction thanks to the spatialization engine. I quickly realized that none of my sound softwares will manage this task easily, especially with 8 speakers in a strange double row setup (no sweet spots). So I decided to develop my own tool using an RGB mapping technique inherited from the project « Sound Island » [2]. I created this demo some years ago for Virtools company. This technique gives instant visual feedbacks of sound spatialization in a complex 3D environment (Figure 1).

M. Mancas et al. (Eds.): INTETAIN 2013, LNICST 124, pp. 136–145, 2013.

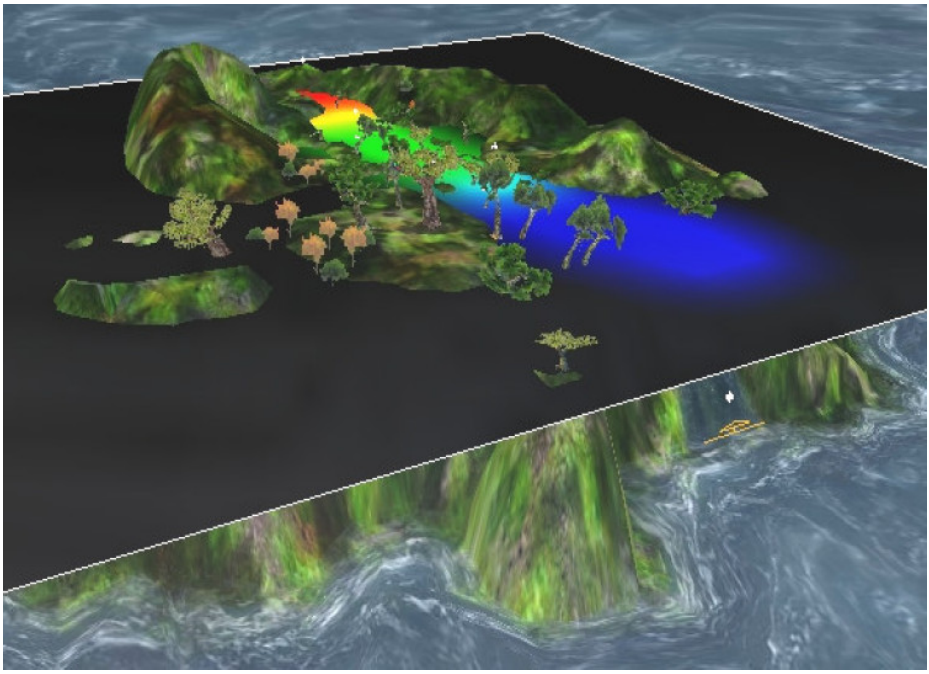


Fig. 1. RGB mapping for « Sound Island » project. Here each color gives its intensity to a sound file, a blend of 2 colors gives a blend of 2 sounds. The whole color pattern can be read as a sound mix driven by the position of an avatar in the 3D world (like GPS coordinate).

2 Existing Approaches

In « Sound Island » project, the color mapping technique gives the sound designer a tool to fulfill the virtual world with sounds and organize its distribution among space. It gives a simple solution to large scale spatialization scenarios. Other softwares work with a visual representation of the spatial composition like the “Holophon” GMEM [3]. This tool works with two windows, one representing trajectories in a virtual space and one showing sound arrangement on a "timeline" (Figure 2a). Iannix software [4] can show sound trajectories in a high resolution 3D representation as floating color ribbons for light paintings art creation (Figure 2b).

One advantage of these approaches is the immediate understanding of the distribution of sounds in space. Another big advantage is the ability to dissociate the visual rendering from the sound rendering, giving the option to decode the virtual world in stereo, 5.1, ambisonic, WFS, binaural... On the other hand, drawbacks comes from their respective paradigm: time domain is heavily constrained. Rethink the space / time representation is a foundation for GASPR software development.

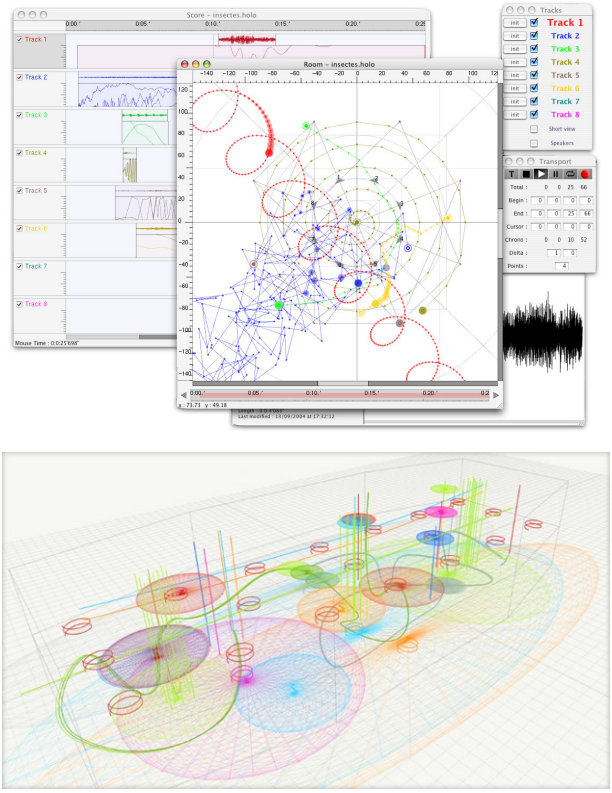


Fig. 2a / 2b. The “Holophon” gives the ability to draw trajectories for a sound setup. One movement and one sound are linked together on the timeline. “Iannix” software can have different style of representation. It is a very programmable approach (here a sound installation for “World Expo 2012”).

3 A New Paradigm

Nicolaus Copernicus [5], has managed to describe planets movements from an other referential than earth. The new referential (the sun) allows both to simplify trajectories representation and, at the same time, to make more accurate measurements. The main idea of Copernicus is to rethink fundamental relationships that link an observation point with perceptual phenomena accepting the idea to get rid of the intuitive model (what we actually see with our eyes).

GASPR software works with an abstract representation of space. In this it is similar to the paradigm used for "Spacemaps" in "CueStation" software [6] which gives relationship maps between speakers dissociated from the “real virtual representation” of an acoustic space. (Figure 6).

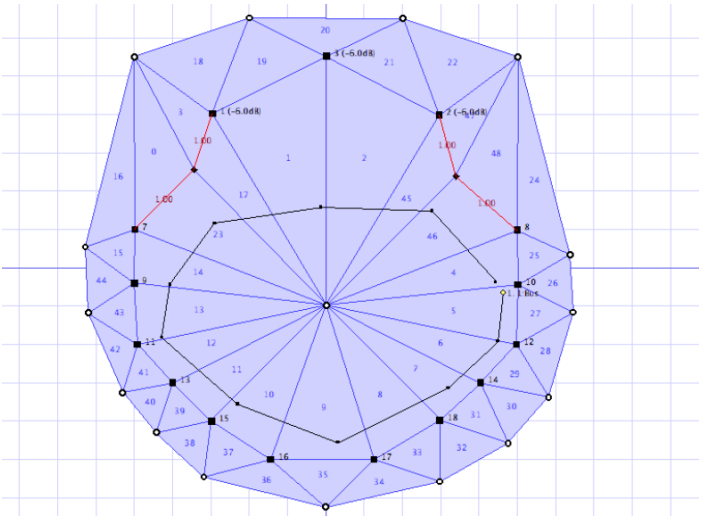


Fig. 3. One “Spacemap” from “CueStation” software (MeyerSound). Black dots represent loudspeakers and the black line is a sound movement. Actual physical loudspeakers position is not equal to this map: with VBAP technique [7], small and large distance between dots gives exactly the same sensation (constant power).

GASPR is based on three video games technologies: a graphic engine (DirectX) a behavioral engine (GameMaker) and a sound engine (FMOD). Visual informations are mapped to control sound informations. In other words, the performance of the graphic engine has a direct influence on the quality of the sound experience ("A new visual paradigm for surround sound mixing" [8]). Spatialization relies on amplitude panning and gives the user the freedom to create weird loudspeaker setups (like for “canopy” project). It allows to work with up to 8 sound outputs. Application field for GASPR is mainly sound art installation but can also work in live situation like concerts and performances. GASPR is a windows stand alone software (not a plug in).

3.1 Graphics Conducting the Sound Experience

Creating a GASPR composition is assembling blocks of different colors and shapes on a 2D timeline (Figure 4). The colors used are red, green and blue (RGB system) and all blends. Blocks shapes can be short or long and can also use transparency. The intensity of each color controls directly the intensity of sound. By mixing different colors together it gives a mix of different sound intensities: a composition. Like in "Sound Island" project, there are real-time position tracking systems which extract color information in order to conduct 8 sound outputs independently. Frame-rate is an important consideration but does not need to be very high for smooth results, most of my compositions are set to only 12 frames per second and give me the opportunity to work even with laptops and limited video cards.

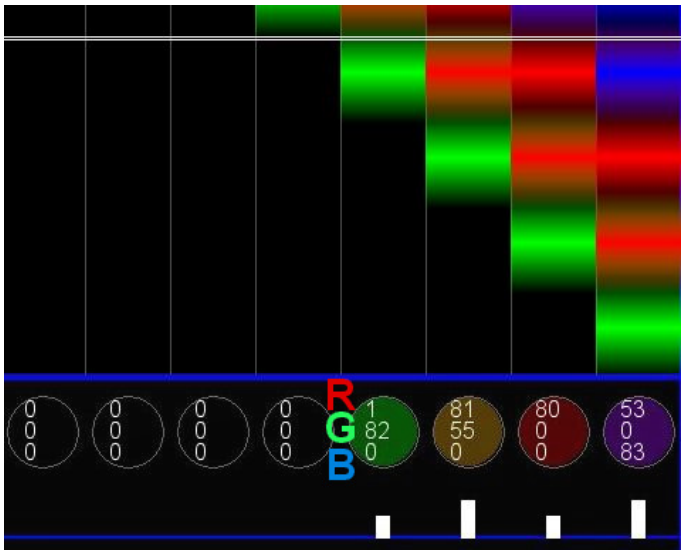


Fig. 4. Close up of a GASPR composition. You can see several colors blended in a vertical disposition. Lower part of the illustration shows the real-time RGB tracking system. Here only 4 outputs are fed with sounds (5, 6, 7, and 8).

A close work is the one from Memo Akten with "Simple Harmonic Motion study # 2a" [9]. A 3D geometrical structure, moves around a center line. When the nodes of the structure meet the line a sound is produced. The graphic engine conducts the sound experience. The behavioral model used is very important because it transforms the structure and the sound creation. The visual representation in GASPR is dynamic: color blocks can move, disappear, change in shape...

3.2 Free Loudspeakers Positionning

GASPR deals only with physical sound card outputs: 8 outputs for the current version. Typically each output is linked to a physical loudspeaker but we can find other scenarios where we can use more loudspeakers. Each output has its "track" that is filled with sounds. One advantage of this completely free loudspeakers positioning system is at the expense of a non intuitive aspect of the representation. One can put 8 speakers in circle, in line, with the use of height, split in 2 rooms, with different kind loudspeakers (several sub-woofers...). This is very similar to a stage lighting system (for theater) where the light immersion comes from the addition of discrete light sources set in a custom way.

In the previous illustration, loudspeakers are arranged from loudspeaker 1 (left) to 8 (right). It is important to decide clearly the layout in conjunction with the physical space. For example a diagonal line in block composition can represent a sound moving around us (Figure 5) only if the loudspeakers are put in circle. The same composition can give other perceptual results depending on loudspeakers disposition.

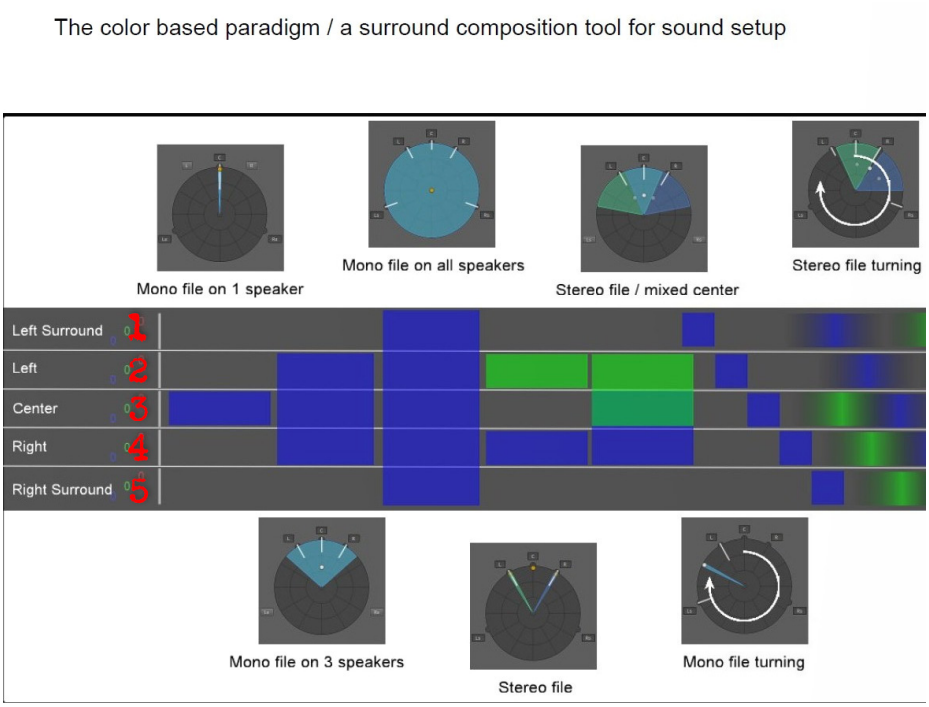


Fig. 5. GASPR paradigm represented for traditional 5.1 surround sound setup. (the sub-woofer is not represented here). Pay attention to the loudspeakers layout on the left part (vertical axis).

3.3 Absolute and Relative Time

The horizontal axis of the timeline represents the temporal component of the creation. There is one playback bar (white vertical bar) which linked the real-time RGB color tracking system.

Time can be absolute: at constant speed with a movement of the playback bar from left to right. The composer can create a sound piece of 5 minutes for example.

Time can also be relative: sound blocks are not triggered with the playback bar (like in traditional sound software), instead they are constantly played in loop with their intensity driven by their opacity. So even if the playback bar stops in the middle of a block, sound is still heard. If you add the option that blocks can move independently, you can create non predictable sound creations.

One strong aspect of GASPR is the ability to create a behavioral sound creation and navigate into it. Then the playback bar can be considered as a virtual listener in a sound world in constant evolution. For example you walk deep in the jungle, go back, stay somewhere in order to discover a strange animal very shy or even run quickly through the forest. A GASPR composition is virtually endless in the time domain.

4 Behavioral Sound Creation

GASPR is a dynamic environment, each block has its own life. They are considered as video games sprites by the engine. They can receive behavior scripts such as random generation, dynamic color change, dynamic color opacity, artificial intelligence... They are born, die, meet, hate... Like in “Spatium” software [10] where sound objects use gravity linked to spatialization.

The composer must programmed these behaviors at first and then “test” his creation. If lot of interactions take place in the composition, he will not be able to listen to all sound combinations. It is possible to even create emergent creations that the composer, himself, can not predict.

4.1 SoundObject

These are the main building blocks. They are linked with a sound file playing in loop. The sound file can be a mono file or an 8 channels file. Depending on the sound card, it is possible to play up to 24bits 96kHz. wav files with virtually no limits on length. Number of blocks is limited by the processing power especially if they contain lot of dynamic behaviors.

There are 3 main colors used in GASPR: Red, Green and Blue. It is possible to mix these 3 colors together, and make a mix of 3 sounds simultaneously (per output). In the current GASPR version there are 3 timelines in sync. It gives a maximum of 9 sounds mixed in real-time per output that is 72 sounds at one time for the whole 8 outputs. One limit of this model is that you can link separate Red blocks (for example) with separate sound files but you will not be able to mix it (same color).

Color coding is performed on 8 bits (256 values) and if 2 blocks of different colors meet a priority system takes place: Red is above Green and Green is above Blue. Maximum value is still maintained to 256 and a real time color fade occurs. It is similar to an “auto-ducking” effect in classic sound production. It is a solution to complex behavioral creations by giving the composer the ability to make sounds to always come from above or below.

4.2 SoundStructure

A "SoundStructure" is a collection of "SoundObjects" sharing an overall behavior. In a social context, a "SoundObject" is an individual and a "SoundStructure" a population. For example, a group of 4 blocks distributed over 4 outputs with a back and forth movement is a "SoundStructure." This may be related to sounds of wind appearing and disappearing gradually from the loudspeakers. Another structure is “monkeys”. Here, SoundObjects are attracted by the playback bar if it does not move too fast, else “monkeys” are moving away.

The "SoundStructure" may also be useful in order to play multichannel files (up to 8 channels) or perform sound premix.

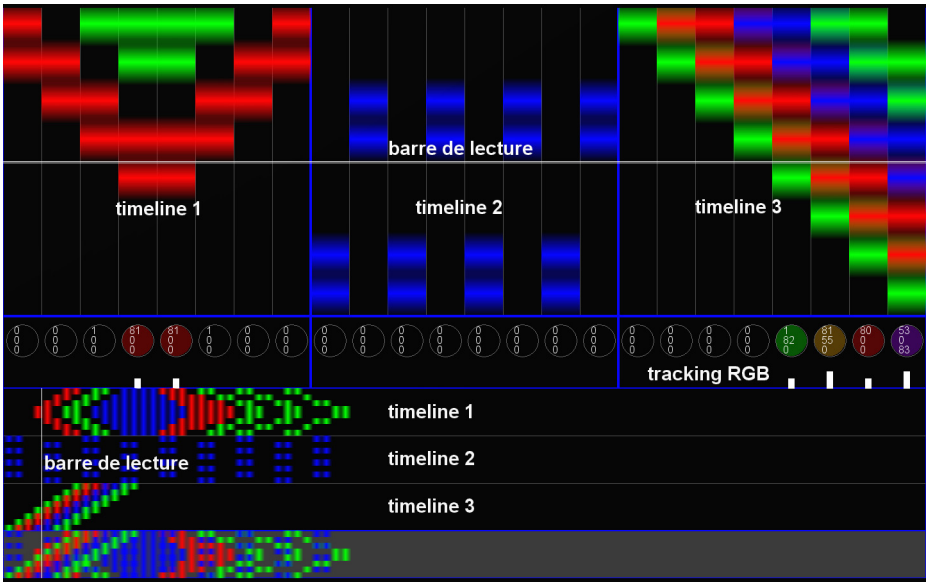


Fig. 6. GASPR actual version. You can see here 3 timelines horizontally and vertically (close up). Playback bar is the vertical white line in the lower left side and the horizontal white line from left to right (upper part).

4.3 MixMap

The "MixMap" represents the overall distribution of sounds over the 8 outputs. It is the gathering of all “SoundStructures” and “SoundObjects”. It is represented simultaneously in two ways (horizontally and close up vertically). Just by looking to a “Mixmap”, one can say that the creation will occupy very rarely all the speakers in the whole composition for example. It is possible to read a "MixMap" like a musical score but you do not read note pitch but sound spatialization among time (absolute or relative).

With GASPR built-in editor, one can create a whole “Mixmap” easily by adding color blocks, move and duplicate them. Built-in script language is also part of GASPR in order to create sound behaviors and interactions.

Several “Mixmaps” can be prepared in advance and launched in real-time. For example, one “MixMap” with birds can change into a “MixMap” with insects at the end of the day or during a performance you can follow a musician, changing the musical mood. In live situations it is important to attach a controller (like a gamepad) or map actions on the computer keyboard.

For art installation, thanks to a special dll, we can make communicate GASPR with Arduino. It opens a whole world of possibilities, making interactions between sensors, motors, lights... and real-time behavioral spatialization.

5 Conclusions

Developing a new paradigm for time, space, surround sound and interactions is an important quest that needs to rethink fundamentals of our intuitive models. By moving to more abstract representation of the acoustic space we can put in front the time and the interactive domain.

As a human being, our visual perception (and brain processing) is limited to a given amount of information at a given time. That's why it's important to help the composer brain giving him useful informations first and quality feedbacks. Sound composition tools and composers interact together mainly with visual informations. The beauty of this is at the end of the process, the audience will receive sound only.

GASPR software is developed with the idea of giving the composer a creative environment that will not be more and more messy if the composition is more an more dense. The interface has no scrolling bar, or menus and you can still have a virtually endless creation with surround sound (on 8 outputs) opened to interactivity.

Current GASPR version has an option to render an interactive composition into a single .exe file. All sound files are encrypted (with FMOD technology) inside a separate folder in order to protect them from being stolen. It gives the opportunity to share these creations to the public quite easily: there is no installation and it uses the Windows OS sound layer in order to output sound. Client framerate is automatically tested and limited in order to have the same quality of interactions.

The only thing to do from the user side is to set his sound card to 7.1 mode in Windows (if he has a 7.1 soundcard). Even if he has a built in stereo sound card in his computer, GASPR composition will play and he will hear only the first 2 outputs. One future development is to integrate several versions of a composition inside one GASPR file (stereo, quad, 5.1, 7.1) in order to give interesting results for a larger audience.

One option currently being tested, is to let the public replace sounds from the encrypted folder with their own sounds. It is then possible to create a new piece out of a current piece but keeping the actual behavioral model: the composer is no more a sound only composer but also a behavioral composer. By giving a degree of interaction opened to the public (able to interact with the piece through the computer keyboard for example), listeners are no more passive and they become: sound designers.

Interactive spatial composition is a huge space of expression for today's artists. I hope this proposition will help the community.

References

1. Shadow_Collectif, Art collective group from Paris, created in (2010), http://shadowcollectif.wordpress.com/
2. Sound Island, project, Thierry Dilger (2002), http://www.virtools.co.kr/dc/minisite/MiniSite/html/e03_DEMO_view_SoundIsland.htm

3. Holophon software from GMEM, real time software for sound trajectories, http://dvlpt.gmem.free.fr/web/static.php?page=Holophon_main
4. Iannix software, open source graphical sequencer, http://www.iannix.org/fr/
5. Copernicus, N.: mathematician and astronomer, father of the heliocentric model, http://en.wikipedia.org/wiki/Nicolaus_Copernicus
6. Matrix3's CueStation, spatialization software for large venue / show, http://www.meyersound.com/products/matrix3/
7. Virtual sound source positioning using vectore base amplitude panning, Ville Pulkki, AES (1997), https://aaltodoc.aalto.fi/bitstream/handle/123456789/2345/article1.pdf?sequence=2
8. Dilger, T.: A new visual paradigm for surround sound mixing. In: Poster for the ICSA First International Convention on Spatial Audio, Detmold, Germany (November 2011), and Engineering Brief Presentation for the 132 AES Convention, Budapest, Hungary (April 2011)
9. Atken, M.: visual artist, http://www.msavisuals.com/
10. Spatium, sound spatialization tool with physical models, http://spatium.ruipenha.pt/

Accuracy Study of a Real-Time Hybrid Sound Source Localization Algorithm

Fernando A. Escobar Juzga[1], Xin Chang[1], Christian Ibala[2], and Carlos Valderrama[1]

[1] Université de Mons, Belgium
{fernando.escobarjuzga,xin.chang,carlos.valderrama}@umons.ac.be
[2] University of Limerick, Ireland
sibala@acm.org

Abstract. Sound source localization in real time can be employed in numerous applications such as filtering, beamforming, security system integration, etc. Algorithms employed in this field require not only fast processing speed but also enough accuracy to properly cope with the application requirements. This work presents accuracy benchmarks of a hybrid approach previously proposed, which is based on the Generalized Cross Correlation (GCC), and the Delay and Sum beamforming (DSB). Tests were performed considering a linear microphone array simulated in MATLAB. Analysis through variations in array size, number of microphones, spacing and other characteristics, were included. Results obtained show that the proposed algorithm is as good as the DSB under some conditions that can be easily met.

Keywords: Accuracy, Sound localization, Generalized Cross Correlation, Beamforming, Computational Complexity, Real Time.

1 Introduction

Numerous applications can be encountered when dealing with sound source localization such as filtering or speech recognition [10], [19], [20], with the use of microphone arrays; by applying beamforming techniques [17], [1], [12], noise, reverberation effects and interference sounds can be filtered by focusing the beam towards the selected sound source location.

One of the well known algorithms to locate sound sources is the Generalized Cross-Correlation with Phase Transform (GCC-PHAT) [9], that can provide an angle ϕ which is the sound source direction of arrival (DOA); to compute it, the GCC uses the temporal shift estimation between a pair of microphones i and j that leads to the maximum cross-correlation between them. On the other hand, the Delay and Sum beamforming (DSB) algorithm [12], [3], can be used to build an acoustic energy map (Steered Response Power - SRP) of a predefined Field of View (FoV); under certain conditions it yields the exact position of the sound source. Finally, the Minimum Variance Distortion-less Response (MVDR) is often used to listen to large frequency band signals [2], but implies a big computational cost.

M. Mancas et al. (Eds.): INTETAIN 2013, LNICST 124, pp. 146–155, 2013.

In order to optimize and speed up the localization process, we proposed in previous work [7] and [8], a hybrid algorithm that combines the GCC-PHAT with the DSB-SRP to reduce the search area and decrease computational complexity. Our theoretical analysis showed a computational advantage of the hybrid algorithm over the DSB-SRP, thanks to the reduction of evaluated points where the energy response is computed.

Because we are interested in obtaining the exact position of a sound source, the GCC-PHAT solution resulted insufficient as it only provides the source's (DOA); on the other hand, since the SRP computation requires the output from the GCC-PHAT algorithm, it was straightforward to combine them both and optimize its execution. The basic idea of this approach is to create a reduced detection zone by drawing two lines: one above and the other below the GCC detected angle, with an inclination ε chosen by the user; then, the SRP can be computed on the constrained area as shown in Figure 1.

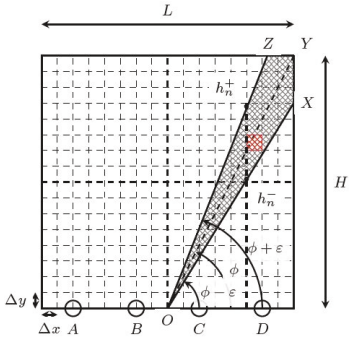


Fig. 1. 2-dimensional, 3x3 m, Field of View (FoV). The constrained area is enclosed by the upper and lower lines and the borders of the FoV.

The main contribution of this work is presented in Section 4 where we show the algorithm's response in terms of localization accuracy and number of detected maxima when changing microphone spacing, the epsilon value and inducing an artificial error to the angle obtained with the GCC algorithm. We employ the toolbox provided by the University of Kentucky [13], which generates synthetic scenarios to analyze with microphone arrays; several simulations were executed to derive accuracy measurements among them all.

The rest of the paper is organized as follows: Section 2 will summarize the latest research on sound localization using microphone arrays. In Section 3, we describe the proposed hydrid algorithm; Section 4 presents our simulated accuracy analysis and finally Section 5 lists our conclusions and suggested further work.

2 Related Work

Several works have been reported in the field of sound source localization. Valin et al. use an 8 microphone array to estimate the time delay of arrival (TDOA) using the GCC-PHAT technique on a moving robot [18] and report an average of 3 degrees accuracy in computing the direction of arrival (DOA); the exact position of the detected target is left as future work though. On a second work [19], the authors use a particle filtering technique for tracking moving sound sources with 2, 8-microphone array configurations. They report high accuracy detecting both elevation and azimuth angles.

Another common technique to locate sound mainly in human shaped robots, is called Head Related Transfer Function (HRTF) which estimates the difference in level intensity between the two ears (microphones); however, whilst human ears are shaped in a special way to enhance localization, the algorithm is rather complex for real time implementation. Some related work can be found in [11], [14] and [6].

There are some researches working on three dimensional sound localization such as [16], [5] and [15]. Reports in [16] describe the first working, scalable and cost-effective array that offers high-precision localization of conversational speech in large semi-structured spaces; it achieves high throughput for real-time updates of tens of active sources. Yoko et. al [15], proposed a spherical microphone array design for spatial sound localization. This structure has 64 microphones arranged in a 350-mm-diameter sphere. It is designed to be mounted on a mobile robot with omni-directional directivity in both azimuth and elevation angles. In order to achieve better accuracy, the number of microphones is substantially raised in this kinds of designs. Since the computation load increased, high performance processors are required. Adittionally, bigger areas are occupied.

3 Proposed Hybrid Algorithm

Since our algorithm has already been presented in previous articles [4], [7], [8], only a brief explanation of its operation will be provided; readers are encouraged to revise the cited references for more detailed information.

The Generalized Cross Correlation between two signals provides the estimation of the temporal shift between two microphones i and j that leads to the maximum cross-correlation between them as in Equation 1:

$$\Delta_{ij} = arg_k \ \ max \ R(k) \tag{1}$$

The cross correlation between two microphones is computed by taking the inverse Fourier transform of the product of the first microphone FFT (Fast Fourier Transform) and the conjugated FFT of the second one. To correct the effect of phase, and improve robustness against noise and other undesired effects, there is a correction called PHAT, i.e. phase transform that can be applied yielding Equation 2:

$$R(k) = \quad IFFT \quad \left(\frac{FFT(f(t)).FFT^*(g(t))}{\mid FFT(f(t)).FFT^*(g(t))\mid^{\beta}} \right) \tag{2}$$

Equation 2 is defined as the GCC-PHAT; β is a coefficient factor in the interval $(0, 1)$. The IFFT is performed to go back to the time domain and extract the corresponding value of index k. The value of k can be computed by taking the index of the maximum value from the GCC-PHAT output. Using the far field approximation, the cosine of the angle of arrival, measured by microphones i and j, can be computed as in Equation 3:

$$cos(\phi)_{ij} = \quad \frac{kv_s}{f_s d_{ij}} \tag{3}$$

Where f_s is the sampling frequency, d_{ij} is the distance between microphone i and j, and v_s is the sound speed.

The output of the GCC algorithm can be used to compute the energy response of a predefined FoV. By assuming the sound source to be located at a certain point in space, it is possible to establish the theoretical delay of the signal between every pair of microphones; using such delays, we can extract and add up the energy contribution of every pair from the GCC output. Under certain microphone array configurations and, assuming a single sound source, there will only be one point in the space where, all delays will match the maximum energy possible, that is, the real source location.

As shown in Figure 1, we can restrain the search region by focusing on the relevant part of the FoV, using the computed angle. Since the hybrid approach only restricts the search area, the output is expected to be as accurate as the regular algorithm. In terms of computational cost, great reduction is obtained by considering less points but other computations are required to establish the restricted area boundaries. More specifically, it's necessary to perform some tangents and cotangents whose amount, depend on the size of the small squares. As shown in Figure 1, the number of small squares to be evaluated in each column, depend on the heights h_1^- and h_1^+ which are defined as follows:

$$h_n^- = \quad n \cdot \Delta_x \cdot \tan(\phi - \varepsilon) \tag{4}$$

$$h_n^+ = \quad n \cdot \Delta_x \cdot \tan(\phi + \varepsilon) \tag{5}$$

In summary, the total number of tangents for a specific FoV of length X and resolution R is:

$$NumTangents = \quad \left(\frac{2 \cdot X}{R} \right) \tag{6}$$

The following subsection will present our accuracy estimations for the proposed algorithm.

4 Accuracy Analysis

The hybrid approach has two error sources: the one introduced by the GCC algorithm when obtaining the angle of arrival, and the one inherent to the DSB-SRP

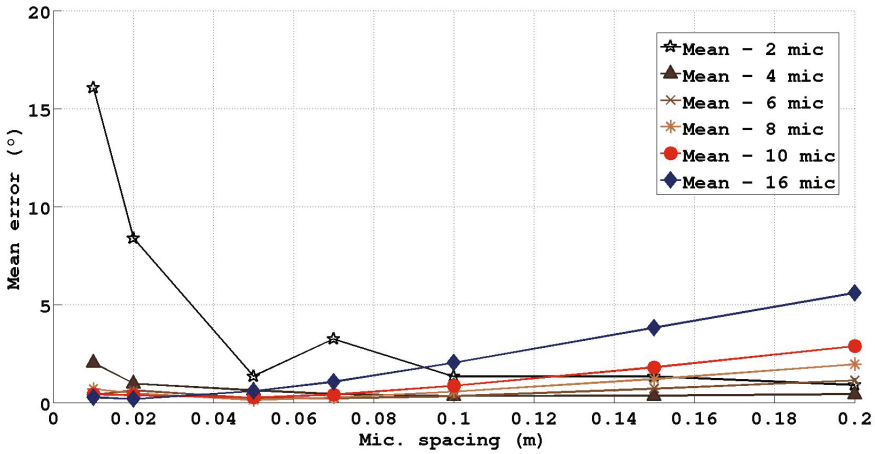


Fig. 2. GCC mean error for different spacings and number of microphones. The algorithm provides a good accuracy with more than 2 microphones; increasing their number does not provide better results. The error obtained is less than 5 degrees for the majority of cases.

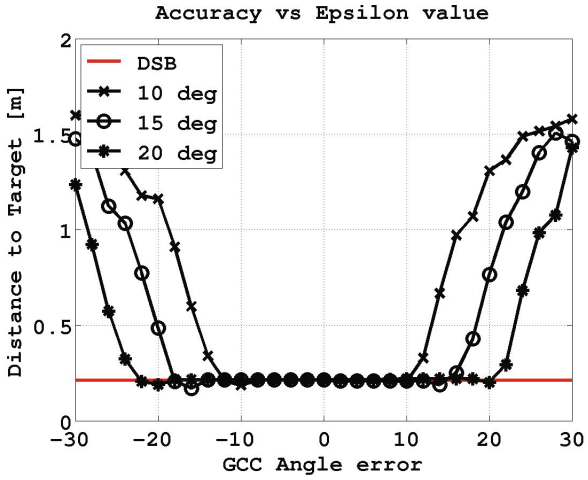
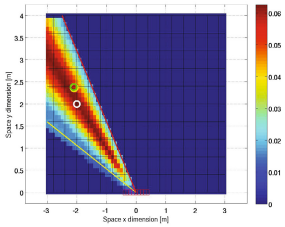


Fig. 3. Accuracy for different values of epsilon vs GCC error. The hybrid algorithm performs as good as the DSB when the value of epsilon is greater or equal than the error of the GCC, if any.

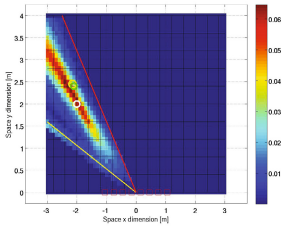
itself. Because both algorithms can yield errors at the same time we present the results considering them both. For all the following simulations we varied the target position for a better generalization and present the mean value as the final result.

We conducted a first study related to GCC precision using the aforementioned toolbox [13]; after considering different scenarios for linear arrays, we obtained an average error of $2-4°$ for arrays of at least 4 cms long with more than 3 microphones. Arrays that did not respect such conditions yielded up to 16-degrees errors. This results are shown in Figure 2; when the number of microphones increases, so does the error. Since every pair of microphones provide an estimated angle, the error increase can be caused by the averaging of all measured values.

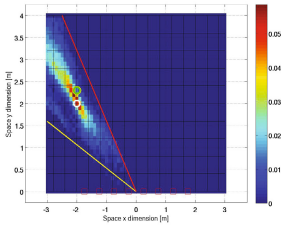
The hybrid algorithm was analysed in terms of the GCC error; a few parameters can be changed to tune up its output but according to our results, only some of them are relevant. Initially, we varied parameter ε and measured the



(a) 10 cm



(b) 30 cm

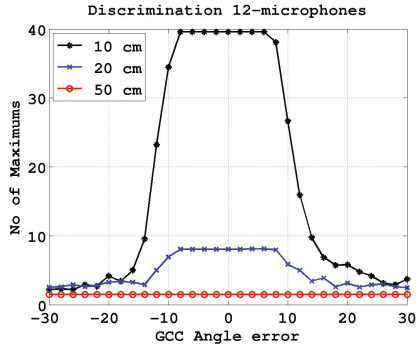


(c) 50 cm

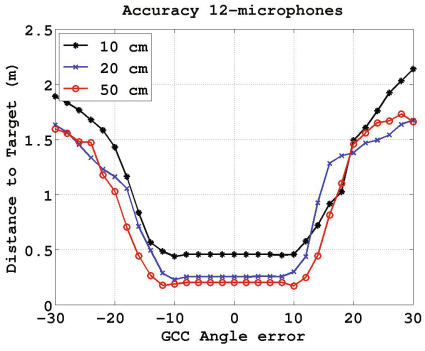
Fig. 4. SRP response of an array of 8 microphones at different spacings using the hybrid GCC-DSB algorithm

distance of the detected point to the real source location. The test was performed under a configuration of 8 microphones, 50 cm apart, and assuming a GCC error range of $[-30°, 30°]$; we obtained Figure 3. The horizontal red line represents the output when using the DSB-SRP, while the black lines show the behavior of the GCC-DSB for different epsilon values; when the GCC angle error is greater than the value of epsilon, accuracy tends to be rapidly lost. On the other hand, if the GCC error lies whithin the range comprised by epsilon, the accuracy is the same as with the DSB-SRP.

Through the simulations performed, we noticed that one of the most important parameters that affected localization accuracy was the microphone spacing; for instance, consider Figure 4, where the same scenario is presented for three different spacings; in Figure 4a the microphones are only 10 cm apart and a big red fringe of points are detected as possible source location. On the contrary, Figures 4b and 4c show that, when increasing their separation, the energy plot is much clearer, thus enabling better accuracy. Since more than one point can be detected as the maximum, we decided to select the mean point between all of them as the source location. In Figure 5 the algorithm is tested with 12 microphones and a 10° epsilon, changing their spacing. In Figure 5a we can see



(a) Number of maximums detected



(b) Mean distance to target obtained

Fig. 5. Performance on a 12-microphone array for different spacings between each one

the amount of points that are detected for three different cases; although the number of maximums between 20 and 50 cms does not change much, the difference in array size greatly does. In accordance with this result, yet with smaller differences, Figure 5b shows that the accuracy improves for longer spaced arrays. Once more, the validity of the result holds as long as the GCC error is less than the value of epsilon.

A final study was carried on to understand the behavior of the algorithms in relatively small arrays; experimental results using the aforementioned database showed that for arrays of 10 cms long or less, even with no GCC angle error, is not possible to obtain a good precision. According to our estimations, sufficiently accurate results can be obtained with arrays of at least 20 cms long; although, as shown in Figure 4, a considerable amount of maximums will be detected, on average, good estimations can still be obtained with this configuration. Our results show that precision obtained tends to slightly improve with more microphones but is practically the same in all cases.

5 Conclusions

Through the analysis performed in this work, we could verify that the proposed hybrid algorithm can perform as accurately as the traditional DSB-SRP with linear microphone arrays. An important factor that can drastically change the output from the algorithm is the error induced by the GCC-PHAT algorithm; when the angle error is greater than the value of parameter epsilon, the algorithm loses its accuracy.

Linear microphone arrays whose microphone spacing is less than $40 - 50$ cms present difficulties to establish the exact coordinate of the sound source; when the array length is fixed, irrespective of the number of microphones, the localization accuracy is very similar. Since accuracy changed with target position and angle, in general terms we consider the best results were obtained when using between 4 to 10 microphones, for both the GCC and the GCC-DSB.

The aforementioned analysis were done at the theoretical level using an artificial database, which generated an impulse response, noised signal, for each microphone however, we consider that a similar study, with real microphone and signals is necessary to verify the robustness of the simulator engine, and to confirm the validity of the conclusions reached in this paper.

The severity of the localization error of the algorithm can vary depending on the application domain; if employed for detecting small objects, an error of 10 cm can be unacceptable, but the contrary might occur for locating speakers, since a person's personal space can be at least $30cm^2$. Improvements can be also obtained by changing the decision function when different maxima have been found or by applying other methodologies to compute the SRP after the angle has been measured. This is however out of the scope of the presented work.

Acknowledgements. This work was funded by the Wallonia Region DG06 Belgium under grant 917005-CTEUC 2009 Eureka ITEA DiYSE.

References

1. Benesty, J., Chen, J., Huang, Y., Dmochowski, J.: On microphone-array beamforming from a mimo acoustic signal processing perspective. IEEE Transactions on Audio, Speech, and Language Processing 15(3), 1053–1065 (2007)
2. Cox, H., Zeskind, R., Owen, M.: Robust adaptive beamforming. IEEE Transactions on Acoustics, Speech and Signal Processing 35(10), 1365–1376 (1987)
3. Dmochowski, J.P., Benesty, J., Affes, S.: A generalized steered response power method for computationally viable source localization. IEEE Transactions on Audio, Speech, and Language Processing 15(8), 2510–2526 (2007)
4. Escobar, F.A., Ibala, C., Chang, X., Valderrama, C.: Fast accurate hybrid algorithm for sound source localization in real time. International Journal of Sensors and Related Networks 1(1), 1–7 (2013)
5. Fréchette, M., Letourneau, D., Valin, J.-M., Michaud, F.: Integration of sound source localization and separation to improve dialogue management on a robot. In: 2012 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS), pp. 2358–2363. IEEE (2012)
6. Hörnstein, J., Lopes, M., Santos-Victor, J.: Sound localization for humanoid robots building audio-motor maps based on the hrtf (2006)
7. Ibala, C., Vachaudez, J., Fourtounis, G., Possa, P., Valderrama, C.: Combining sound source tracking algorithms based on microphone array to improve real-time localization. In: 2012 Proceedings of the 19th International Conference on Mixed Design of Integrated Circuits and Systems (MIXDES), pp. 478–483 (May 2012)
8. Ibala, C., Escobar, F.A., Chang, X., Valderrama, C.: Hybrid algorithm computation methodology to accelerate sound source localization. International Journal of Microelectronics and Computer Science 3(3), 99–110 (2012)
9. Knapp, C., Carter, G.: The generalized correlation method for estimation of time delay. IEEE Transactions on Acoustics, Speech and Signal Processing 24(4), 320–327 (1976)
10. Li, Q., Zhu, M., Li, W.: A portable usb-based microphone array device for robust speech recognition. In: IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP 2009, pp. 1301–1304 (April 2009)
11. MacDonald, J.A.: An algorithm for the accurate localization of sounds (2005)
12. McCowan, I.: Robust speech recognition using microphone arrays (2001)
13. University of Kentucky. Performance analysis of srcp image based sound source detection algorithms (2010)
14. Rothbuncher, M., Kronmuller, D., Durkovic, M., Habigt, T., Diepold, K.: Hrtf sound localization (2011)
15. Sasaki, Y., Kabasawa, M., Thompson, S., Kagami, S., Oro, K.: Spherical microphone array for spatial sound localization for a mobile robot. In: 2012 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS), pp. 713–718. IEEE (2012)
16. Sun, D., Canny, J.: A high accuracy, low-latency, scalable microphone-array system for conversation analysis (2012)
17. Tashev, I.J.: Sound Capture and Processing: Practical Approaches. Wiley Publishing (2009)

18. Valin, J.-M., Michaud, F., Rouat, J., Letourneau, D.: Robust sound source localization using a microphone array on a mobile robot. In: Proceedings of the 2003 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS 2003), vol. 2, pp. 1228–1233 (October 2003)
19. Valin, J.-M., Michaud, F., Rouat, J.: Robust localization and tracking of simultaneous moving sound sources using beamforming and particle filtering. Robotics and Autonomous Systems 55(3), 216–228 (2007)
20. Zwyssig, E., Lincoln, M., Renals, S.: A digital microphone array for distant speech recognition. In: 2010 IEEE International Conference on Acoustics Speech and Signal Processing (ICASSP), pp. 5106–5109 (March 2010)

Image Surround: Automatic Projector Calibration for Indoor Adaptive Projection

Radhwan Ben Madhkour[1], Ludovic Burczykowski[2] Matei Mancaş[1], and Bernard Gosselin[1]

[1] Numiediart Institute, University of Mons, Bd. Dolez 31, Mons, Belgium
radhwan.benmadhkour@umons.ac.be
[2] University of Paris 8, Saint-Denis, France
ludovic.czy@gmail.com

Abstract. In this paper, we present a system able to calibrate projectors, perform 3D reconstruction and project shadow and textures generated in real-time. The calibration algorithm is based on Heikkila's camera calibration algorithm. It combines Gray coded structured light patterns projection and a RGBD camera. Any projection surface can be used. Intrinsic and extrinsic parameters are computed without a scale factor uncertainty and any prior knowledge about the projector and the projection surface. The projector calibration is used as a basis to augment the scene with information from the RGBD camera. Shadows are generated with lights. Their position is modified in real-time to follow a user position. The 3D reconstruction is based on the Kinect fusion algorithm. The model of scene is used to apply texture on the scene and to generate correct shadows.

Keywords: projection, calibration, tracking, scene augmentation.

1 Introduction

Video projectors are mostly known for their classical use: a projection on a planar screen with the projector located in front of it. The homography integrated in the menu of all new the projectors has enabled to slightly change the projector position but the screen is still a planar surface.

During the last years new developments in video projector calibration arose. With structured light scanning, projection on complex surfaces can be performed [13] but the correction is perfect from only one point of view: the camera. Furthermore, if any object moves, the process has to be restarted.

To provide more capabilities to projectors in terms of projection surface, multiple methods for projector calibration have already been proposed. Audet and Okutomi [2] method provides a good way to calibrate the intrinsic parameters of the projector but it does not solve the problem of the extrinsic calibration. The method uses a planar board to calibrate a camera and a projector at the same time. If the projector is not close to the camera, it is difficult to project on the board and at the same time, put the board in a good position for the

M. Mancas et al. (Eds.): INTETAIN 2013, LNICST 124, pp. 156–162, 2013.

camera detection. A solution is to increase the size of the board but the method becomes less user-friendly.

In [16], Yamazaki et al. presented a method for the geometric calibration of a video projector using an uncalibrated camera and structured light. Nevertheless, the method performs the calibration up to a scalar factor. Moreover, a prior knowledge of the principal point is needed.

With the rise of intelligent TV and social gaming, most of the applications need to provide a visual feedback to the user. Microsoft's Kinect sensor allows to track people easily and to develop intuitive human to computer interactions [5]. Tracking moving objects or people is easier [4,12,9] but to project on them, a full geometric calibration of the projector is required. The need for an easy-to-use projector calibration is growing.

In this paper, we propose a fully automatic method for the geometric calibration of a projector. The process is based on Heikkila's algorithm [7] but it is extended to projector calibration with the use of structured light and a RGB-Depth (RGBD) camera. We apply the calibration to augment a scene with shadows and textures.

The rest of the paper is organised as follows. Section 2 describes the projector model and the projector calibration method. Section 3 gives the results. Section 4 shows how the scene is augmented. Finally, Section 6 concludes the work and gives some perspectives to improve the method.

2 Projector Calibration

The mathematical model of the projector used in this paper is the pinhole model [6]. Indeed, a projector is the same as a camera, the only difference being the light ray direction [10]. This model is represented mathematically by equation 1.

$$x \sim P\ X_{world} = K[R|t]X_{world} \tag{1}$$

In this equation, $x(u, v, 1)$ is the pixel position in the projected 2D image and $X_{world}(X, Y, Z, 1)$ is a 3D position where the pixel x lights up. The matrix K is called the projector calibration matrix. R—t is the coordinate transform from the world coordinate frame to the camera coordinate frame. R is the rotation matrix and t, the translation vector.

The projector calibration needs multiple couples of 3D coordinates and pixel coordinates. We propose to use Heikkila's algorithm [7] to perform the calibration. Heikkila algorithm is based on the direct linear algorithm and does not impose a constraint on the surface to use. A structured light projection gives pixel to pixel correspondences between the projector and the camera, while the use of a RGBD camera gives the 3D coordinates of the projected points. Therefore, couples of 3D and pixels coordinates are retrieved. The proposed method is represented in figure 1.

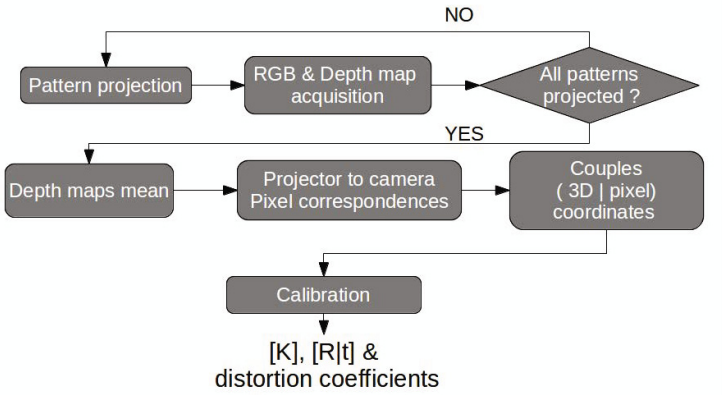


Fig. 1. Calibration process

The process is decomposed in different steps:

1. Project the Gray-coded binary patterns
2. Acquire a RGB and a depth map for each projected pattern
3. Compute the correspondences between the pixel of the projector and the RGBD camera
4. Average the depth maps to eliminate possible noise on the depth mesure
5. Compute the couple of pixel 2D coordinates and its 3D coordinates
6. Apply Heikkila's algotrithm

The method does not impose a planar surface constraint and is fully automated. For more details, the method is further described in [3].

3 Calibration Results

We performed multiple calibrations for different camera positions and for different zooms of the projector. The average reprojection error from multiple calibration tests is presented in the table 1.

Compared to state of the art methods [16,2], the method provides a higher reprojection error. Those high values are explained by:

- the error introduced during the structured light correspondences estimation,
- the error introduced by the RGBD camera,
- Heikkila's algorithm which is less accurate.

Nevertheless, the proposed algorithm has the important advantage to be fully automatic which is not the case of the other methods, and there is no need of any a priori knowledge: only a non planar surface is needed. Those advantages are central in an application which could be used by non specialists in their living room.

Table 1. Average reprojection error RMSE (in pixel)

Average reprojection error	
u	v
2.5368	2.3558

4 Scene Augmentation

It is complex to handle textures and shadows directly in an OpenGL scene. To simplify this process, we used Unity [14] rendering engine. It manages the shadows and textures in real time and simplifies the light management compared to pure OpenGL.

To perform the real-time rendering, the projector is modelled in the virtual world by a perspective projection. The perspective describes a pyramid in which every object is rendered (see Figure 2 [11][1]). Equation 2 provides a way to transform the projector matrix value into a perspective transform.

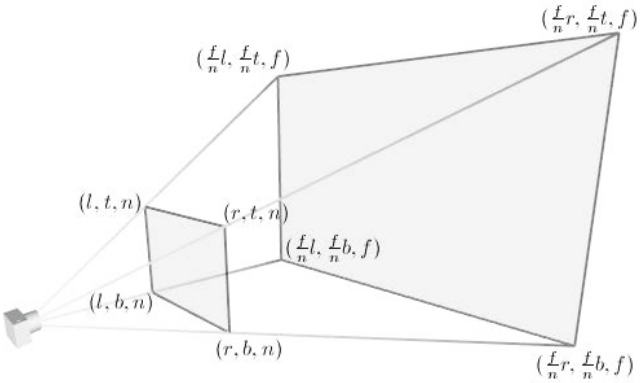


Fig. 2. Calibration process

$$perspective\ transform = \begin{pmatrix} \frac{2n}{r-l} & 0 & \frac{r+l}{r-l} & 0 \\ 0 & \frac{2n}{t-b} & \frac{t+b}{t-b} & 0 \\ 0 & 0 & -\frac{f+n}{f-n} & -\frac{2fn}{f-n} \\ 0 & 0 & -1 & 0 \end{pmatrix} \quad (2)$$

Equation 2 uses six parameters: near (n), far (f), left (l), right (r), top (t), bottom (b). The left (right) is the position of the left (right) plane of the pyramid along the x axis. The top (bottom) is the position of the top (bottom) plane of the pyramid along the y axis. Those values are obtained from the K matrix (f_u,f_v, u_0, v_0, width and height).

The rendering engine communicates with the tracking and calibration system via an OSC (Open Sound Control) communication [15]. This UDP protocol is fast and efficient for control commands like in this case and a variety of other software have OSC communication already implemented.

5 First Results

In a first experiment, a video projection is achieved in real time on a moving human. Figure 3 shows the reprojection of a 3D tracking information from OpenNI [12]. OpenNI library allows to extract a human silhouette from the depth map of the Kinect sensor. We use this silhouette as a blob on which the video projector will project red pixels while it projects white pixels on the background. The red blob projection follows the user in real-time regardless of the user position. The error can be seen as red edges around the background shadow of the silhouette. This error is due both to the Kinect blob which is not perfect and has a lag between the real motion and the detected movement and to the reprojection error.

In addition to user body, information of his position are projected on the ground, in front of the user. The system also works with several users and interaction information can be projected on the ground around them.

The first user reactions are positive and people are astonished by how reactive the projector is to their movements. The only lag is due to the delay of blob detection due to the Kinect sensor.

Fig. 3. Results of the projection in real time on a user

In a second experiment, a video projection is achieved on a complex 3D surface. The shadow of a 3D object is modelled and projected on the ground: in that way the shadow can artificially be moved and create an impression of illumination change. The surface is reconstructed with the RGBD camera using Kinect fusion algorithm [8]. The mesh is obtained by slightly moving the Kinect camera (figure 4) and the area to be projected is manually selected at this stage.

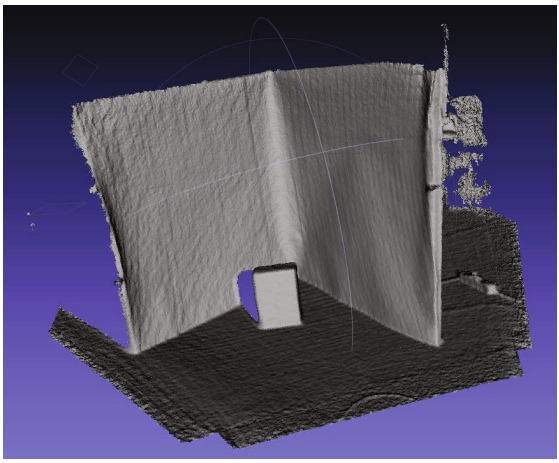


Fig. 4. Surface reconstructed with Kinect Fusion

6 Conclusion and Future Works

We have described our method for the geometric calibration of a projector and applied it to real-time projections. An application to projecting directly on a human and a novel application to projecting moving shadows on complex 3D objects were shown. Those two preliminary experiments show the feasibility of an application which uses a Kinect sensor and a classical projector to augment in real time a complex scene. An interesting scenario for those applications is in modifying in real-time the ambiance of the living room in front of a TV depending on the content displayed on the TV. Images which can be triggered by the content can be projected on objects in front of the TV or on people passing between the viewer and the TV surrounding the viewer with images.

Despite higher reprojection errors than in the state of the art of projector calibration methods, the proposed method has multiple advantages.

First, the planar surface constraint introduced by most of the state of the art techniques is removed by the combination of Heikkila's algorithm [7], the structured light and the RGBD camera. The RGBD camera simplifies the calibration process, thanks to the depth map. In the same time, the RGB sensor allows to acquire images of the projected Gray coded patterns and then, to calculate the projector to camera pixel correspondences. With this combination, the calibration can be performed on any complex surface in real time.

Second, the method is fully automated and does not require any user intervention, which is a key step towards consumer-oriented applications.

Finally, no prior knowledge about the projection surface and the projector are needed to achieve the calibration which virtually opens adaptive projections to any complex indoor scene such as living rooms.

The applications described here are preliminary and they need to be tested in several scenarios and to get more viewer feedback on the results. Also, the

calibration method needs optimization to reduce reprojections errors and cameras with a faster frame rate than the Kinect will be used to reduce the delay between projection and object movements.

References

1. Ahn, S.H.: Opengl (2011) (last viewed February 1, 2013)
2. Audet, S., Okutomi, M.: A user-friendly method to geometrically calibrate projector-camera systems. In: Computer Vision and Pattern Recognition Workshop, pp. 47–54 (2009)
3. Madhkour, R.B., Mancas, M., Gosselin, B.: Automatic geometric projector calibration: Application to a 3d real-time visual feedback. In: Proceedings of the 8th International Conference on Computer Vision Theory and Apllications (VISIGRAPP 2013), pp. 420–424 (2013)
4. Bradski, G., Kaehler, A.: Learning OpenCV: Computer Vision with the OpenCV Library. O'Reilly (2008)
5. Harrison, C., Benko, H., Omnitouch, A.D.W.: wearable multitouch interaction everywhere. In: Proceedings of the 24th Annual ACM Symposium on User Interface Software and Technology (UIST 2011), pp. 441–450. ACM (2011)
6. Hartley, R., Zisserman, A.: Multiple view geometry in computer vision. Cambridge University Press (2004)
7. Heikkila, J.: Geometric camera calibration using circular control points. IEEE Transactions on Pattern Analysis and Machine Intelligence 22(10), 1066–1077 (2000)
8. Izadi, S., Kim, D., Hilliges, O., Molyneaux, D., Newcombe, R., Kohli, P., Shotton, J., Hodges, S., Freeman, D., Davison, A., Fitzgibbon, A.: Kinectfusion: real-time 3d reconstruction and interaction using a moving depth camera. In: Proceedings of the 24th Annual ACM Symposium on User Interface Software and Technology (UIST 2011), pp. 559–568. ACM (2011)
9. Kalal, Z., Matas, J., Mikolajczyk, K.: P-N Learning: Bootstrapping Binary Classifiers by Structural Constraints. In: Conference on Computer Vision and Pattern Recognition (2010)
10. Kimura, M., Mochimaru, M., Kanade, T.: Projector calibration using arbitrary planes and calibrated camera. In: Proceedings of the 2007 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR 2007), pp. 1–2. IEEE Computer Society (2007)
11. Marsh, D.: Applied Geometry for Computer Graphicsand CAD. Springer (2004)
12. OpenNI. Openni user guide (November 2010) (last viewed January 19, 2011)
13. Tardif, J.P., Roy, S., Trudeau, M.: Multi-projectors for arbitrary surfaces without explicit calibration nor reconstruction. In: Proceedings of the Fourth International Conference on 3-D Digital Imaging and Modeling (3DIM 2003), pp. 217–224 (2003)
14. Unity Technologies. Unity rendering engine (2013)
15. Matt Wright. Open sound protocol specification 1.0 (2002)
16. Yamazaki, S., Mochimaru, M., Kanade, T.: Simultaneous self-calibration of a projector and a camera using structured light. In: Proceedings of the 2011 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR 2011) - Workshops (Procams 2011), pp. 67–74. IEEE Computer Society (2011)

EGT: Enriched Guitar Transcription

Loïc Reboursière and Stéphane Dupont

Laboratoire de Théorie des Circuits et Traitement du Signal (TCTS),
Faculté Polytechnique de Mons (FPMs), Belgique
{loic.reboursiere,stephane.dupont}@umons.ac.be

Abstract. EGT (Enriched Guitar Transcription) is a real-time and automatic guitar playing transcription software. Unlike most of the automatic score transcription software, not only the *note on*, *note off* events and pitch tracking are performed, but all the main guitar playing techniques are detected as well, providing a more complete transcription of the playing. These detections are made possible thanks to the use of an hexaphonic pickup (one pickup per string) enabling a string-by-string analysis. These transcriptions can then be used in many different contexts and / or embedded in different tools in order to obtain high-level information on the instrumentalist playing. This paper will demonstrate two use cases: a complete and realtime tablature writer and a 3D neck model controlled by the detected guitar playing events.

Keywords: Guitar playing techniques, hexaphony, music information retrieval, automatic score transcription, augmented guitar, guitar controller.

1 Introduction

From instrument-controlled synthesizer (may they be[1]analog or [2]digital), to automatic score writing, the detection of instrumentalist's playing and gestures have been the focus of many studies. These range from Music information Retrieval domain to augmented instruments area in which the instrumentalist playing is analyzed, characterized and used for different purposes. Regarding the guitar, research also range from MIR [2], [1]to augmented instruments [5], [3] and elements of playing at different scale have been studied, [4], [9], [6] .

In [8], we presented algorithms to detect each of the major guitar playing techniques. EGT (i.e, Enriched Guitar Transcription), integrates all these algorithms into a single software, performing real-time hexaphonic analysis of the string signals. The following playing techniques can be detected: hammer-on, pull-off, bend, harmonic, slide, palm muting. In addition to these, *note on* and *note off* events are reported, as well as the pitch of the played note. Besides, the plucking point (i.e, the position where the string is plucked) is although calculated.

[1] http://www.joness.com/gr300/GR-500.html
[2] http://windsynth.net/basics.html

M. Mancas et al. (Eds.): INTETAIN 2013, LNICST 124, pp. 163–168, 2013.

The detection software has been wrapped within a VST audio plugin. This standard has been chosen due to its integration in most audio software. The detected guitar playing events can be outputted in both OSC and MIDI formats.

This plugin doesn't only compute all the guitar events detection but provides other useful functionalities, which will be fully describe in section 2. In section 3, we describe two uses of this plugin: a complete and real-time tablature writer and a 3D guitar neck representation which reacts accordingly to the various detected guitar playing events.

2 Sofware Implementation

2.1 VST Choice and Output Formats

The VST audio plugin format was chosen because of its popularity and its availability on most music software, from Digital Audio Workstation, i.e ProTools, to any blank page real-time music software making, i.e Pure Data. As previously mentioned, all detected events can be fully outputted through OSC and the MIDI standard can be used as well to drive existing MIDI synthesizer. These communication options then make the plugin open to the majority of audio software. It needs to be pointed out that the use of MIDI standard entails a reduction of the set of techniques that are transcribed, as it does not provide a way to describe techniques such as palm muting, harmonics, etc. Some flexibility can be offered through the provision of MIDI control change messages though. When using OSC, the richness of all detected guitar playing events is kept however.

2.2 Software Elements

Setting Up. The software is made up of several elements organized in 4 tabs: General, Detector, Region and Behaviour. Several general options can be adjusted to set up the detection system: an **audio device** (and then sampling frequency and block size) as well as an external **MIDI device** can be specified. **Input gains** can be set up individually for each string but an automatic gain system can compute each gain after the instrumentalist has played it's loudest. The **tuner** is calibrated by defining the original tuning of the guitar. Eight different **OSC senders** can be defined to transmit the detected events through the network.

Predefined setups for the **detection parameters** will be available regarding the type of guitar and/or playing style and/or pickup used. A learn function will be available for each playing techniques if predefined settings don't fit the user.

Time Discretization. One of the key element when transcribing what an instrumentalist is playing is the notion of timing and duration of the notes. As digital notes and amplitudes, digital rhythms and durations need to be quantized. Present in most of the music software, MIDI transport is the norm used for discrete time representation. It has to be noticed here, that as a VST plugin, our detection software is meant to be incorporated into existing audio software,

using, as a matter of fact, its host MIDI transport function.However not all audio software use MIDI transport and it seemed relevant to also include our own time quantization module. In the future, the software will allow the user to choose between an internal generation of discrete time and the one from its host.

For the moment, timing information is output as float values with *noteOn* and *noteOff* OSC messages. A subdivision parameter is accessible under the General tab in order to define the smallest quantization unit for notes duration (e.g, quarter, eight, sixteenth, etc.).

Preset. The preset system of EGT is handling 2 elements:

- a complete back-up of the system: all modified or created elements (input gains, detectors, behaviours, etc.) are saved.
- combination of those elements (preset) can as well be defined and managed (added, deleted or modified).

A XML file format was used to serialize the preset data, and the open source TinyXML[3] library has been implemented to enable this functionality.

Region and Behaviour. A six strings real-time playing techniques detection system can be CPU consuming and may be too much information to work with, depending on the situation. The two remaining tabs, Regions and Behaviours, enable the user to define, respectively, **where** on the fretboard he wants the detected events to be available and **what** he wants to detect. The idea behind these two tabs is to filter the flow of detection to what was really useful for the user. In both of our application cases however, all detections have been used.

The region concept was already prototyped in [7] and was expanded in this project as two types of regions can be defined: picking regions and fretboard regions. Picking regions are linked to plucking point detection. Three picking regions are defined by default (bridge, soundhole and neck) but others can be graphically created and spread all over the neck. A fretboard region is simply defined in a global sense as a **group of notes** which can be further characterized as chord, arpeggio or free depending on the time between each note of the group.

Figure 1 shows an example of behaviour and the region on which the behaviour is applied. To define a behaviour, the filtered playing techniques need to be selected (**what**) as well as the picking and fretboard regions (**where**) previously defined. MIDI and OSC outputs are available for the filtered events. Finally, a VST audio effect plugin can be chosen to be activated each time the behaviour is detected.

The combination of the region and the behaviour tabs can be used in many different contexts. A remote control zone, e.g, could be defined on specific notes (especially on hard to reach high end notes of the lower strings, but not only) in order to change presets or effects depending on the use case. Another example could be the one of different places on the fretboard linked to specific sounds or effects: the first five frets on all the strings could be setup with an overdrive effect and notes between the 12th and 15th fret on the three high strings could

[3] `http://www.grinninglizard.com/tinyxml/`

be setup with a reverb effect which amount could be controlled by the amount of bend. One last example could be chord recognition: as chords can be defined as a group of notes played under a defined amount of time, a specific chord is to be detected if the notes are played in the right order (upward or downward).

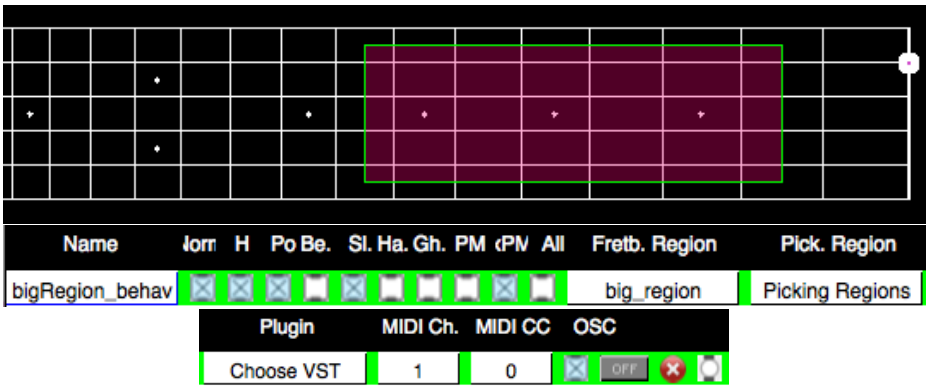


Fig. 1. This behaviour filters normal notes as well as hammer-on, pull-off and slide techniques performed in the fretboard region showed on the top part of this graphic

3 Use Cases

Two use cases are demonstrated here: a real-time and complete guitar score transcription in[4]TuxGuitar, an open-source tablature software and a real-time manipulation of a 3D representation of a guitar fretboard done in[5]openFrameworks.

3.1 TuxGuitar Input: Real-Time Score Writing

Tablature is the most common type of guitar score available to guitarists due to its easy readability. Indeed, on tablature, notes are represented by a number of fret on a specific string. All the guitar's specific techniques have their own representation as well. We decided to build a plugin for TuxGuitar tablature software which would receive OSC messages from EGT and display them on the tablature editor. This opensource software was chosen among others (i.e, KGuitar and DGuitar) because of its continuity in terms of development as well as its rich plugin API for which previous code examples were available. [6]MusicXML format has been investigated as well. Although it is implemented in several notation software i.e Finale, Sybellius, or specific guitar tablature

[4] http://tuxguitar.herac.com.ar/
[5] http://www.openframeworks.cc/
[6] http://www.musicxml.com/

ones, i.e Guitar Pro, the implementation quality appeared uneven, and we hence decided to postpone the integration. The Guitar Pro file format may also be consider in future development.

The current TuxGuitar plugin implementation gives a first basic framework whose accuracy result tends to depend on the complexity of the played track as well as on the quality of the interpretation. It has to be noticed that no precise study has yet been performed on the accuracy of the system. From the first tryouts though, a stronger and global robustness step is needed regarding the time management and events detection. TuxGuitar software may as well be a source of mistakes regarding the display as the real-time entry plugin we built is more a hack than a native API function.

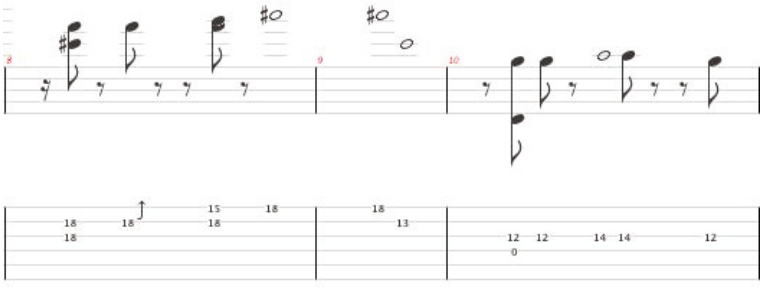


Fig. 2. Some of the first measures of Shine On You Crazy Diamond by Pink Floyd. The detected guitar playing events have been performed by EGT.

3.2 OpenFrameWorks: Control of a Guitar Neck 3D Representation

For this use case, we wanted to have an aesthetic representation of a guitar neck which would react to EGT's detected events. An OSC receiver has been implemented in this OpenFrameWorks program in order to interpret EGT's messages and make them interact with the 3D model. The model is based on an hexagonal tube, representing the fretboard, inside which hexagonal planes represents frets. A plucked note is represented by a burst on the string whose amplitude depends on guitar's notes amplitude. When a note is plucked the tube turns on itself so that the string corresponding edge appears in the foreground. The different playing techniques have their own graphical representations. Examples of this interaction can be found[7]online.

4 Conclusion and Perspectives

EGT is a guitar transcription software upon which higher-level tools can be developed. It detects the following techniques: hammer-on, pull-off, harmonic, slide, bend, palm muting, as well as note start and end timings, pitch and plucking point. Additionally, a concept of regions and behaviors has been designed in

[7] `https://vimeo.com/34504237`

order to filter a set of guitar events happening on a specific place of the guitar fretboard. The software has been wrapped into a VST plugin and the detected events can be sent through OSC or in a reduced version through MIDI. Two use-cases have been presented: a real-time tablature writer and a 3D guitar neck representation animated using parameters from the detected events.

This tool appears to be a first good base for guitar's playing transcription. However, a precise user study with different playing styles needs to be done to make the detection system more robust and adapted to the instrumentalist's playing. EGT is also a first step towards higher-level concepts: one can, e.g, easily imagine a complete musical phrase record system to help studying and analyzing the playing at a macro structure level. More globally, these detections open up to a substantial study on mappings, to see how they can be used to interact with other media (e.g, sound or video synthesis parameters, light control, etc.). A lot of strategies can now be investigate to make the control of guitar effects evolve to a more refine and instrumentalist linked control.

Acknowledgments. Numediart is a long-term research program centered on Digital Media Arts, funded by Région Wallonne, Belgium (grant N°716631).

References

1. Barbancho, A.M., Klapuri, A., Tardon, L.J., Barbancho, I.: Automatic transcription of guitar chords and fingering from audio. Trans. Audio, Speech and Lang. Proc. 20(3), 915–921 (2012)
2. Fiss, X., Kwasinski, A.: Automatic real-time electric guitar audio transcription. In: Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing, ICASSP 2011, Prague Congress Center, Prague, Czech Republic, May 22-27, pp. 373–376. IEEE (2011)
3. Graham, R.: A live performance system in pure data: Pitch contour as figurative gesture. In: Proc.of Pure Data Convention, Bauhaus-Universität, Weimar, Germany (August 2011)
4. Guaus, E., Ozaslan, T., Palacios, E., Ll Arcos, J.: A left hand gesture caption system for guitar based on capacitive sensors. In: Proc. of NIME (2010)
5. Lähdeoja, O.: An approach to instrument augmentation: the electric guitar. In: Proceedings of the 2008 Conference on New Interfaces for Musical Expression, NIME 2008 (2008)
6. Penttinen, H., Välimäki, V.: Time-domain approach to estimating the plucking point of guitar tones obtained with an under-saddle pickup. Applied Acoustics 65(12), 1207–1220 (2004)
7. Reboursière, L., Frisson, C., Lähdeoja, O., Mills III, J.A., Picard, C., Todoroff, T.: Multimodal guitar: A toolbox for augmented guitar performances. In: Proc. of NIME (2010)
8. Reboursière, L., Lähdeoja, O., Drugman, T., Dupont, S., Picard, C., Riche, N.: Left and right-hand guitar playing techniques detection. In: Proc. of NIME (2012)
9. Traube, C., Depalle, P.: Extraction of the excitation point location on a string using weighted least-square estimation of comb filter delay. In: Proceedings of the Conference on Digital Audio Effects, DAFx (2003)

Performative Voice Synthesis for Edutainment in Acoustic Phonetics and Singing: A Case Study Using the "Cantor Digitalis"

Lionel Feugère[1,2], Christophe d'Alessandro[1], and Boris Doval[3]

[1] LIMSI-CNRS, F-91403 Orsay Cedex, France
[2] UPMC Univ Paris 06, F-75005 Paris, France
[3] Equipe lutheries - acoustique - musique, UPMC Univ Paris 06, UMR 7190, Institut Jean Le Rond d'Alembert, F-75005 Paris, France

Abstract. A real-time and gesture controlled voice synthesis software is applied to edutainment in the field of voice pedagogy. The main goals are teaching how voice works and what makes the differences between voices in an interactive, real-time and audio-visual perspective. The project is based on "Cantor Digitalis", a singing vowel digital instrument, featuring an improved formant synthesizer controlled by a stylus and touch graphic tablet. Demonstrated in various pedagogical situations, this application allows for simple and interactive explanation of difficult and/or abstract voice related phenomena, such as source-filter theory, vocal formants, effect of the vocal tract size, voice categories, voice source parameters, intonation and articulation, etc. This is achieved by systematic and interactive listening and playing with the sound of a virtual voice, related to the hand motions and dynamics on the tablet.

Keywords: edutainment, voice synthesis, performative synthesis, graphic tablet.

1 Introduction

Like for any kind of knowledge to be taught, the learning process becomes easier and more entertaining when using various media as teaching materials. Besides, if the students can interact with the media in real-time, success is almost guaranteed. How voice works can be one of this knowledge to be taught.

Splitting a system into several subparts can help for understanding it. Concerning the voice, two observations can be noted. First, a part of the organs of the vocal apparatus is hidden from outside, and it may be dangerous to modify it further than what we can do naturally for understanding its behaviour. Second, despite it is of everyday usage, voice is a complex instrument which involves abstract concepts difficult to understand by the general public.

In this paper, an interactive and real-time application is presented, based on the Cantor Digitalis [1], a musical instrument for singing vowels synthesis implemented in Max/MSP [2].

M. Mancas et al. (Eds.): INTETAIN 2013, LNICST 124, pp. 169–178, 2013.

Owing to a signal-type approach and a physically meaningful mapping, it is easy to modify a large number of high level model parameters. The parallel formant [1] synthesis and the source filter model used in this synthesizer allow us to deconstruct the voice model to listen to abstract acoustic phenomena such as individual formants or vocal fold sounds.

A few voice models are available for teaching purposes, such as VocalTractLab [3] or Benoit Project [4]. Their synthesis method can not allow to listen to abstract phenomena, as they are often based on physical models of the vocal apparatus, then reflecting concrete physical phenomena. Also, the original feature of our application is above all the capability to play with the model in real-time and then to listen to the dynamic transformation of the vocal tract [2] and/or the glottal source [3]. The initial goal of the Cantor Digitalis is music, so we can easily use it in an entertaining way for pedagogical purposes by using its control interface while modifying the model parameters and listening to the effects.

After presenting the Cantor Digitalis instrument, we will deal with the decomposition of the source-filter model for a given voice and explain how we use it for pedagogical purposes. Then, still in a real-time interactive perspective, a voice is transformed from one to another through continuous and reactive transformation. We will finally conclude by the main contributions and the perspectives of our work.

2 Cantor Digitalis: Performative Singing Synthesis

Cantor Digitalis is a digital musical instrument allowing for control of pitch, vowel color, strength and quality of a synthetic voice model. Synthesis is controlled in real-time, like a musical instrument, hence the expression "performative synthesis". A general view of the instrument is given at the figure 1. It is based on an improved formant synthesis model, bi-manually controlled by the position and pressure of a stylus and a finger over a graphic tablet. It has been used in concerts within the Chorus Digitalis ensemble [1] [5].

2.1 Formant Synthesis Using the RT-CALM Source Model

The production model of Cantor Digitalis uses the source-filter theory: the glottal source flow is modeled using the RT-CALM [6] and the vocal tract resonances by parallel bandpass filters.

RT-CALM is a real-time version of the CALM model [7]. The spectral properties of the glottal flow model (GFM) are described by the shape of the spectrum derivative: a resonance in low frequencies, and a spectral slope for mid and high frequencies. A white noise modulated by the shape of the GFM represents voice

[1] A formant is the result of one or several vocal tract resonances that contribute to the perception of a vowel.

[2] The shape of the vocal tract changes with the location of the articulators (lips, tong, jaws, ...)

[3] The glottal source is the sound signal created just above the vocal folds.

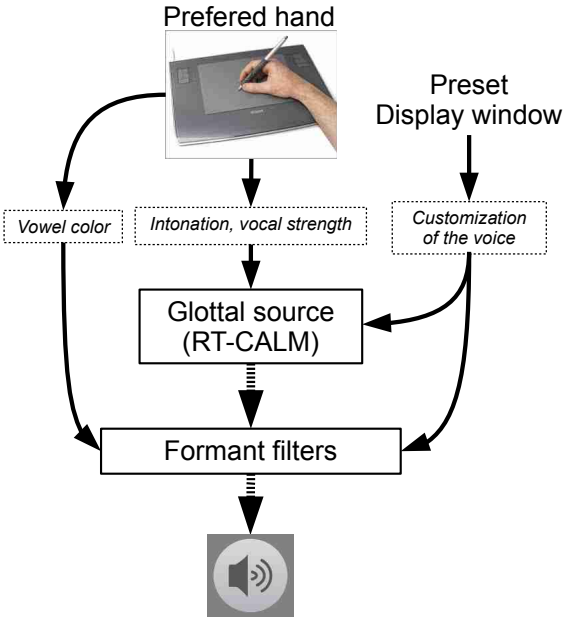


Fig. 1. Functional diagram of Cantor Digitalis

aspiration. Voice strength is realized by an increase in signal intensity, a decrease in the spectral slope of the GFM derivative, and a position shift of the spectrum low frequency maximum toward higher frequencies.

Five bandpass filters model the vocal tract resonances and a bandstop filter the anti-resonance of the piriform sinus [8]. The bandpass filters represent the formants of the vowels. Thus, a database of formants defines the vowels by a set of amplitude / bandpass / frequency values. A scale factor is applied to the formant central frequencies to model the vocal tract size. Then, the different singer types (bass, tenor, alto, soprano) are only characterized by two parameter shifts in our model: this global scale factor and the pitch range.

2.2 Source-Filter Interactions and Automatic Source Perturbation

For singing, a number of rules has been added to the source-filter model, concerning source-filter interactions and automatic source perturbation.

Source-filter interactions are modeled by specific dependencies between:

– Fundamental frequency F_0 and formant central frequencies F_i (i indicates the rank of the formant from the lowest to the highest central frequency).

The literature on acoustics shows that singers modify the frequencies of their formants to adapt them with F_0, in particular F_1 and F_2 are increased

with F_0 in the upper part of their range so that they remain greater than F_0 [9]. Indeed, formants are the key of the voice sound level.

– Voice effort and the first formant central frequency F_1

  It has been demonstrated [10] that F_1 increases by 3.5 Hz/dB between soft and loud voice, or approximately 50 Hz over 15 dB range. For $F_1 = 600$ Hz, a scaling factor proportional to vocal effort can be applied to F_1 , in order to get a 10% increase of F_1 from soft vocal effort to maximum vocal effort (voice effort parameter approximately evolves linearly with sound level).

Automatic source perturbation are divided into deterministic and non deterministic perturbations. We implemented the deterministic heart pulse perturbation on amplitude and frequency of the glottal source vibration, as demonstrated by Orlikoff [11]. They showed that vocal sound pressure and fundamental frequency across a heart cycle has a deterministic perturbation component and that its deviation depends on vocal effort: 14% (soft), 8% (moderate), 3% (loud) for the amplitude variation; 1.4% (soft, moderate), 0.8% (loud) for frequency variation. Along a heart cycle, the perturbation looks like very coarsely a damped sinusoid around the average value and then was modeled as such. The sound result is a small perturbation of pitch and sound level giving more naturalness to the synthetic voice.

Among the non deterministic perturbation are jitter and shimmer which are modeled by a white noise perturbation over the amplitude and frequency of the GFM fundamental pattern.

2.3 Chorus Digitalis: Choral Performative Synthesis

Cantor Digitalis is controlled by one or two graphic tablets. A pen is used to control the pitch very accurately along the X-dimension with the preferred hand, taking advantage of our writing skills. The stylus pressure over the tablet is mapped to voice effort. The tablet is visually augmented with a continuous pitch keyboard to help playing accurately.

The vowel color has been mapped in several ways. By using a second tablet or a part of the main tablet, we can control the vowels in a 2D space with the second hand (figure 2), where four vowel formants are interpolated to get a continuous vocalic space. An alternative to bi-manuality is to control the vowel color along the Y-axis of the single tablet with the preferred hand.

The Chorus Digitalis musical ensemble has performed several times, using 1 to 6 Cantor Digitalis, each controlled by a musician and with a customized voice. The figure 3 represents the Chorus Digitalis publicly performing in May 2012.

3 Edutainment: The Source-Filter Model of Vocal Production

After having designed an application for gestural control of voice synthesis for some years, we realized that in order to teach how the voice works, a good way was to *deconstruct* what we have done.

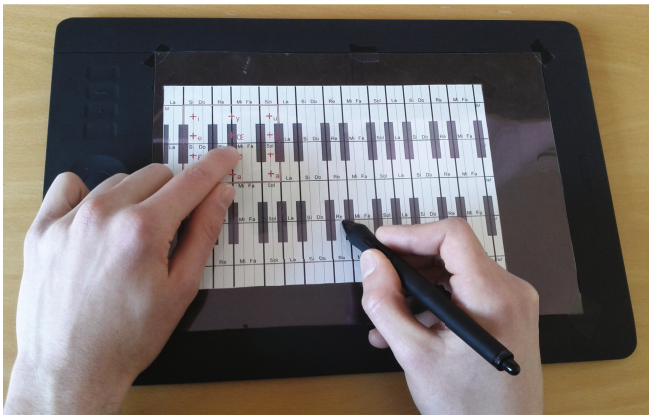


Fig. 2. The bi-manual controller

Fig. 3. The Chorus Digitalis in concert

We implemented a software interface to easily listen to certain parts of the voice model. The figure 4 is a screen capture of the software interface. It is separated into a clickable area and a visualization area. The clickable area allows one to build or deconstruct a voice using the source-filter approach. The visualization area displays a real-time spectrum of the output, i.e. what is heard. Below are detailed possible uses of the software for teaching purposes.

3.1 Playing with the Glottal Source

Pitched voice source is an acoustic effect of the vibration of the vocal folds. However, this intrinsic sound is never heard separately, as the voice sound is the convolution of the source and the vocal tract. Besides, it is impossible to remove

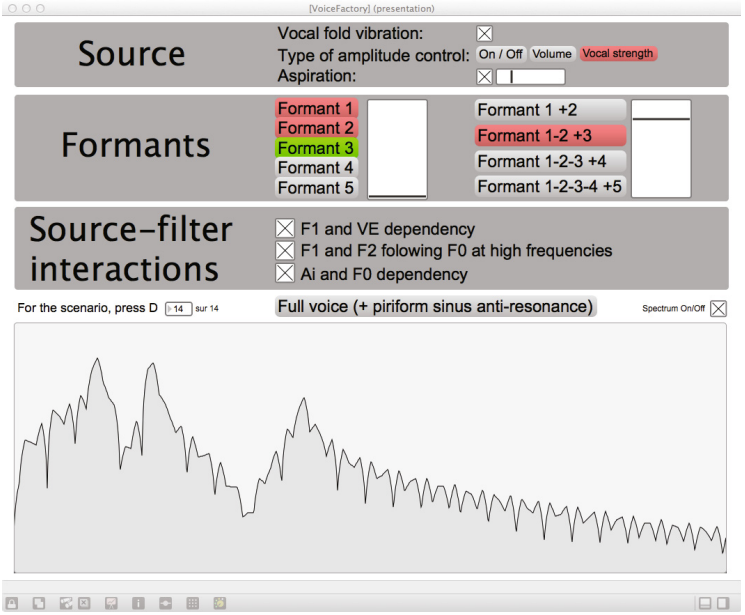


Fig. 4. Screen capture of the "voice factory" interface

the vocal tract, whereas by analogy it would be possible to do it with a saxophone for example, where the mouthpiece can be removed from the instrument body.

With a voice model using the source-filter theory, it is possible to listen to the source, as if the vocal tract had been removed. Then, by using the same interface as with the Cantor Digitalis, the player can listen to how the glottal source sounds. The glottal control is limited to pitch and vocal effort. Then changing vocal tract configuration (vowels) does not affect the sound.

The software interface allows for choosing the mapping between the stylus pressure and the voice intensity, and then to test it while listening to the different mapping: an on/off mapping depending on stylus contact with the tablet; a volume mapping, by linearly controlling the signal amplitude with the stylus pressure; a more natural control, vocal effort mapping, which acts on amplitude and spectral slope in high frequencies.

Finally, the source noise, due to turbulent flow around the vocal fold, can be added to the harmonic source sound or can be listened independently, an effect hardly achieved with a real voice.

3.2 Listening to Individual Formant Motions along Articulatory Trajectories

Using the glottal source sound, the interface allows for continuously moving each formant to listen to its filtering effect. Formant is a common term in language science, but it can be difficult to understand, above all for students who are not

familiar with signal processing. Moreover, an isolated formant never occurs in speech, as the global formants pattern results of the global shape of the vocal tract. Then it is difficult to identify each formant in the sound of a real voice.

While listening to one of the 5 formants, the vocal tract can be modified by exploring the vowel space and then the contribution of this formant to the chosen vowel can be identified. The real-time control of the center frequency of the formant allows to roughly identify articulatory movements: jaw aperture to formant 1, tongue position to formant 2, lip aperture to formant 3.

3.3 Combining Formants to Make Vowel Identification Emerging

After having listened to each formant independently from each other, the formants can be added one by one from the one with the lowest to the highest central frequency. Each additional formant is continuously controlled owing to a slider to highlight the sound effect. The addition of the 2^{nd} formant to the 1_{st} one allows for identification of almost all vowels. The addition of the 3_{rd} formant mainly resolves ambiguities between a few high vowels. The 4^{th} and 5^{th} formant improve voice quality, but without changing the vowel quality.

It is very interesting to listen to the emergences of human voice and vowels, while looking at the spectrum evolution. It really enables to understand the link between sound and spectrum, through concept of resonances/formants.

The harmonic contribution of the glottal source sound can be removed, allowing to listen to the effects of the formants on the turbulent noise of the source, while moving into the vowel space.

3.4 Synchronizing the Voice Source and Vocal Tract Motions

This software has been used several times with general public (science festival, science & music day, ...) and in classes with scientific students.

Besides the above presented applications, it was sometimes used to illustrate the coordination task of the glottal source and the vocal tract to produce small words with a given expression. A person was asked to control the glottal source (pitch and vocal effort) while an other had to control the vocal tract (vowel color). They were asked to reproduce short expressive words like *Oh yeah* or *Oui-oui* (*Yes-yes* in French): first by analysing the evolutions of the pitch and the vowel color; second by trying to reproduce the appropriate gesture; last by synchronizing their gestures to produce the targeted expressive word. This allows people to understand how their own voice is produced by controlling a synthesized voice. The enthusiasm was clear and the users understand fast how to improve their first trial and to generalize the method to produce other short words and different expressions.

The source-filter interaction presented in the section 2.2 can be switched on or off, in order to appreciate its effect.

4 Edutainment: Voice Settings and Individualization

In the preceding section, we presented the way of building / deconstructing a given voice to teach how the vocal apparatus works. Now, we are talking about how to use voice individualization for teaching purposes.

For this sake, we use the voice individualization window, a part of the Cantor Digitalis application. The window interface, as shown in figure 5 is divided into presets and manual settings. The upper part concerns voice types and the bottom part concerns formant values of the vowels. All these settings apply immediately to the voice sound.

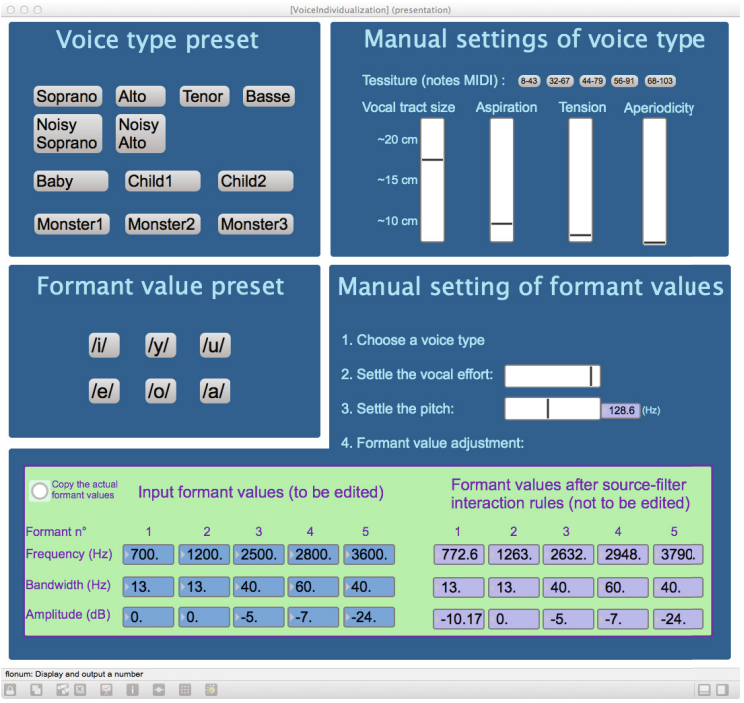


Fig. 5. Screen capture of the individualization voice interface

4.1 Adjustment of Formants

A basic set containing 6 vowels is available. By interpolations of these six vowels, continuous intermediate vowels are obtained, using gestural control over the 2-dimensional plan of the graphic tablet.

Each bandpass filter, representing a formant, can be manually adjusted through its central frequency, bandwidth, and amplitude. Then 3 values for the 5 formant filters can be set independently at the bottom left of the interface window, and the effective values (after the source filter interaction) are given at bottom right. The modification is realized in real-time but it is not possible

for the user to play this new vowel with the tablet (pitch and vocal effort are controlled via a slider). This can be used to demonstrate that different vowel colors (i.e. different formant configurations) can be perceived as a same vowel, an effect demonstrating the categorical perception of vowels.

4.2 Vocal Tract Size

The smaller the vocal tract, the greater the frequency resonances, and then the greater the central frequency of the vocal tract formants. Then the central frequency of its formants will be larger. Thus, we can change the apparent vocal tract size of our synthesized voice by a common multiplier applied to the central frequencies of the formant filters. In real-time, we can pass from a baby voice (small vocal tract) to an adult voice (large vocal tract). This is even more effective if the pitch range consistently decreases with an increase of the vocal tract size.

4.3 Voice Quality Parameters

Several voice qualities parameters are available for controlling the glottal source model, such as aspiration noise rate, voice tension (inversely proportional to the frequency of the glottal flow spectrum maximum) or aperiodicity (jitter/shimmer of the vocal fold vibration).

4.4 Beyond Human Voice

By modifying all these parameters in real-time, it is easy to individualize the synthetic voice by trial and error. It is possible to demonstrate the vocal apparatus similarities between humans and some animals, by changing the vocal tract size, aspiration noise and vocal fold tension. An impressive example is obtained by the transformation of human voice into a beast roar by increasing a lot the vocal tract size, adding aspiration and vocal tension.

5 Conclusion

We presented an application derived from the Cantor Digitalis instrument and intended for acoustic phonetics teaching. It allows to use gestural control to interact with a real-time building and individualization of a voice model.

Two main paths can be explored: 1. How does the voice basically work using source-filter model approach? 2. What are the differences and similarities between human voices (or even some animal voices)?

Abstract phenomena like formants, vowel identification, voice decomposition into source and filter can be understood with the help of audio and gestural real-time interaction.

This application has already been demonstrated in several edutainment contexts like general public science festivals or during university classes. But no proper objective or subjective evaluation has been done yet.

6 Perspectives

Being first of all intended for voice teachers and in order to help as many teachers/students in voice education as possible, the next step would be to make the application available as free or open-source software. Now, the software is not yet available.

New features could be added to the present prototype, according to the needs expressed by users like voice or music teachers.

Finally, this type of performative synthesis is ideally suited to the creation and design of new voices ("human" or "monster-like") in a fast way, because of the real-time response of the system to any parameter modification. A specific and powerful feature of performative synthesis is its ability to play with vocal dynamics and vocal motion, that are often the key for natural sounding voice synthesis.

References

1. Feugère, L., Le Beux, S., d'Alessandro, C.: Chorus digitalis: polyphonic gestural singing. In: 1st International Workshop on Performative Speech and Singing Synthesis, Vancouver (2011)
2. Max/MSP, `http://cycling74.com/products/max/` (visited on February 28, 2013)
3. Vocal Tract Lab, `http://www.vocaltractlab.de/` (visited on February 28, 2013)
4. Benoit Project, `http://benoit.susannefuchs.org/tutorial3.html` (visited on February 28, 2013)
5. Le Beux, S., Feugère, L., d'Alessandro, C.: Chorus digitalis: experiment in chironomic choir singing. In: Proceedings of the Conference on 12th Annual Conference of the International Speech Communication Association (INTERSPEECH 2011), Firenze, Italy, August 27-31, pp. 2005–2008 (2011) ISSN: 1990-9772
6. d'Alessandro, N., d'Alessandro, C., Le Beux, S., Doval, B.: Real-time calm synthesizer: new approaches in hands-controlled voice synthesis. In: Proc. of New Interfaces for Musical Expression, Paris, pp. 266–271 (2006)
7. Doval, B., d'Alessandro, C., Henrich, N.: The voice source as a causal/anticausal linear filter. In: Proceedings of Voqual 2003: Voice Quality: Functions, Analysis and Synthesis, ISCA (2003)
8. Dang, J., Honda, K.: Characteristics of the piriform fossa in models and humans. J. Acoust. Soc. Am. 101(1), 456–465 (1997)
9. Henrich, N., Smith, J., Wolfe, J.: Vocal tract resonances in singing: Strategies used by sopranos, altos, tenors, and baritones. J. Acoust. Soc. Am. 129(2) (2011)
10. Liénard, J.-S., Di Benedetto, M.-G.: Effect of vocal effort on spectral properties of vowels. J. Acoust. Soc. Am. 106(1), 411–422 (1999)
11. Orlikoff, F.R.: Heartbeat-related Fundamental Frequency And Amplitude Variation In Healthy Young And Elderly Male. Journal of Voice 4(4), 322–328 (1990)

MAGEFACE: Performative Conversion of Facial Characteristics into Speech Synthesis Parameters

Nicolas d'Alessandro, Maria Astrinaki, and Thierry Dutoit

Institute for New Media Art Technology, University of Mons
Signal Processing Laboratory, 31 Boulevard Dolez, B-7000 Mons
nda@numediart,{maria.astrinaki,thierry.dutoit}@umons.ac.be

Abstract. In this paper, we illustrate the use of the MAGE performative speech synthesizer through its application to the conversion of realtime-measured facial features with FaceOSC into speech synthesis features such as vocal tract shape or intonation. MAGE is a new software library for using HMM-based speech synthesis in reactive programming environments. MAGE uses a rewritten version of the HTS engine enabling the computation of speech audio samples on a two-label window instead of the whole sentence. Only this feature enables the realtime mapping of facial attributes to synthesis parameters.

Keywords: speech synthesis, software library, performative media, streaming architecture, HTS, MAGE, realtime audio software, face tracking, mapping.

1 Introduction

Speech is the richest and most ubiquitous modality of communication used by human beings. Through vocal expression and conversation, we realize a complex process, highly interactive and social. For decades, algorithms for the production of synthetic speech have focused on converting a static text into an intelligible and natural waveform, lately with great success. Ten years ago, the trend of creating expressive or emotional speech brought researchers to realize that such properties were not only a matter of sound quality. Expressivity in speech is contextual, interactive, social, coming in response to other ongoing processes, reaching across most of other human being modalities, and therefore other disciplines. Through this large interdisciplinary redefinition of expressive speech, one property seems to emerge and reach some consensus: speech is a performance; it starts with a gesture and ends up as a message, conveying both informative and affective contents.

As these new trends in understanding expressivity in speech are being explored, one might notice that a real solid platform is missing. Indeed Text-To-Speech (TTS), as a platform, has been tackling and greatly solving other problems: similarity between original and synthetic waveforms, segmental and supra-segmental qualities, intonation modeling, etc. However most of existing TTS systems require a significant amount of text in advance (typically a sentence) and process it into sound as a whole target. Most of the time, the ability to influence the synthesis process has been

M. Mancas et al. (Eds.): INTETAIN 2013, LNICST 124, pp. 179–182, 2013.

limited, disabled or discouraged, as the resulting sound quality quickly degrades. If we consider that expressivity is related to the ability to interact with the artificial speech production process at various production levels and time scales, as it would happen with real speech, then the requirements for such a platform are different: we need a so-called reactive programming architecture, applied to speech synthesis.

2 Performative HMM-Based Speech Synthesis

For the last decade, HMM-based speech synthesis has been constantly improving and became a serious alternative to non-uniform unit selection (NUU), especially when a more lightweight and flexible synthesis engine is required. Particularly, the HTS system [6] is now reaching a reasonably high synthesis quality. Moreover, the model-based approach used in HTS to generate the speech production parameters enables a whole new category of techniques, such as speaker adaptation, speaker interpolation, voice cloning, voice reconstruction, etc.

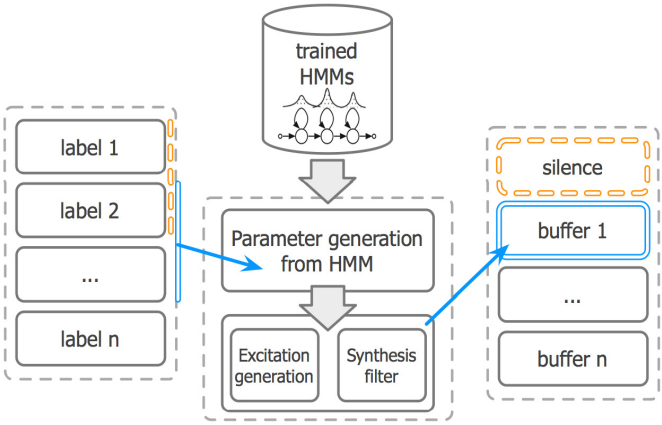


Fig. 1. Performative implementation of HTS: speech signal is computed on a sliding two-label window instead of the whole sentence. This enables the parameters to be changed on the fly.

One fundamental aspect of HTS is the use of context-dependent HMMs. Before the statistical models are trained, each phoneme is characterized by its linguistic context, e.g. previous phoneme, next phoneme, current syllable, previous syllable, next syllable, etc. The trend in synthesizing natural speech trajectories from text has led to add as much phonetic context as possible, to capture variations that are encountered in real speech. As a consequence, the need for a large look-ahead in the future (next phoneme, next syllable, next word) has brought the accessible time scale at run- time in HTS to the current sentence. Therefore, scenarios such as starting to synthesize a sentence with one speaking style and terminating that ongoing sentence with another speaking style based on an unpredictable user control command is impossible.

2.1 Performative Implementation of HTS (pHTS)

In pHTS, we have developed a series of modifications, enabling a much more reactive control of speech parameter trajectories. The main modification is the optimization of the generation of speech parameters on a sliding window of 2 labels rather than on the whole sentence, as shown in Fig. 1. When speech parameters have been generated for such a 2-label window, audio samples corresponding to the past label can be synthesized right away. If these samples are used within a realtime audio architecture, it means that modifications achieved on pHTS models will have an impact on the ongoing speech audio output with a delay of only one label.

2.2 MAGE: Flexible API for Speech Synthesis

MAGE is the software umbrella that provides the appropriate real- time audio architecture, in order to plug the pHTS speech synthesis engine. Indeed it schedules the various tasks encountered in the pHTS synthesis, so that the sound is constantly synthesized from an ongoing stream of asynchronous user-provided phonemes:

- on the fly generation of label-formatted streams;
- scheduling of model selection from the database;
- scheduling of speech parameters generation;
- scheduling of MLSA filtering.

The MAGE framework transparently uses concurrent programming techniques in order to guarantee the reactivity and flexibility of the application to unpredictable inputs, such as new labels, modification of F0 models, duration models or MGC models. MAGE is also the opportunity to encapsulate the HTS functionalities into a user-friendly API, so that more developers can integrate our new speech synthesis features in their applications.

3 Prototypes

MAGE as a software library has already been used in various prototypes. These prototypes tend to highlight a common aspect of our research work: exploring how HMM-based speech synthesis can be gesturally controlled. The concept of gesture or performance is here considered in a very large sense. Indeed we work with any user input that can have a meaningful impact on speech production properties. Our primary interest is the manipulation of speech phonemes and prosody with hand gestures, as in [5] and [1], but here we consider more indirect causes, such as facial expression.

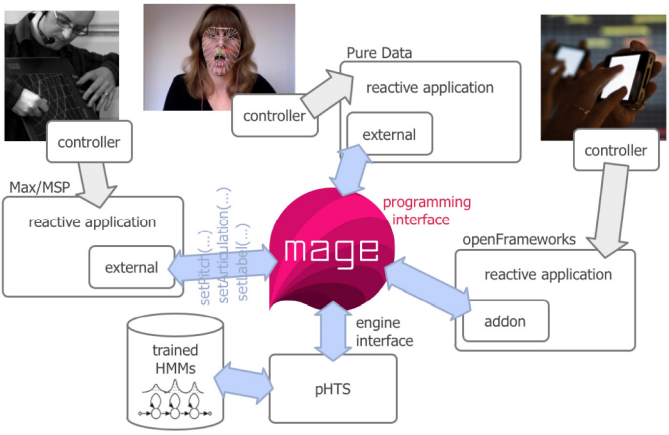


Fig. 2. Various scenarios of MAGE used to create a speech synthesis external object or add-on, as to enable the use of various controllers for impacting on the speech synthesis output

Fig. 2 gives an overview of integrating MAGE into various software environments, with the aim of connecting controllers to the speech synthesis. We can highlight various prototypes that have already been built, following this process. Firstly, the HandSketch musical instrument [1], formerly used for singing synthesis, is now able to change speaking speed and intonation with pen or finger gestures [3]. We can demonstrate a version in Max/MSP with a tablet and an iPhone version. We also built a virtual speaker, triggering syllables directly from mouth motion, using a face tracking software [2] in openFrameworks and synthesizing the speech in PureData. Face features come from FaceOSC [7]. We use various mouth opening sequences and eye broses position to influence the speech intonation and the vocal tract length.

References

1. d'Alessandro, N., Dutoit, T.: HandSketch Bi-Manual Controller: Investigation on Expressive Control Issues of an Augmented Tablet. In: Proc. International Conference on New Interfaces for Musical Expression, pp. 78–81 (2007)
2. MAGE and Face Tracking, `https://vimeo.com/39567236`
3. MAGE and HandSketch, `https://vimeo.com/39558917`
4. MAGE website, `http://mage.numediart.org`
5. Nordstorm, K., et al.: Developing Vowels Mappings for an Interactive Voice Synthesis System Controlled by Hand Motions. Journal of the Acoustical Society of America 127, 2021 (2010)
6. Zen, H., Tokuda, K., Black, A.: Statistical Parametric Speech Synthesis. Speech Communications 51(11), 1039–1064 (2009)
7. FaceOSC, `https://vimeo.com/26098366`

Multimodal Analysis of Laughter for an Interactive System

Jérôme Urbain[1], Radoslaw Niewiadomski[2], Maurizio Mancini[3], Harry Griffin[4], Hüseyin Çakmak[1], Laurent Ach[5], and Gualtiero Volpe[3]

[1] Université de Mons, Place du Parc 20, 7000 Mons, Belgium
jerome.urbain@umons.ac.be
[2] LTCI UMR 5141 - Telecom ParisTech, Rue Dareau, 37-39, 75014 Paris, France
[3] Università degli Studi di Genova, Viale Francesco Causa, 13, 16145 Genova, Italy
[4] UCL Interaction Centre, University College London,
Gower Street, London, WC1E 6BT, United Kingdom
[5] LA CANTOCHE PRODUCTION, rue d'Hauteville, 68, 75010 Paris, France

Abstract. In this paper, we focus on the development of new methods to detect and analyze laughter, in order to enhance human-computer interactions. First, the general architecture of such a laughter-enabled application is presented. Then, we propose the use of two new modalities, namely body movements and respiration, to enrich the audiovisual laughter detection and classification phase. These additional signals are acquired using easily constructed affordable sensors. Features to characterize laughter from body movements are proposed, as well as a method to detect laughter from a measure of thoracic circumference.

Keywords: laughter, multimodal, analysis.

1 Introduction

Laughter is an important signal in human communication. It can convey emotional messages, but is also a common back-channeling signal, indicating, for example, that we are still actively following the conversation. In dyadic conversations, each participant laughs, on average, every 2 minutes [1]. Recent works have also discovered the positive impact of a laughing virtual agent on users experiencing human-machine interactions [2].

Our long-term objective is to integrate laughter into human-machine interactions, in a natural way. This requires building an interactive system able to efficiently detect human laughter, analyze it and synthesize an appropriate response. The general system architecture of our application is displayed in Figure 1. We distinguish 3 types of components: input components, decision components and output components.

The input components are responsible for multimodal data acquisition and real-time laughter analysis. In our previous experiments [2], only the audio modality was used for laughter detection. This resulted in two types of detection errors: a) false alarms in presence of noise; b) missed detections when the

M. Mancas et al. (Eds.): INTETAIN 2013, LNICST 124, pp. 183–192, 2013.

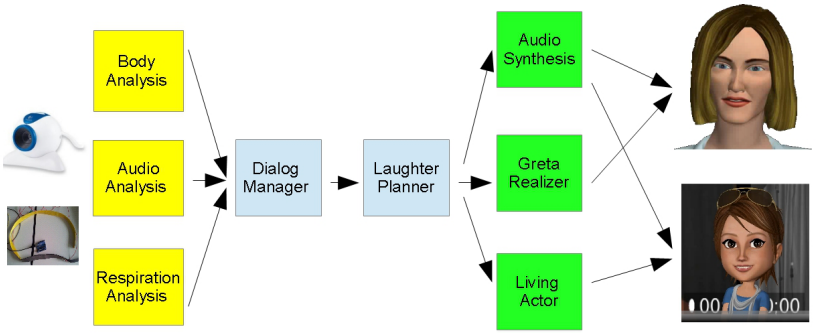


Fig. 1. Overall architecture composed of input components (in yellow), decision components (in blue) and output components (in green)

laugh is (almost) silent. This is why in this work we are introducing new modalities to make the laughter detection more robust. The input components now include laughter detection from body movements and respiration in addition to audio detection and intensity estimation. The data on user behavior (see Table 1) are captured with two devices: a simple webcam and the respiration sensor developed at University College London (see Section 4).

Table 1. Recorded signals

Recording device	*Captured signal*	*Description*
Webcam	Video	RGB, 25 fps
	Audio	16 kHz, 16 bit, mono
Respiration Sensor	Respiration	120Hz, 8 bit

The laughter-enabled decision making modules decide, given the information from the input components, when and how to laugh so as to generate a natural interaction with human users. At the moment, two decision components are used to decide the agent audiovisual response. The first one (Dialog Manager) receives the information from the input components (*i.e.*, laughter likelihoods and intensity) as well as contextual information and it generates the instruction to laugh (or not) with high-level information on the laugh to produce (*i.e.*, its duration and intensity). The second component, Laughter Planner, controls the details of the expressive pattern of the laughter response by choosing, from the lexicon of pre-synthesized laughter samples, the most appropriate audiovisual episode, *i.e.* the episode that best matches the requirements specified by the Dialog Manager module.

Finally, the output components are responsible for the audiovisual laughter synthesis that generates avatar laugher when the decision components instruct them to do so. For this purpose two different virtual characters are used: Greta

Realizer [3] and Living Actor by Cantoche[1]. At the moment the acoustic and visual modalities of laughter are synthesized separately using the original audiovisual signals from the AVLaughterCycle (AVLC) corpus of human laughter [4]. All synthesized episodes are stored in the agent lexicon, and can then be displayed in real-time. In more details, audio is synthesized with the use of the HMM-based Speech Synthesis System (HTS). HMMs have been trained on the AVLC database and its phonetic annotations [5]. The facial animation in the Greta Realizer was created with two different approaches [6]. First, a procedural approach was used: the AVLC videos were manually annotated with FACS [7], then the animations were resynthesized with the Greta system, able to control the intensity and duration of each action unit. The second approach - a data-driven synthesis - was realized by applying a freely available face tracker to detect facial landmarks on the AVLC videos and then by mapping these landmarks displacements to the facial animation parameters of the virtual character.

The facial animation of Living Actor virtual characters is similar to speech synthesis, where information about phonemes or visemes is sent by the Text to Speech engine along with the audio signal. For laughter, the visemes are composed of lip deformation but also cheek and eye movements. Pseudo-phoneme information is sent using a chosen nomenclature of sounds depending on the synthesis functions. Figure 2 displays examples of laughter poses.


```
<?xml version="1.0" encoding="iso-8859-1" ?>
<LivingActor>
<item time_offset="0.223" morph_name="Laugh_Hum" intensity="1"/>
<item time_offset="0.521" morph_name="Laugh_a" intensity="0.5"/>
<item time_offset="2.190" morph_name="Laugh_a" intensity="0.3"/>
<item time_offset="3" morph_name="Laugh_a" intensity="0.1"/>
</LivingActor>
```

Fig. 2. Using laughter visemes for facial animation

A demo of the system can be viewed at `https://www.youtube.com/watch?v=fElP2_c8vJU`. Further details on the components (acoustic laughter detection, decision and audiovisual synthesis), the communication middleware as well as experimental results can be found in [2].

The rest of this paper focuses on the new input components of our system, with the objective of improving laughter detection robustness through multimodal decisions. In Section 2 we present related work for laughter detection. Section 3

[1] `http://www.cantoche.com`

discusses laughter detection from body cues while Section 4 shows how we can use respiration which is a very important element of laughter expressive pattern. Finally, Section 5 presents the conclusions and future works.

2 Related Work

In the last decade, several systems have been built to detect laughter. It started with audio-only classification.

Kennedy and Ellis [8] obtained 87% accuracy with Support Vector Machines fed with 6 MFCCs; Truong and van Leeuwen [9] reached slightly better results (equal error rate of 11%) with Neural Networks fed with Perceptual Linear Prediction features; Knox and Mirghafori [10] obtained better performance (around 5% error) by using temporal feature windows.

In 2008, Petridis and Pantic started to enrich the so far mainly audio-based work in laughter detection by consulting audio-visual cues for decision level fusion approaches [11,12]. They combined spectral and prosodic features from the audio modality with head movement and facial expressions from the video channel. They reported a classification accuracy of 74.7% to distinguish three classes, namely unvoiced laughter, voiced laughter and speech.

Since laughter detection robustness increases when combining audio and facial features [12], including other modalities can probably further improve the performance. First, the production of audible laughter is, in essence, a respiratory act since it requires the exhalation of air to produce distinctive laughter sounds ("Ha") or less obvious sigh- or hiss-like verbalizations. The respiratory patterns of laughter have been extensively researched as Ruch & Ekman [13] summarize. A distinctive respiration pattern has emerged of a rapid exhalation followed by a period of smaller exhalations at close-to-minimum lung volume. This pattern is reflected by changes in the volume of the thoracic and abdominal cavities, which rapidly decrease to reach a minimum value within approximately 1s [14]. These volumetric changes can be seen through the simpler measure of thoracic circumference, noted almost a century ago by Feleky [15]. Automatic detection of laughter from respiratory actions has previously been investigated using electromyography (EMG). Fukushima et al. [16] analyzed the frequency characteristics of diaphragmatic muscle activity to distinguish laughter, which contained a large high-frequency component, from rest periods, which contained mostly low-frequency components. In this paper, we will explore automatic laughter detection from the measure of the thoracic circumference (Section 4).

Second, intense laughter can be accompanied by changes in postures and body movements, as summarized by Ruch [17] and Ruch & Ekman [13]. Throwing the head backwards will ease powerful exhalations. The forced expiration movements can cause visible vibrations of the trunk and shoulders. This is why we propose features characterizing such laughter-related body movements, that are presented in Section 3.

3 Body Analysis

The EyesWeb XMI platform is a modular system that allows both expert (*e.g.*, researchers in computer engineering) and non-expert users (*e.g.*, artists) to create multimodal installations in a visual way [18]. The platform provides modules, that can be assembled intuitively (*i.e.*, by operating only with the mouse) to create programs, called patches, that exploit system resources such as multimodal files, webcams, sound cards or multiple displays. The body analysis input component consists of an EyesWeb XMI patch performing analysis of the user's body movements in real-time. The computation performed by the patch can be split into a sequence of distinct steps, described in the following paragraphs.

Currently, the task of the body analysis module is to track the user's shoulders and characterize the variation of their positions in real-time. To this aim we could use a sensor like Kinect to provide the user's shoulders data as input to our component. However, we observed that the Kinect shoulders' position do not consistently follow the user's actual shoulder movement: in the Kinect skeleton, shoulders' position is extracted via a statistical algorithm on the user's silhouette and depth map and usually this computation cannot track subtle shoulder movement, for example, small upward/downward movements.

This is why in this paper we present a different type of shoulder movement detection technique: two small and lightweight green polystyrene spheres have been fixed on top of the user's shoulders. The EyesWeb patch separates the green channel of the input video signal to isolate the position of the two spheres. Then a tracking algorithm is performed to follow the motion of the sphere frame by frame, as shown in Figure 3. However, the above technique can be used only in controlled environments, i.e., it can not be used in real situations when users are free to move in the environment. So we plan to perform experiments to compare the two shoulder movement detection techniques: the one based on Kinect and the one based on markers. Results will guide us in developing algorithms for approximating user's shoulder movement from Kinect data.

Fig. 3. Two green spheres placed on the user's shoulders are tracked in real-time (red and blue trajectories)

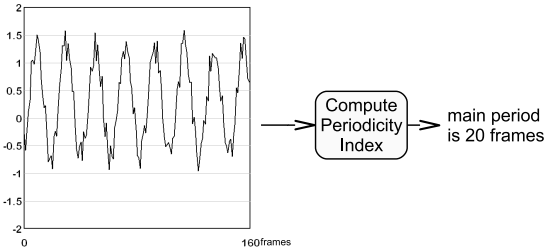


Fig. 4. An example of Periodicity Index computation: the input time-series (on the left) has a periodicity of 20 frames

The position of each user's shoulder is associated to the barycenter of each sphere, which can be computed in two ways. The first consists in the computation of the graphical barycenter of each sphere, that is, the mean of the pixels of each sphere's silhouette is computed. The second option includes some additional steps: after computing the barycenter like in the first case, we consider a square region around it and we apply a Lukas-Kanade [19] algorithm to this area. The result is a set of 3 points on which we compute the mean: the resulting point is taken as the position of the shoulder. From this shoulder tracking, several laughter-related features can be computed:

- **Correlation:** The correlation ρ is computed as the Pearson correlation coefficient between the vertical positions of the user's shoulders. Vertical positions are approximated by the y coordinate of each shoulder's barycenter.
- **Kinetic energy:** The kinetic energy is computed from the speed of user's shoulders and their percentage mass as referred by [20]:

$$E = \frac{1}{2}(m_1 v_1 + m_2 v_2) \tag{1}$$

- **Periodicity:** Kinetic energy is serialized in a sliding window time-series having a fixed length. Periodicity is then computed, using *Periodicity Transforms* [21]. The time-series is decomposed into a sum of its periodic components by projecting data onto periodic subspaces. Periodicity Transforms also output the relative contribution of each periodic signal to the original one. Among many algorithms for computing Periodicity Transforms, we chose *mbest*. It determines the m periodic components that, subtracted from the original signal, minimize the residual energy. With respect to the other algorithms, it provides a better accuracy and does not need the definition of a threshold. Figure 4 shows an example of computation of the Periodicity Index in EyesWeb for a sinusoidal signal affected by a uniform noise in the range $[0, 0.6]$.
- **Body Laughter Index:** Body Laughter Index (BLI) stems from the combination of the averages of shoulders' correlation and kinetic energy, integrated with the Periodicity Index. Such averages are computed over a fixed range of

frames. However such a range could be automatically determined by applying a motion segmentation algorithm on the video source. A weighted sum of the mean correlation of shoulders' movement and of the mean kinetic energy is carried out as follows:

$$BLI = \alpha\bar{\rho} + \beta\bar{E} \tag{2}$$

As reported in [13], rhythmical patterns produced during laughter usually have frequencies around 5 Hz. In order to take into account such rhythmical patterns, the Periodicity Index is used. In particular, the computed BLI value is acknowledged only if the mean Periodicity Index belongs to the arbitrary range $[\frac{fps}{8}, \frac{fps}{2}]$, where fps is the input video frame rate (number of frames per second), 25 in our case.

Figure 5 displays an example of analysis of user's laugh. A previously segmented video is provided as input to the EyesWeb XMI body analysis module. The green plot represents the variation of the BLI in time. When the BLI is acknowledged by the Periodicity Index value the plot becomes red. In [22] we present a preliminary study in which BLI is validated on a corpus of laughter videos. A demonstration of the Body Laughter Index can be watched on `http://www.ilhaire.eu/demo`.

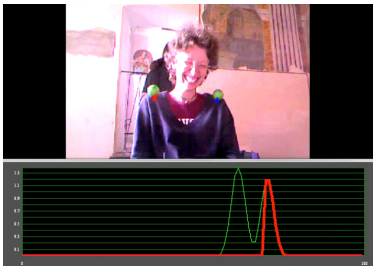


Fig. 5. An example of Body Laughter Index computation

4 Respiration

In order to capture the laughter-related changes in thoracic circumference (see Section 2), we constructed a respiration sensor based on the design of commercially available sensors: the active component is a length of extensible conductive fabric within an otherwise inextensible band that is fitted around the upper thorax. Expansions and contraction of the thorax change the length of the conductive fabric causing changes in its resistance. These changes in resistance are used to modulate an output voltage that is monitored by the Arduino prototyping platform[2]. A custom written code on the Arduino converts the voltage to a 1-byte serial signal, linear with respect to actual circumference, which is passed to a PC over a USB connection at a rate of approximately 120Hz.

[2] `http://www.arduino.cc/`

While Fukushima et al. [16] designed a frequency-based laughter detection module (from EMG signals), our approach is time-based. Laughter onset is identified through the appearance of 3 respiration events (see Figure 6):

1. A sharp change in current respiration state (inhalation, pause, standard exhalation) to rapid exhalation.
2. A period of rapid exhalation resulting in rapid decrease in lung volume.
3. A period of very low lung volume.

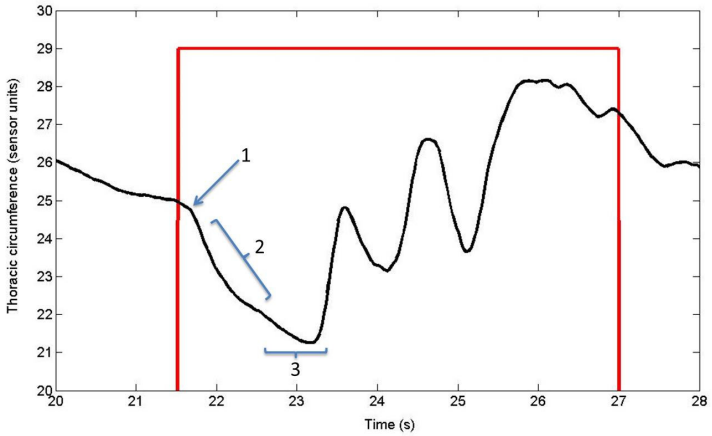


Fig. 6. Example of thoracic circumference, with laughter episode marked in red, and notable features of laughter initiation. Feature 1 - a sharp change in current respiration state to rapid exhalation; feature 2 - a period of rapid exhalation; feature 3 - a period of very low lung volume.

These appear as distinctive events in the thoracic circumference measure and its derivatives:

1. A negative spike in the second derivative of thoracic circumference.
2. A negative period in the first derivative of thoracic circumference.
3. A period of very low thoracic circumference.

These were identified by calculating a running mean (λ_f) and standard deviation (σ_f) for each measure. A running threshold (T_f) for each measure was calculated as: $T_f = \lambda_f - \alpha_f \sigma_f$, where α_f is a coefficient for that measure, empirically determined to optimise the sensitivity/specificity trade-off. Each feature was determined to be present if the value of the measure fell below the threshold at that sample. Laughter onset was identified by the presence of all three features in the relevant order (1 before 2 before 3) in a 1s sliding window. This approach restricts the number of parameters to 3 ($\alpha_{1—3}$) but does introduce lag necessary for calculating valid derivatives from potentially noisy data. It also requires a period for the running means and standard deviations, and so the thresholds, to stabilise. However, this process would be jeopardised by the

presence of large, rapid respiratory event such as coughs and sneezes. The robustness of this detection module remains to be investigated, as well as what it can bring in multimodal detection.

5 Conclusion and Future Work

In this paper we have focused on the development of two new modalities to detect and characterize laughs that are integrated in a broader, fully functional, interactive application. These two modalities are affordable to include in multimodal systems and offer real-time monitoring. The proposed features are related to laughter behavior and will provide useful information to classify laughs and measure their intensity.

This is ongoing work. We will go on developing robust laughter detection. For example, the rules for laughter detection from respiration features, currently determined empirically, will be optimized in a larger study. In addition, other modalities will be included, for example facial tracking. For this purpose we plan to include another sensor, *i.e.* a Kinect camera. The latest version of the Microsoft Kinect SDK not only offers full 3D body tracking, but also a real-time 3D mesh of facial features tracking the head position, location of eyebrows, shape of the mouth, etc. Action units of laughter could thus be detected in real-time.

Secondly, our analysis components need formal evaluation. For this purpose we have recently captured using our analysis components the data of more than 20 people participating in laughter-eliciting interactions. The collected data will now be used to validate these components. In the future, we will also perform a methodical study of multimodal laughter detection and classification (*i.e.*, distinguishing different types of laughter), to evaluate the performance of each modality (audio, face, body, respiration) and measure the improvements that can be achieved by fusing modalities. The long term aim is to develop an intelligent adaptive fusion algorithm. For example, in a noisy environment audio detection should receive a lower importance.

This additional information will allow our decision components to better tune the virtual character reactions to the input, and hence enhance the interactions between the participant and the virtual agent.

Acknowledgment. The research leading to these results has received funding from the European Union Seventh Framework Programme (FP7/2007-2013) under grant agreement $n°270780$. H. Çakmak receives a Ph.D. grant from the Fonds de la Recherche pour l'Industrie et l'Agriculture (F.R.I.A.), Belgium.

References

1. Vettin, J., Todt, D.: Laughter in conversation: Features of occurrence and acoustic structure. Journal of Nonverbal Behavior 28(2), 93–115 (2004)
2. Niewiadomski, R., Hofmann, J., Urbain, J., Platt, T., Wagner, J., Piot, B., Cakmak, H., Pammi, S., Baur, T., Dupont, S., Geist, M., Lingenfelser, F., McKeown, G., Pietquin, O., Ruch, W.: Laugh-aware virtual agent and its impact on user amusement, Saint Paul, Minnesota, USA (May 2013)

3. Niewiadomski, R., Bevacqua, E., Le, Q.A., Obaid, M., Looser, J., Pelachaud, C.: Cross-media agent platform. In: Web3D ACM Conference, Paris, France, pp. 11–19 (2011)
4. Urbain, J., Niewiadomski, R., Bevacqua, E., Dutoit, T., Moinet, A., Pelachaud, C., Picart, B., Tilmanne, J., Wagner, J.: AVLaughterCycle: Enabling a virtual agent to join in laughing with a conversational partner using a similarity-driven audiovisual laughter animation. JMUI 4(1), 47–58 (2010)
5. Urbain, J., Dutoit, T.: A phonetic analysis of natural laughter, for use in automatic laughter processing systems. In: D'Mello, S., Graesser, A., Schuller, B., Martin, J.-C. (eds.) ACII 2011, Part I. LNCS, vol. 6974, pp. 397–406. Springer, Heidelberg (2011)
6. Niewiadomski, R., Pammi, S., Sharma, A., Hofmann, J., Platt, T., Cruz, R., Qu, B.: Visual laughter synthesis: Initial approaches. In: Interdisciplinary Workshop on Laughter and other Non-Verbal Vocalisations in Speech, Dublin, Ireland (2012)
7. Ekman, P., Friesen, W., Hager, J.: Facial action coding system: A technique for the measurement of facial movement (2002)
8. Kennedy, L., Ellis, D.: Laughter detection in meetings. In: NIST ICASSP 2004 Meeting Recognition Workshop, Montreal, pp. 118–121 (May 2004)
9. Truong, K.P., van Leeuwen, D.A.: Automatic discrimination between laughter and speech. Speech Communication 49, 144–158 (2007)
10. Knox, M.T., Mirghafori, N.: Automatic laughter detection using neural networks. In: Proceedings of Interspeech 2007, Antwerp, Belgium, pp. 2973–2976 (August 2007)
11. Petridis, S., Pantic, M.: Fusion of audio and visual cues for laughter detection. In: Proceedings of the 2008 International Conference on Content-Based Image and Video Retrieval, pp. 329–338. ACM (2008)
12. Petridis, S., Pantic, M.: Audiovisual discrimination between speech and laughter: Why and when visual information might help. IEEE Transactions on Multimedia 13(2), 216–234 (2011)
13. Ruch, W., Ekman, P.: The expressive pattern of laughter. In: Kaszniak, A. (ed.) Emotion, Qualia and Consciousness, pp. 426–443. World Scientific Publishers, Tokyo (2001)
14. Filippelli, M., Pellegrino, R., Iandelli, I., Misuri, G., Rodarte, J., Duranti, R., Brusasco, V., Scano, G.: Respiratory dynamics during laughter. Journal of Applied Physiology 90(4), 1441 (2001)
15. Feleky, A.: The influence of the emotions on respiration. Journal of Experimental Psychology 1(3), 218–241 (1916)
16. Fukushima, S., Hashimoto, Y., Nozawa, T., Kajimoto, H.: Laugh enhancer using laugh track synchronized with the user's laugh motion. In: Proceedings of CHI 2010, pp. 3613–3618 (2010)
17. Ruch, W.: Exhilaration and humor. Handbook of Emotions 1, 605–616 (1993)
18. Camurri, A., Coletta, P., Varni, G., Ghisio, S.: Developing multimodal interactive systems with eyesweb xmi. In: NIME 2007, pp. 302–305 (2007)
19. Lukas, B., Kanade, T.: An iterative image registration technique with an application to stereo vision. In: Proceedings of the 7th IJCAI (1981)
20. Winter, D.: Biomechanics and motor control of human movement (1990)
21. Sethares, W., Staley, T.: Periodicity transforms. IEEE Transactions on Signal Processing 47(11), 2953–2964 (1999)
22. Mancini, M., Varni, G., Glowinski, D., Volpe, G.: Computing and evaluating the body laughter index. In: Salah, A.A., Ruiz-del-Solar, J., Meriçli, Ç., Oudeyer, P.-Y. (eds.) HBU 2012. LNCS, vol. 7559, pp. 90–98. Springer, Heidelberg (2012)

@scapa: A New Media Art Installation in the Context of Physical Computing and AHRI Design

Andreas Gernemann-Paulsen[1], Claudia Robles Angel[2], Lüder Schmidt[1], and Uwe Seifert[1]

[1] Institute of Musicology , University of Cologne, Germany
{andreas.gernemann,lueder.schmidt,u.seifert}@uni-koeln.de
[2] New Media & Audiovisual Artist
post@claudearobles.de
www.claudearobles.de

Abstract. In this paper **@scapa**, an installation developed in the context of Artistic Human-Robot Interaction design (AHRI design) is introduced. AHRI design is a methodological approach to realize cognitive science's research paradigm of situated or embodied cognition within cognitive musicology to investigate social interaction in artistic contexts [11], [12], [13] using structured observation [2], [6]. Here we focus on design aspects the course of development of **@scapa** using procedures of Physical Computing and the aspects to develop such a New Media Installation in the framework of AHRI design [5].

Keywords: situated cognition, cognitive science, research methodology, New Media Art, Physical Computing, artistic human-robot interaction design, structured observation.

1 Description

@scapa, an installation for robot, light and sound, is developed by Andreas Gernemann-Paulsen, Claudia Robles Angel and Lüder Schmidt.

The artistic idea is apparent behind the sonic conception as well as in the iterative, continuously approximating realization process of **@scapa**, which conforms to the notion of tinkering as known in the context of Physical Computing. In particular the artistic impulses for lighting and the robot's behavior resulted from this approach. The system behavior is controlled by the values of the IR distance sensors of Nibo 2 robot, which can move on a small 'stage' which consists of a circular table made by Tekno. The robot can exhibit four movement patterns labeled *rest*, *tremor*, *escape* and *panic*, depending on certain threshold values. The visitors can activate these different states by approaching the robot with the hands. This also changes in a continuous manner the sound as well as the color of the lights surrounding the table.

M. Mancas et al. (Eds.): INTETAIN 2013, LNICST 124, pp. 193–198, 2013.

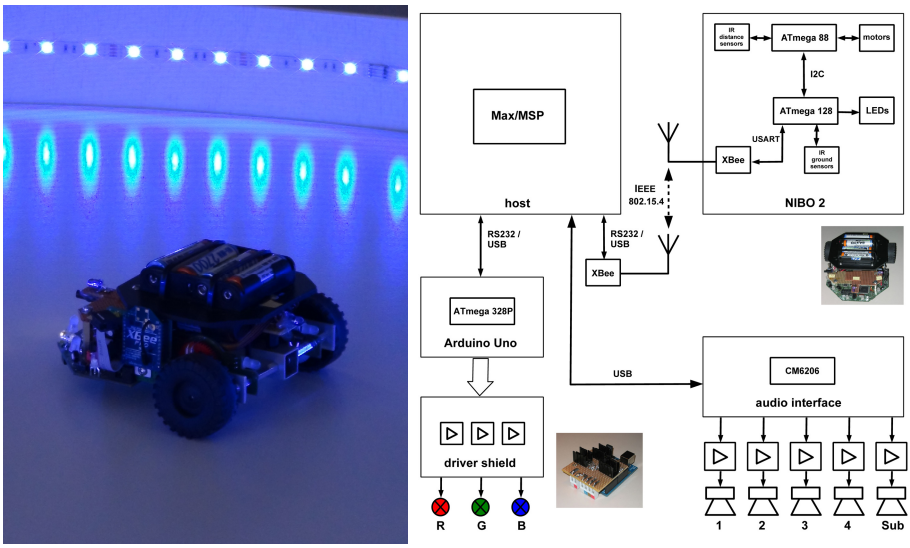


Fig. 1. The modified Nibo2 robot and the block diagram of **@scapa**

1.1 Sonic Aspects

With particular regard to the sonic aspect of this installation, the sound increases the body-awareness of listeners/visitors by using two adjustable sound files containing, on the one hand a heartbeat and, on the other hand, human breathing. The selection of these sounds is due to the fact that in everyday life, people tend not to listen to internal body sounds. Thus, the sound conception of the work allows for listeners/visitors to perceive these generally ignored internal sounds in an extremely audible and perceptible manner in situations such as changes in psychological and emotional affections due to external stimuli. In this way, the robot and its internal sounds, becomes a mirror of the human being.

The aforementioned sounds are assigned to the robot, which reacts according to an external human presence. Those sounds, which seem to be originated by the robot, are exteriorised in a quadrophonic and immersive sound environment inviting the visitors/participants to reflect about unconscious and rather imperceptible internal human reactions.

The artistic intention of **@scapa** is based on the idea of making visible and/or audible the invisible and/or inaudible, the aesthetical position of artist Claudia Robles Angel, one of the authors herewith, who, in her work, seeks to transform what is imperceptible into a perceptible sensation.

1.2 Technical Aspects: Nibo-2 Robot, MAX and Arduino Board

The artist's intention is supported by a built-in table lighting, which is realized by a custom-made border with about 150 RGB-LEDs integrated. Corresponding to the four

states of the robot, the lighting changes constantly and with a slight delay between blue (*rest*), purple (*tremors*) and red (*escape* and *panic*). The light is controlled from a Max patch via the serial interface and an Arduino Uno board with a suitable sketch (Arduino program) and a custom-made driver shield. The sound is realized in the Max patch too. The sensor data of the Nibo 2 are sent via a modified Xbee connection (Nicai NXB2) to a laptop and are further processed in Max. For a better performance the hardware regarding to the IR sensors have been modified: the sensors are placed on a separate board, one of them is now located on the rear side.

All mechanical and electronic components of the installation can be disassembled so that they can be transported in an estate car.

2 Physical Computing

2.1 Physical Computing in New Media Art

A core aspect of Physical Computing is the use of specific hardware and software [1], [4] [5]. In the context of New Media Art the practical aspects emphasize the easy handling and a tinkering approach. Particularly the Arduino project - originating from the artistic context and often used in interactive New Media Art - exemplifies this. Especially the artistic creative use of current low-cost chip technology, the relationship between software and hardware and the human interaction could be mentioned [7]. Precisely the projects from Physical Computing can be an approach to enable humans to express themselves physically and to provide artists with tools to capture and convey their ideas [8]. The realization of hardware and software in a simple, tinkering way and the low-cost design as a continuously approximating approach where sometimes a goal is not clearly defined emphasizes this point of view [1], [4]. Thus, by the use of microcontroller technology not only new artistic ideas can be realized (which could not without this technology) but this practical and iterative approach also results in new creative impulses.

2.2 Implication of Physical Computing in @scapa

The importance of Physical Computing's tinkering approach during the design of @scapa can be observed in the realization and development of the aesthetic and artistic ideas concerning lighting, sound, and robot behavior. The continuously approximating realization of the lighting, the robot's behavior and the modification of the hardware results in an iterative way of using typical parts like the Arduino board and procedures from Physical Computing. Concerning the light effects this process started from the initial idea to use a simple lamp and resulted in the extensive design and preparation of the border lighting of the installation, whose states correspond to the behavior of the robot. Moreover, concerning the robot's final behavior and environment one could observe a mutual dependency mediated by tinkering during the development of the installation of technical realization and artistic ideas. The robot's environment – a round table with light effects – and it's simple and

transparent behavior can be viewed as the artistic attempts to bypass the limited possibilities of movement by two wheels and also the robot's small size in order to facilitate the visitors' interactions with the robot by an attractive appearance creating a specific atmosphere to enhance aesthetic expectations. Additionally, the white cotton gloves invite the visitors to wear and to approach the robot with their prepared hands. The final realization of the aesthetic core idea to let the visitors experience and reflect on vital sounds - inspired by the two sound files - resulted from constant tinkering in connection with aesthetic explorations.

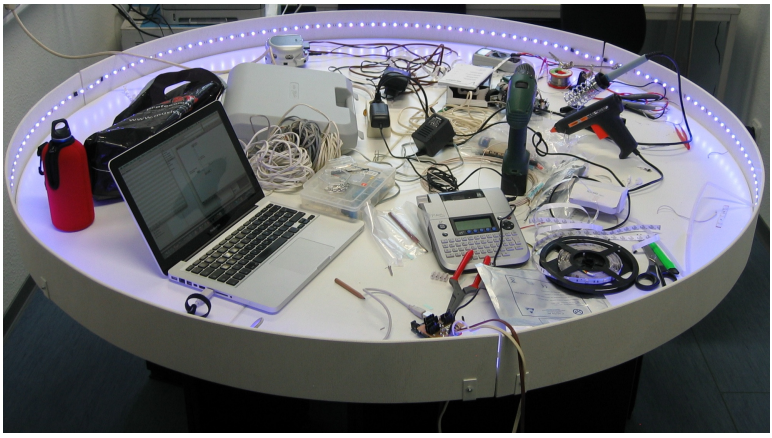


Fig. 2. Many electronic parts and tools referring to the tinkering process during the design of @scapa

3 AHRI Design

3.1 AHRI Design: Situated Cognition

Artistic Human-Robot Interaction design (AHRI design) is an attempt to develop an empirical research methodology which strives for combining computational cognitive modeling and traditional empirical research strategies to study the human mind from the perspective of situated or embodied cognition [11], [12], [13], [6], [3], [5]. One basic assumption of situated cognition is that in studying mental phenomena interactions of a cognitive system with its environment need to be taken into account [9], [10]. In particular, for human cognitive capacities such as language, music, and art social interaction and communication with conspecifics come into play. Therefore, AHRI design uses New Media Art installations "to test theories and collect empirical data in semi-artificial social environments" ([11], p. 67) to study social interaction. Such an approach has to deal with conceptual as well as empirical and practical problems from science, art, and technology at different levels. For example, computational models of interaction and cognition need to be developed and implemented in robotic systems and, then, integrated in an art installation. During

such installations empirical data relevant to the cognitive capacity under study and its related social interactions are gathered from interactions of humans with robotic systems by means of structured observation. Structured observation as a well-established method for data acquisition in psychology, sociology and ethology was chosen as a research method because methodologically it allows to investigate and to develop functional units of analysis for interaction by developing observational categories for further research. As an observational method it offers a more open and flexible way to investigate complex situations such as social interaction than measurements based on interval and ratio scales and yet allows using statistical methods for data analysis. These analyses are used to test or develop ideas concerning the investigated cognitive capacity's underlying processes and mechanisms.

3.2 AHRI Design and @scapa

This installation was developed in the context of the ideas for the research method artistic human-robot interaction design (AHRI design) at the Institute of Systematic Musicology at the University of Cologne. In more detail it is suitable to study social interaction in artistic contexts within the framework of cognitive musicology [11]. Particularly **@scapa** has a distinct artistic approach. This is apparent in the sonic conception mentioned above as well in an iterative „tinkering" way as a continuously approximating realization described in chapter 1.2 and 2.2.

This project provides new aspects in conjunction with the methods of cognitive science. Investigations and reflections on interactive processes should be mentioned in the near future, so the installation would be appropriated in terms of observational studies on interactive processes between humans and machines. In particular, the nature and the process of social interaction should be investigated in the framework of interactive installations of New Media Art.

References

1. Banzi, M., Cuartielles, D., Igoe, T., Mellis, D., Martino, G.: Getting Started with Arduino. Attribution-NonCommercial-ShareAlike 2.5, Milano (2011), http://wiki.digitalarts.wits.ac.za/uploads/2/24/Getting_Started_With_Arduino.pdf
2. Bartneck, C., Kulic, D., Croft, E.: Measuring the Anthropomorphism, Animacy, Likeability, Perceived Intelligence, and Perceived Safety of Robots. In: Proceedings of the Metrics for Human-Robot Interaction Workshop in Affiliation with the 3rd ACM/IEEE International Conference on Human-Robot Interaction (HRI 2008), Technical Report 471, Amsterdam, March 12-15 (2008)
3. Buch, B., Coussement, P., Schmidt, L.: 'playing_ robot': An Interactive Sound Installation in Human Robot- Interaction Design for New Media Art. In: Proceedings of the Conference on New Interfaces for Musical Expression (NIME 2010), Sydney, pp. 411–414 (2010)

4. Gernemann-Paulsen, A., Robles Angel, C., Seifert, U., Schmidt, L.: Physical Computing and New Media Art - new challenges in events. In: 27. Tonmeistertagung, 2012, Proceedings, Cologne 2012, pp. 822–832 (2012), http://www.vdtshop.de/2012ST3 (May 2013)
5. Gernemann-Paulsen, A., Schmidt, L., Seifert, U., Buch, B., Otto, J.A.: Artistic Human-Robot Interaction Design and Physical Computing: New Directions in Music Research and Their Implications for Formal Scientific Education. In: 26. Tonmeistertagung, 2010, Proceedings, Leipzig, pp. 574–586 (2012), http://www.tonmeister.de/tmt/2010/tmt_dl.php?tmtid=2010&lang=de&pid=99&v=pr (available May 2013)
6. Kim, J.H., Chang, S.-H., Schmidt, L., Otto, J.-A., Buch, B., Seifert, U., Coussement, P.: 'playing_robot': An investigation of situated cognition in the context of (artistic) human-robot interaction design. In: Proceedings of the 1st IROS 2010 Workshop on Robots and Musical Expressions (IWRME 2001), Taipei, pp. 65–72 (2010)
7. Odendahl, M., Finn, J., Wenger, A.: Arduino - Physical Computing für Bastler, Designer und Geeks. O'Reilly, Cologne (2010)
8. O'Sullivan, D., Igoe, T.: Physical Computing: Sensing and Controlling the Physical World with Computers. Course Technology Cengage Learning, Boston (2004)
9. Robbins, P., Aydede, M.: A short primer on situated cognition. In: Robbins, P., Aydede, M. (eds.) The Cambridge Handbook of Situated Cognition, pp. 3–10. Cambridge University Press, Cambridge (2009)
10. Shapiro, L.: The embodied cognition research program. Philosophy Compass 2(2), 338–346 (2007)
11. Seifert, U.: Investigating the musical mind: Cognitive musicology, situated cognition, and artistic human- robot interaction design, and cognitive musicology. In: EWHA HK International Conference: Principles of Media Convergence in the Digital Age, LG Convention Hall, Ewha Womans University, June 24-25, pp. 61–74. Ewha Women's University Press, Seoul (2010)
12. Seifert, U., Kim, J.H.: Entelechy and embodiment in (artistic) human-computer interaction. In: Jacko, J.A. (ed.) HCI 2007. LNCS, vol. 4550, pp. 929–938. Springer, Heidelberg (2007)
13. Seifert, U., Kim, J.H.: Towards a conceptual framework and an empirical methodology in research on artistic human-computer and human-robot interaction. In: Pavlidis, I. (ed.) Advances in Human- Computer Interaction, pp. 177–194. In-Tech Education and Publishing, Vienna (2008)

Author Index