Search and retrieval of audiovisual content by integrating non-verbal multimodal, affective, and social descriptors

Antonio Camurri

Casa Paganini – InfoMus Intl Research Centre
DIST- University of Genova
Piazza Santa Maria in Passione 34, 16123 Genova, Italy

antonio.camurri@unige.it

Abstract. One of the research challenges for future search engines concerns the integration of multimodal and cross-modal, nonverbal, full-body, affective, social, and enactive interaction in the process of search and retrieval of audiovisual content. The paper gives a short presentation of the three-year EU project I-SEARCH (EU 7FP ICT STREP), aiming at creating a novel unified framework for multimodal and cross-modal content indexing, sharing, search and retrieval of audiovisual content. A couple of scenarios developing multimodal paradigms of search and retirieval of audiovisual content are introduced and briefly discussed to explain in concrete terms some of the main research challenges that are addressed in I-SEARCH. Finally, the paper presents preliminary results on a specific research challenge: analysis of non-verbal expressive and social behaviour to extract useful information from users for the retrieval of audiovisual content.

Keywords: non-verbal full-body multimodal interfaces; cross-modal descriptors; emotion; social signals; sound and music computing.

1 Introduction

Internet is quickly evolving towards providing richer and immersive experiences, in which the user interact seamlessly and transparently with digital and physical artefacts. Due to the widespread availability of digital recording devices, improved modelling tools, advanced scanning mechanisms as well as display and rendering devices, even on mobile environments, users are more and more empowered to have a more immersive and interactive experience. Digital media are moving to "User Centric Media" [2,3], enabling adaptive and active experiences of audiovisual content (see for example the EU ICT SAME Project www.sameproject.org). Users become "prosumers", and are more and more participating in the updating process of the information and in improving the resolution and richness of the media repositories. The emergence of embodied and social interaction with content, enabled by the dramatic advances of multimodal/intelligent/natural interfaces, enriches this scenario, providing users with further degrees of freedom and channels to access the content, in terms of full-body, non-verbal, expressive [7], social [5] interaction with content.

It is therefore now possible for users to rapidly move from a mainly textual-based to a media-based "embodied" Internet, where rich audiovisual content (images, graphics, sound, videos, 3D models, etc.), 3D representations (avatars), virtual and mirror worlds, serious games, lifelogging applications, multimodal yet affective utterances (gestures, facial expressions, eye movements,…) etc. become a reality. See for example [1] for an extensive survey on content-based multimedia information retrieval, and the "white papers" from EU on Future internet and on User-Centric Media [2,3] for an in depth analysis of future internet and emerging user-centric media.

Traditional search of music archives usually include descriptive methods (mainly textual, e.g., author, title, etc.). Recent search engines also include alternative querying modalities such as audio (Shazam, Google China Music) or image (Google goggles, Google similar images). However, these search engines are suited for query-by-content or query-by-example, where the research objective is clearly defined or where a specific information is targeted.

We aim at developing alternative, yet complementary, querying modalities, e.g., integrating expressive gesture and affective, emotional cues, that facilitate more explorative and creative search.

This paper presents some insights and preliminary research results on the problem of search and retrieval in cases where textual information is either missing or it is not sufficient or adequate, and therefore the need for the integration of non-verbal multimodal, cross-modal, full-body, affective, social descriptors emerges.

Use case scenarios on music content search and retrieval are adopted in this paper to discuss main research challenges and approaches.

2 The I-SEARCH EU Project

The EU 7FP ICT STREP I-SEARCH project aims to create a novel unified framework for multimedia and multimodal content indexing, sharing, search and retrieval. I-SEARCH is coordinated by Dimitrios Tsovaras by ITI-CERTH (Centre for Research and Technology Hellas – Informatics and Telematics Institute), and partners include JCP-Consult, INRIA, Athens Technology Center, Engineering Ingegneria Informatica S.p.A., Google, University of Genoa, Exalead, Erfurt University of Applied Sciences, Accademia Nazionale di Santa Cecilia, EasternGraphics. The project started in January 2010, with a duration of three years.

"I-SEARCH aims to create a novel unified framework for multimedia and multimodal content indexing, sharing, search and retrieval. I-SEARCH aims to be the first search engine able to handle specific types of multimedia (text, 2D image, sketch, video, 3D objects, audio and combination of the above) and multimodal content (gestures, face expressions, eye movements) along with real world information (GPS, temperature, time, weather sensors, RFID objects,), which can be used as queries and retrieve any available relevant content of any of the aforementioned types and from any end-user access device. Towards this aim, I-SEARCH proposes the research and development of an innovative Rich Unified Content Description (RUCoD). RUCoD will consist of a multi-layered structure, which will integrate descriptors of all of the above types of

content, real-world information, even non-verbal yet implicit, emotional cues and social descriptors, in order to better express what the user wants to retrieve. Another objective of I-SEARCH is the development of intelligent content interaction mechanisms, including personal, social-based and recommendation-based relevance feedback and novel interoperable multimodal interfaces. This will result in a highly user-centric search engine, able to deliver to the end-users only the content of interest, satisfying their information needs and preferences, which is expected to dramatically improve end-user experience and offer new market opportunities. Furthermore, I-SEARCH introduces the use of advanced visual analytic technologies for search results presentation in order to facilitate their fast and easy interpretation and also to support optimal results presentation under various contexts (i.e. user profile, end-user terminal, available network bandwidth, interaction modality preference, etc.). Finally, the search engine will be dynamically adapted to end-user's device, which will vary from a simple mobile phone to a high-performance PC." (from cordis.europe.eu projects archive)

3 Scenarios

3.1 Music retrieval through expressive embodied queries

Chiara is a music-lover, looking for music material that share common affective features. Here, search aims at discovering unexpected filiations and similarities across music artworks. It takes place in an environment equipped with devices enabling the user to express herself through voice, hands and body gesture. Pre-recorded multimedia content can be uploaded from an external device (e.g., a mobile phone). For each digital content in the collection, descriptors related to low-level features, real-world context data, and expressive/emotional/social cues that compose the RUCoD (Rich Unified Content Description of the I-SEARCH project) standard are stored.

Inputs to the search module include text queries, audio capture of live user singing incipit of music piece, audio file recorded on handheld device such as mobile phones. Beat tracking can be captured by tapping on a microphone or through accelerometers embedded in mobile devices. User gestures can be captured using either video camera or accelerometers embedded in user's mobile devices.

Chiara wants to explore music artworks that share affective features with the Ravel's Bolero. She starts by using the I-SEARCH framework to retrieve audio information that share similarities with this audio pattern. Using a tangible acoustic interface [6], she taps the beating of the rhythm, a constant 4/4 time with a prominent triplet on the second beat of every bar, or smaller rhythmic cells (for example the beat and triplet).

The recorded audio fragment is used by the I-SEARCH engine as an audio query to initialize the search. Specifically, the I-SEARCH framework extracts low-level descriptors from the audio content (the tapping resulting audio) and create a query based on the RuCoD format. Through template matching techniques [4], similar audio results related to Bolero rhythmic pattern are retrieved. Related video files and music scores are also retrieved through multimodal annotation propagation. Results are

displayed via visual analytics techniques on Chiara's terminal, using clusters annotated with information like modality type, population size and others.
Chiara picks one of the results returned by the query, listens to it but decides that what she needs is something more energetic, so she closes her fists and starts making sharp, sudden vertical movements on the same Bolero rhythm. Through a video camera or embedded accelerometers the environment captures such expressive features of the gesture, and refine the search, resulting in changes in the displayed results to match it (either by removing the items that don't convey that expression, or by moving the suitable items closer to build a different cluster configuration). One of the results captures Chiara's attention: a drum recording from Italian ethnomusicological repertoire, the 'Ritmo di tamburo', where various drum rhythms are played, sharing indeed the same triplet of Bolero. A little further away she also finds a voice recording.

3.2 Collective DJ - Social music retrieval through expressive embodied queries

Four friends at a party wish to dance together, and to accomplish this they search some music pieces resonating with their (collective) mood. They do not know in advance the music pieces they want, and they use the I-SEARCH tool collaboratively to find their music, and possible associated video. Alternatively, the search process might not necessarily be 'conscious' or intentional, but simply a part of a social game, i.e., in a more fun/entertainment approach.
The friends are at a party, but not necessarily they share the same physical environment. One or more of them can be remotely connected via audiovisual links. GPS and context aware information are available to the ISEARCH search engine.
Users have devices that enable to express themselves through voice, movement, face, and full-body movements.
For each digital music content of the archive under consideration, descriptors related to low-level features, real-world data and expressive/emotional/social cues that compose the RuCoD (Rich Unified Content Description) standard are available.
Users inputs include the following: (i) Rhythmic queries, using hands, clapping, full-body movement; (ii) Context data (GPS, compass, proximity of others etc); (iii) Entrainment/synchronisation and dominance/leadership among users, measured by on-body sensors (eg accelerometers or/and videocameras on their mobiles, or game interfaces) and/or environment videocamera(s), to find a shared, collective information to build the query based on non-verbal social signals; (iv) Gestures to shape the query: again, the gesture are captured using either video camera and/or accelerometers embedded in mobile devices (carried by the user or kept in hand) or in environmental videocameras.
The output of the experience include a shared enjoyment of performances of the retrieved music pieces, possible video clips associated to such music pieces (music videos).

The experience may be described as follows:

1. The four users A, B,C, D start to dance;
2. Their movement acts as selector of music pieces coded in relation to their motoric-affective-social behaviour (slow, fast, dionisiac, ...);
3. If the movements of A,B,C,D are not sympathetic (i.e., low entrainment), an overlapping of different music pieces will emerge, in a rather chaotic sound environment (the different music pieces are heard simultaneously at different changing levels according to users behaviour). As the joint experience goes on, the music continuously changes, and may start to converge to a piece corresponding to the user who results to dominate in the group (dominance/leadership features which are automatically extracted by the system);
4. When the movement of the group obtains a sufficient uniformity (contagion from the one who results to act as a leader), the group will converge to dance on the same music, chosen by such "collective gesture": this is the first level of query result;
5. This entertaining task is integrated and is part of the search task: for example, the search can take the priority as soon as one of the users, once obtained a shared agreement and a single music, tries to trigger a change in the shared general emotion of the group (with a consequent change of musical choices), for example a perturbation to the current situation by a user, to explore music pieces similar to the one obtained and experienced by the group. The effect can be a sort of "game on leadership": the user who is able to triggers a general perturbation of emotion in the group causes a sort of "collective DJ-like" real-time interactive editing of heterogeneous music fragments on the main music piece. The user who is able to act as a leader (detected by the system), has the possibility to inject the consequences of her gestural/movement/affective choices, which will take the power to determine the change of the music context only in relation to the capability of her contagion on the other users. If the others will be captured, they will follow her moving toward a new music piece, by means of an audio and possibly video cross-fade or sudden change of scene. Otherwise, if the user will not have enough power or dominance on the others, her associated new music piece will fade off.

This is a sort of a "Collective DJ" example, in which the collective behaviour provides the source data for the search of the music. It is the opposite of the traditional music experience: here the movement determines its own correct music frame [8].
The music retrieval may also keep into account of GPS and spatial locations of users. For example, GPS information may be kept into account to select music pieces keeping into account their geographical region.

3.3 Discussion

Social search can occur simultaneously or in different moments. In the first case, which is considered in the scenario 2, users collaborate to a common shared objective. In the latter, a user leaves a track of her activity, which can be used later by other users to take inspiration for their search. Of course, mechanisms allowing the

intentional control of access to personal search schemas and data, similarly to shared data in social networks, must be introduced.
In situations of "affective non-verbal querying" the following three types of difficulties emerge:

(i) The user is not familiar to non-verbal search by means of "affect", "expressivity", "embodiment": she is not sure if she is able to perform correctly the task, she does not know if she is able to link an affect with a content. For example, if she sings/hums a song but she has poor singing skills, this can be a case in which the user is not able to express the affect she wants to convey (but associated gesture may help). In a "game-like" scenario, the users may be more inclined to dancing even if they are not professional dancers, for the sake of fun and social interaction. In this case, the results of the scenario should be measured more in terms of user experience than on the technical 'efficiency' and performance of the search task.

(ii) The user may be not capable to express what she is searching for;

(iii) The machine might not be able to execute the search according the user intentions.

Despite these difficulties, emotional/expressive social non-verbal queries, with their inner "blurred" characterisation, can enhance "serendipitous discovery", and can lead to stimulating exploring-style querying, complementary to traditional querying paradigms.

4. Multimodal queries based on non-verbal expressive and social features

Several research challenges emerge from the scenarios sketched above, including the understanding of users multimodal inputs, and in particular the non-verbal multimodal cues conveying users' intentions, and the mapping between user multimodal inputs to features in the audiovisual content.

In this section we focus on how to exploit users' non-verbal expressive and social signals useful to build multimodal queries.

The proposed approach consists of two phases:

(i) Extraction of an array of expressive features describing each user behavior [12];

(ii) Using such expressive features as the inputs to modules which extract social features related users behavior, with particular focus on entrainment and dominance [13, 14, 15].

The resulting array of individual and social features will be a subset of the user component RuCoD descriptors.

4.1 Expressive features

As for the extraction of descriptors on non-verbal expressive behaviour, a particular focus is on full-body movement and gesture, i.e., on recognizing how a gesture is performed, including expressive and affective content [12]. Typically, a single gesture can be performed in several different ways (e.g., fluid, hesitant, impulsive). A collection of features characterizing the expressive qualities of a gesture has been defined, starting from biomechanics, psychology, and humanistic theories [12].
Expressive features include the following:

Quantity of Motion (QoM) is an index of motoric activation that provides an estimation of the amount of overall movement (variation of pixels) the video-camera detects. QoM computed on translational movements only (TQoM) provides an estimation of how much the user is moving around the physical space. Using Laban's Effort [17] terminology, whereas Quantity of Motion measures the amount of detected movement in both the Kinesphere and the General Space, its computation on translational movements refers to the overall detected movement in the General Space only. TQoM, together with speed of barycentre (BS) and variation of the Contraction Index (dCI) are introduced to distinguish between the movement of the body in the General Space and the movement of the limbs in the Kinesphere. Intuitively, if the user moves her limbs but does not change her position in the space, TQoM and BS will have low values, while QoM and dCI will have higher values.

Impulsiveness (IM) is extracted using a model measuring it as a combination of other features, mainly derived from QoM. The first one is the variance of QoM in a sliding time window of 3s, i.e., a user is considered to move in an impulsive way if the amount of movement the video-camera detects on her body changes considerably in the time window. A second group of features is related to the analysis of the shape of QoM along time. Such features include e.g., the ratio between the main peak of QoM in the time window and the time duration of such a peak, the steepness of the attack of a movement phase, the steepness of the main peak, the number of peaks detected in the time window, the ratio between the main peak and the second biggest one, the distribution of the peaks in the time window (i.e., whether they are uniformly distributed along the time window or concentrated over a specific time range). Another feature is related to the content of the QoM spectrum in the frequency band over 5 Hz. Each feature is then weighted and combined in the model in order to provide an overall index of Impulsiveness.

Vertical and horizontal components of velocity of peripheral upper parts of the body (VV, HV) are computed starting from the positions of the upper vertexes of the body bounding rectangle. The vertical component, in particular, is used for detecting upward movements that psychologists (e.g., [16]) identified as a significant indicator of positive emotional expression.

Space Occupation Area (SOA) is computed starting from the movement trajectory integrated over time. In such a way a bitmap is obtained, summarizing the trajectory followed along the considered time window (3s). An elliptical approximation of the shape of the trajectory is then computed. The area of such ellipse is taken as the Space Occupation Area. Intuitively, a trajectory spread over the whole space gets high SOA values, whereas a trajectory confined in a small region gets low SOA values.

Directness Index (DI) is computed as the ratio between the length of the straight line connecting the first and last point of a trajectory (in this case the movement trajectory in the selected 3s time window) and the sum of the lengths of each segment composing the trajectory. It is inspired by the Space dimension of Laban's Effort Theory.

Space Allure (SA) measures local deviations from the straight line trajectory. It is inspired by composer Pierre Schaeffer's Morphology. Whereas DI provides information about whether the trajectory followed along the 3s time window is direct or flexible, SA refers to waving movements around the straight trajectory in shorter time windows. Currently, SA is approximated with the variance of DI in a time window of 1s.

The *Amount of Periodic Movement* (PM) provides a preliminary information about the presence of rhythmic movements. Computation of PM starts from QoM. Movement is segmented in motion and pause phases using an adaptive threshold on QoM [12]; inter-onset intervals are then computed as the time elapsing from the beginning of a motion phase and the beginning of the following motion phase. The variance of such inter-onset intervals is taken as an approximate measure of PM.

Symmetry Index (SI) is computed from the position of the barycenter and the left and right edges of the body bounding rectangle. That is, it is the ratio between the difference of the distances of the barycenter from the left and right edges and the width of the bounding rectangle:

$$SI = \frac{\big||x_B - x_L| - |x_B - x_R|\big|}{|x_R - x_L|}$$

where x_B is the x coordinate of the barycentre, x_L is the x coordinate of the left edge of the body bounding rectangle and x_R is the x coordinate of the right edge.

The expressive features, can be analyzed using video input from videocameras, but other sensor inputs can be considered, e.g. the 3D accelerometers embedded in mobile systems.

Current work aims at refining and extending the set of expressive features, to contribute to the RuCoD standard.

The expressive features are implemented as real-time software modules in the open software platform EyexWeb XMI (www.eyesweb.org).

4.2 Social features

Research on the analysis of social descriptors include the development of models and techniques for measuring entrainment, empathy, dominance, leadership, and salient behaviour in small groups of users. We obtained preliminary results on entrainment and dominance, based on theories of synchronization [13,14,15].

A basic assumption to approach nonverbal social behavior consists of the modeling of a small group of users as a complex system consisting of single interacting components able to auto-organize and to show global properties, which are not obvious from the observation of their individual dynamics.

A number of algorithms and related software modules for the automated analysis in real-time of non-verbal cues related to expressive gesture in social interaction are

currently studied at our centre. Analysis of entrainment and dominance is based on Phase Synchronisation and Recurrence Quantification Analysis. We start from the hypothesis that phase synchronisation is one of the low-level social signals explaining empathy and dominance in a small group of users. Another direction, based on Multi Scale Entropy and other approaches is currently adopted to measure saliency and rarity index in small group of users. Real time implementation of the algorithms developed so far is available in the EyesWeb XMI Social Signal Processing Library (www.eyesweb.org).

5 Conclusions

Some of the main research challenges faced in the EU 7FP ICT I-SEARCH project, and preliminary results in the analysis of users behavior in terms of expressive and social features have been presented, as a contribute to the RuCoD standard in I-SEARCH. Current work includes research on empathy, emotional entrainment, leadership, co-creation, and attention.

Other important directions of the research in I-SEARCH concern the study of descriptors in audiovisual content, and the study of cross-modal descriptors [11].

Acknowledgements

I am deeply grateful to my colleagues and friends Corrado Canepa, Paolo Coletta, Nicola Ferrari, Alberto Massari, Gualtiero Volpe, Donald Glowinski, Maurizio Mancini, Giovanna Varni.
This research is partially supported by the 7FP EU-ICT three-year project I-SEARCH no. 248296.

References

1. Lew, Michael S., Sebe, N., Djeraba, C., and Jain, R.. Content-based multimedia information retrieval: State of the art and challenges. ACM Trans. Multimedia Comput. Commun. Appl., 2(1):1–19, (2006).
2. Laso Ballestreros, I. (Ed.) Research on Future Media Internet, Future Media Internet Task Force, European Commission, 7FP ICT Networked Media Unit, January 2009.
3. Laso Ballestreros, I. (Ed.) User Centric Media in the Future Internet, European Commission, 7FP ICT Networked Media Unit, November 2009.
4. Casey, M.A., R. Veltkamp, M. Goto, M. Leman, C. Rhodes, and M. Slaney. Content-based music information retrieval: current directions and future challenges. *PROCEEDINGS-IEEE*, 96(4):668, 2008.
5. Pentland, A.,: Socially aware, Computation and Communication. Computer, IEEE CS Press (2005)

6. Camurri, A., C.Canepa, S.Ghisio, G.Volpe (2009) Automatic Classification of Expressive Hand Gestures on Tangible Acoustic Interfaces According to Laban's Theory of Effort. In M.S.Dias, S.Gibet, M.W.Wanderley, R.Bastos (Eds.), Gesture-Based Human-Computer Interaction and Simulation, pp.151-162, LNAI5085, Springer, ISSN 0302-9743.
7. Camurri, A., De Poli, G., Leman, M., Volpe, G.: Toward Communicating Expressiveness and Affect in Multimodal Interactive Systems for Performing Art and Cultural Applications, IEEE Multimedia, Vol.12, No.1, pp.43-53, IEEE Computer Society Press (2005)
8. Camurri, A.: Interactive Dance/Music Systems, Proc. *Intl. Computer Music Conference ICMC-95*, pp.245-252, The Banff Centre for the arts, Sept.3-7, Canada, ICMA-Intl.Comp.Mus.Association (1995)
9. Camurri, A., Canepa C., Volpe G.: Active listening to a virtual orchestra through an expressive gestural interface: The Orchestra Explorer. Proc. Intl Conf NIME-2007 New Interfaces for Music Expression, New York University, (2007)
10. Varni, G., Camurri, A., Coletta, P., Volpe G.: Emotional Entrainment in Music Performance. Proc. 8th IEEE Intl Conf on Automatic Face and Gesture Recognition, Sept. 17-19, Amsterdam (2008).
11. Camurri, A., P.Coletta, C.Drioli, A.Massari, G.Volpe (2005). Audio processing in a multimodal framework. Proc. Intl. Conf. AES-05 Audio Engineering Society, Barcelona, May 2005.
12. A.Camurri, B. Mazzarino, G. Volpe (2004) Expressive Interfaces. Cognition Technology & Work, special issue on "Presence: design and technology challenges for cooperative activities in virtual or remote environments", P.Marti (Ed.), Vol.6, pp.15-22, Springer-Verlag.
13. Varni G., Mancini M., Volpe G., Camurri A. (2009) "Sync'n'Move: social interaction based on music and gesture". In Proceedings of the 1st Intl. ICST Conference on User Centric Media (UCMedia 2009), Venice, Italy, December 2009. LNICST vol.40, Springer, 2010. (ISBN 978-3-642-12629-1).
14. Camurri A., Varni G., Volpe G. (2009) Measuring Emotional Entrainment in Small Groups of Musicians. In Proceedings of International Conference on Affective Computing & Intelligent Interaction (ACII 2009), Lisbon.
15. Varni G., Camurri A., Coletta P., Volpe G., (2009) Toward Real-time Automated Measure of Empathy and Dominance. In Proceedings of the 2009 IEEE International Conference on Social Computing SocialCom, Vancouver, Canada, August 2009.
16. Boone, R. T., Cunningham, J. G., (1998) Children's decoding of emotion in expressive body movement: The development of cue attunement, *Developmental Psychology*, 34, 1007-1016.
17. Laban, R., Lawrence, F.C., 1947. Effort. Macdonald & Evans Ltd., London.