MUSCLE movie database: A multimodal corpus with rich annotation for dialogue and saliency detection

D. Spachos, A. Zlatintsi*, V. Moschou, P. Antonopoulos, E. Benetos, M. Kotti, K. Tzimouli,

C. Kotropoulos, N. Nikolaidis, P. Maragos*, I. Pitas

Department of Informatics, Aristotle University of Thessaloniki, Box 451, Thessaloniki 541 24, Greece

E-mail: {dspachos, vmoshou, pantopo, empeneto, mkotti, kattzim, costas, nikolaid, pitas}@aiia.csd.auth.gr

* School of Electrical and Computer Engineering, National Technical University of Athens, Athens 157 73, Greece

E-mail: {nzlat,maragos}@cs.ntua.gr

Abstract

Semantic annotation of multimedia content is important for training, testing, and assessing content-based algorithms for indexing, organization, browsing, and retrieval. To this end, an annotated multimodal movie corpus, the so called MUSCLE movie database, has been collected to be used as a test bed for development and assessment of content-based multimedia processing, such as speaker clustering, speaker turn detection, visual speech activity detection, face detection, facial feature detection, face clustering, scene segmentation, saliency detection, and multimodal dialogue detection. All metadata are saved in xml format following the MPEG-7 ISO prototype to ensure data compatibility and reusability by different users and applications. The entire database can be downloaded through the web for research purposes. Furthermore, we describe a novel annotation tool called Anthropos7 Editor.

1. Introduction

The wide prevalence of personal computers, the decreasing cost of mass storage devices, and the advances in compression techniques have fuelled a vast increase in digital multimedia content, giving rise among others to online music and video stores, personal multimedia collections and video on demand. However, the convenience of multimedia libraries and the functionality of the aforementioned applications will be in doubt, unless efficient multimedia data management, necessary for organizing, navigating, browsing, searching, and viewing the multimedia content, is employed (Benetos, 2008). Multimedia standards such as MPEG-4 and MPEG-7 provide important functionality for manipulation and transmission of objects and the associated metadata, but the extraction of the semantic descriptions and the multimedia content is out of the standard scope (Chang, 2001).

In this paper, we present a large multimodal corpus that has been collected and annotated in order to test and assess different algorithms and hypotheses, such as actor clustering, visual speech detection, dialogue detection, or multimodal saliency detection. Rich annotation by multiple human annotators for concepts such as dialogue manifestations in audio and video, based on the level of background audio, presence of faces, presence of lip activity, is offered. Another concept that is defined in the database is saliency. The database covers 4 distinct modalities, namely audio, video, audiovisual, and text and offers annotated examples for the aforementioned concepts. We also describe a novel video annotation tool named Anthropos7 Editor, which offers capabilities for visual reviewing and editing of MPEG-7 data, following the MPEG-7 ISO format.

The outline of this paper is as follows. Section 2 lists some well known video and audio annotation tools and surveys ANVIL. It also provides a general overview of the Anthropos7 Editor with emphasis to data display and editing using Anthropos7 editor. Section 3 provides a description of the collected movie database. Finally, conclusions are drawn in Section 4.

2. Video annotation tools

A number of video annotation tools have been developed the past years. In addition to the tools reviewed in (Garg, 2004), we mention the following ones: IBM-MPEG-7 Annotation Tool, Ricoh – Movie Tool, ZGDV – VIDETO, COALA – LogCreator, and ELAN. Several factors influence the choice of the annotation tool. First, the tool must be able to support the annotation scheme. Second, it must be user friendly and, in many cases, compatible with other tools. Third, it is desired that the tool can transcribe both audio and video data. Finally, the tool must be suitable for several tasks, such as annotation of speakers and addressees as well as several types of dialogue acts (Garg, 2004). In the following, we survey ANVIL and describe the features of a novel annotation tool called Anthropos7 editor.

ANVIL is a free video annotation tool, used at research institutes world-wide. It offers frame-accurate, hierarchical multi-layered annotation driven by user-defined annotation schemes. The intuitive annotation board shows color-coded elements on multiple tracks in time-alignment. Special features include cross-level links, non-temporal objects and a project tool for managing multiple annotations. ANVIL can import data from the widely used, public domain phonetic tools PRAAT and XWaves, which allow precise and comfortable speech transcription. ANVIL's data files are xml-based. Special ASCII output can be used for import in statistical toolkits (like SPSS).

Anthropos7 Editor is an annotation tool for MPEG-7 advanced viewing and/or editing. It makes viewing and editing of MPEG-7 video content description an easy task.

Such a description can be related to time/duration of scenes and shots, frame-based information, such as the Regions of Interest (ROI) that encompass a specific actor in a frame, and high-level information regarding the video, such as the names of persons or actors appearing in the video. In order to visualize and manipulate time/duration-related information, Anthropos7 Editor uses the Timeline Area. Information based on a single frame, is visualized in the Video Area. Other static movie information, as well as duration and frame-based properties appear in the Static Information Area. These areas communicate with each other, automating various tasks and improving the way the user interacts with the Anthropos7 Editor. For example, the Static Information Area automatically shows the properties of the component the user interacts with; the Timeline area follows the playback of the Video Area. The user may also change the video position from the Timeline Area. Anthropos7 Editor uses overlays on top of the Video Area, e.g. it can visualize the ROI of each actor on every frame, if such information is present in the MPEG-7 file. The user can interact with these ROIs using the mouse. Every 2-D image region that encompasses an actor, or parts of actor's body defined in the Anthropos7 file can be overlaid on the corresponding video frame as a Polygon or a Box (rectangle) and the user can modify its position and its properties, such as the size of the box. A ROI (or parts of it) can be moved or deleted and new ROIs can be added. ROI edges can be also deleted or added. The application automatically tracks all these changes and saves them in the corresponding Anthropos7 file, an xml file in the MPEG-7 format. For more accurate editing, one can use the static ROI property window, which is opened as soon as the user clicks on a ROI. In the current version, ROIs are retrieved only according to the Anthropos7 description of the Actor Instance. No user defined schemas are supported. Apart from a drawn ROI, the name of the associated actor is also depicted on screen. This way, the end user can directly identify ROIs and actors, track face detection results and locate errors.

3. MUSCLE movie database specifications

The basic requirement for the movie database annotation is that the concepts (e.g. dialogue, saliency) must be described in each modality independently as well as in a cross-modal manner. This means that there must be audio-only and video-only descriptions, but audio-visual descriptions as well. This fact emerges from the research community needs to process the same data for different applications. Thus, several modalities along with the corresponding dialogue and saliency annotations are supported: audio-only, video-only, text-only, audio-visual. A more detailed description of these annotations is provided in subsections 3.1 and 3.2, respectively. The movie database and the xml annotation files can be downloaded for research purposes through the URL: http://poseidon.csd.auth.gr/EN/MUSCLE_moviedb.

3.1 Dialogue annotation

In total, 54 movie scenes of total duration 42 min and 41 sec have been extracted from 8 movies from different genres (Table 1). The audio language for all selected scenes is English. The duration of each scene is between 24-123 seconds and the scenes have been carefully selected to represent all possible cases. More details on the movie scenes are listed in Table 1. Each movie scene is separated in two different files: an audio file, which contains the audio of the scene and a video file, which contains the video of the scene without audio.

Movie title	Number of Dialogue scenes	Number of non-dialogue scenes	Scenes per Movie
Analyze That	4	2	6
Cold Mountain	5	1	6
Jackie Brown	3	3	6
Lord of the Rings I	5	3	8
Platoon	4	2	6
Secret Window	4	6	10
The Prestige	4	2	6
American Beauty	10	0	10
Total number of scenes	39	19	58

Table 1: MUSCLE movie database description

Different human annotators worked on the audio and video files. The dialogue type label was added to each one of the scenes (audio and video), one label per scene. The dialogue types for audio are as follows. CD (Clean Dialogue): Dialogues with low-level audio background; BD (Dialogue with background): Dialogue in the presence of a noisy background or music. A monologue is classified as either CM (Clean Monologue), i.e. monologue with low-level audio background or BM (Monologue with background), i.e. monologue in the presence of a noisy background or music. All scenes that are not labeled as CD or BD are considered to be non-dialogue (Non Dialogue - ND). The dialogue types for video are as follows. CD (Clean Dialogue): Two actors are present in the scene, their faces appear simultaneously or in an alternating pattern (A-B-A-B), and there is lip activity; BD (Dialogue with background): At least two actors are present, their faces appear simultaneously or in an alternating pattern in the scene and there is lip activity, while other actors, apart from the two that are engaged in the dialogue, appear. Large intervals where no dialogue occurs might be included in the scene. The monologue types for video are labeled as CM (Clean Monologue), i.e. one actor is present in the scene, his face is visible and there is lip activity or BM (Monologue with background), i.e. at least one actor is present, his face is visible and there is lip activity while other actors might appear and

large intervals where no dialogue occurs might be included in the scene. Similar to audio scenes, all video scenes that are not labeled as CD or BD, including monologues, are considered to be non-dialogue (Non Dialogue - ND).

The extracted annotation metadata for the audio files are speech activity data, namely speech intervals, defined from the start and the end time, for each actor in a scene. For the video files, lip activity data are extracted for each actor (2 actors in each scene maximum), defined through intervals specified by the start and end time and frame. The following three states are used to label each lip activity interval: 0 indicates that back of actor's head is visible; 1 indicates that actor's frontal face is visible, but no lip activity occurs; 2 is indicative of actor's frontal face visibility with lip activity. The structure of the annotation is described in xml format, not following the MPEG-7 ISO prototype.

Afterwards, shot cut information, human face detection, and face tracking information are extracted for all scenes. Shot cut information is extracted using the Shot Boundary module of the DIVA3D software package. The module provides shot boundary detection and shot information management capabilities. The extracted information was subsequently processed by a human annotator that corrected the errors. Human face detection and face tracking information is extracted for each frame using the DIVA3D tracking module. The module allows the user to perform either only automatic human face detection, or to combine the face detection process with face tracking. The face of each actor participating in a dialogue or monologue is assigned a bounding box in each frame of the scene. Face tracking results were edited when needed by human annotators using the Anthropos7 Editor. The extracted data are saved in an xml MPEG-7 compliant manner.

Finally, the two xml files (audio, video) are merged into one xml file for each scene following the MPEG-7 format. The annotations for the two modalities are synchronized since they make use of the same timeline, thus providing joint audio-visual annotation information. Furthermore, the annotation data include the captions for the dialogues and monologues in the scene. It should be noted for the time-being dialogue annotation and captions do not exist for the films The Prestige, and American Beauty.

3.2 Saliency annotation

Saliency annotation is being produced based on manual detection of an audio or visual event that “pops-out”, i.e. which has the unique condition or quality of standing out relative to its environment. Attention in audio signals is focused on abrupt changes, transitions and abnormalities in the stream of audio events, like speech, music, environmental noises in real life or sound effects in movies. The salient features that attract more attention can be detected more clearly. The same observations are valid in case of video signals, where outstanding colors (compared to the background color), abrupt scene changes or movements, or sudden events attract the viewer’s attention (Rapantzikos, 2007).

Three movie clips of total duration ~ 27 min have been selected from 3 different movies of different genres (“300”, “Cold Mountain” and “Lord of the Rings 1”). The clips have been selected after careful consideration; to represent all possible cases of saliency, i.e. visual, audio and audiovisual saliency, as well as smooth alternations between action/non action parts, and dialogue/non dialogue parts to be included. The audio content includes speech in various conditions; speech in form of dialogue, speech with background sound which can be music, noise, other speech or environmental sound. The music content can be found in various conditions too, music with background noise, speech or effects. The background sounds in the clips include environmental sounds such as animals (dog barking, birds singing), autos, knockings, sword sounds etc. and sound effects. The visual content includes a variety of different elements, i.e. abrupt scene changes, computer made light effects and other editing effects.

All movie clips are annotated by two different annotators. No strict duration for the annotation elements is specified, yet an audio event is a bounded region in time that is characterized by a variation or transitional state to one or more sound-producing sources (Evangelopoulos, 2008). An event considered salient is annotated separately, as a means to assign a separate saliency factor. The saliency factor for an audio sequence depends on the impact the sound makes in different scenes and its importance for the annotator. No semantic or linguistic consideration of the content is taken for speech saliency, which is only based on the intensity and strength. Visual saliency concerns pop-out events (pop-out color and pop-out motion) and how salient they are considered by the annotator. Abrupt changes and sudden events can also be regarded as salient. Silence on the other hand, meaning that no significant sound is occurring in the segment is not annotated at all.

The annotators, having already predefined all the above, agree on definitions of the audio and visual events but since each one of them can have an individual opinion about what is salient, likable, interesting or outstanding for the senses, they are free to decide the saliency factor of each event based on their own likes and dislikes. Consequently, annotations from different annotators show some analogy; however since the annotators have different likes and dislikes there are variations on the saliency factor. Such disparities are notable at the annotation of generic saliency where the annotator marks only the parts that bear a saliency factor.

Anvil has been used for saliency annotation. A rich annotation scheme has been defined in order to get all possible saliency factors. The three main saliency categories of the annotation scheme are visual saliency, audio saliency and generic saliency.

Audio saliency is annotated using only the auditory sense; visual saliency only the visual sense while generic saliency is annotated using both modalities simultaneously.

Audio saliency includes a description of the audio type found in a scene. The categories that have been chosen to best fit all possible kinds of sounds in movies are: voice/dialogue, music, noise, sound effect, environmental sound, machine sound, background sound, unclassified sound and mix sound. The annotator has the

opportunity to choose more than one of the above sound types to describe every event, since in a movie up to 5 sounds or more can be detected simultaneously. Thereafter, a factor of high, mid, low or none is assigned for the saliency. Speech saliency is measured by the intensity and loudness of the voice (and defined as extra strong, strong, normal, or reduced). Audio and speech saliency features are presented in Table 2.

Audio Saliency	
Audio type	Voice/Dialogue, Music, Noise, Environmental sound, Machine sound, Background Sound, Unclassified sound, Mix sound
Saliency Factor	None, Low, Mid, High
Speech Saliency	
Actor Id	(actor's numeric label)
Visibility	Visible, Non visible, Voice-Over visible, Voice-Over non visible
Saliency Factor	None, Reduced, Normal, Strong, Extra Strong

Table 2: Audio and speech saliency features

Visual saliency includes a description of the object's motion in every scene. Changes of cast and pop-out events are annotated too. Pop-out events, as stated before, can either refer to color or motion (compared to their environment). Visual saliency is measured as high, mid, low or none. In Table 3, all visual saliency features are presented in detail.

Visual Saliency	
Motion	Start-Stop, Stop-Start, Impulsive event, Static, Moving, Other
Changes of cast	(binary decision)
Pop-out event	(binary decision)
Saliency Factor	None, Low, Mid, High

Table 3: Visual saliency features

Generic saliency is a low-level description of saliency, where the description features are: audio saliency, visual saliency and audiovisual saliency, i.e. when both modalities contribute equally to saliency. Saliency can be measured as high, mid or low. Generic saliency features can be seen in Table 4.

Generic Saliency	
Saliency Type	Visual, Audio, Audio Visual
Saliency factor	None, Low, Mid, High

Table 4: Generic Saliency Features

The above selected audiovisual features have already been proven useful and promising in ongoing experiments aiming at comparing human vs. automatic annotations as well as in testing human evaluations of video summaries. The performance comparison of the audiovisual saliency event detector against the manual annotation on the selected clips showed good agreement. The output of this procedure was a saliency indicator function $I_{sal}(n)$ where n is the temporal frame index. The salient regions were computed automatically by selecting a threshold on the median filtered audiovisual saliency curve. Median filters of different length frames were used. Especially for the longer median filter, the correct frame classification (as salient or non-salient) was up to 80% (Evangelopoulos, 2008).

4. Conclusions

In this paper, MUSCLE movie database was described. It is a multimodal annotated movie database. The fact that MUSCLE movie database encompasses 4 modalities, namely audio-only, video-only, text-only, and audiovisual makes it an efficient test bed for the audio and video research communities. Well known annotation tools are surveyed including a novel tool, named Anthropos7 Editor. Future work will focus on the assessment of agreement/disagreement between annotators for the concepts of dialogue and saliency.

5. Acknowledgment

This work is supported in part by European Commission 6th Framework Program with grant number FP6-507752 (MUSCLE Network of Excellence Project).

6. References

Benetos, E., Siatras, S., Kotropoulos, C., Nikolaidis, N., Pitas, I. (2008). Movie Analysis with Emphasis to Dialogue and Action Scene Detection", in P. Maragos, A. Potamianos, & P. Gros (Eds.), *Multimodal Processing and Interaction: Audio, Video, Text*. N.Y.: Springer.

Chang, S. –F., Sikora, T., Puri, A. (2001). Overview of the MPEG-7 standard. *IEEE Trans. Circuits and Systems for Video Technology*, 11(6): 688–695.

Garg, S., Martinovski, B., Robinson, S., Stephan, J., Tetreault, J., Traum, D. R. (2004). Evaluation of Transcription and Annotation Tools for a Multi-modal, Multi-party Dialogue Corpus. In *Proc. 4th Int. Conf. Language Resources and Evaluation*, pp. 2163-2166.

IBM: MPEG-7 Annotation Tool. www.alphaworks.ibm.com/tech/videoannex

Ricoh: MovieTool. www.ricoh.co.jp/src/multimedia/MovieTool/

ZGDV, VIDETO: Video Description Tool. www.rostock.zgdv.de/ZGDV/Abteilungen/zr2/Produkte/videto/

EPFL, COALA: Content-Oriented Audiovisual Library Access – Log Creator. http://coala.epfl.ch/demos/demosFrameset.htm

ELAN: EUDICO Linguistic Annotator. www.let.kun.nl/sign-lang/echo/ELAN/ELAN_intro.html

ANVIL: The Video Annotation Research Tool. www.anvil-software.de

Rapantzikos, K., Evangelopoulos, G. Maragos, P., Avrithis, Y. (2007). An Audio-visual Saliency Model for Movie Summarization. In *Proc. IEEE Workshop Multimedia Signal Processing*, pp 320-323.

Evangelopoulos, G. Rapantzikos, K., Potamianos, A., Maragos, P., Zlatintsi, A., Avrithis, Y. (2008). Movie Summarization Based on Audio-Visual Saliency Detection, *IEEE Int. Conf. Image Processing*, submitted.