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 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, face clustering, scene segmentation, saliency detection, and visual dialogue detection. All metadata are saved in XML format following the MPEG-7 ISO prototype to ensure data compatibility and reusability. The entire MUSCLE movie database is available for download through the web. Visual speech activity and dialogue detection algorithms that have been developed within the software package DIVA3D and tested on this database are also briefly described. Furthermore, we review existing annotation tools with emphasis on the novel annotation tool 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 [1]. New multimedia standards such as MPEG-4 and MPEG-7 [2] provide important functionality for manipulation and transmission of objects and associated metadata, but the extraction of semantic descriptions and annotation of the multimedia content with the corresponding metadata is out of the scope of these standards. Thus a multimodal corpus with rich annotation becomes a necessity.

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 one or 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. The database covers 3 distinct modalities, namely audio, video, and text as well as concepts such as saliency. The MUSCLE movie database is useful to researchers and developers of related analysis tools.

Furthermore, in this paper, we intend to demonstrate the usability of the developed movie database on visual speech detection and visual dialogue detection algorithms that have been developed. We also review some well known available annotation tools and briefly 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 briefly describes some well known video and audio annotation tools. Section 3 provides a general overview of the Anthropos7 Editor with emphasis to data display and editing using Anthropos7 editor. Section 4 provides a description of the designed movie database and Section 5 describes the use of MUSCLE movie database on visual speech detection and visual dialogue detection with the DIVA3D platform. Finally, conclusions are drawn in Section 6.

2. VIDEO ANNOTATION TOOL SURVEY

A number of video annotation tools have been developed during the last years. Several factors influence the choice of the annotation tool. First, the tool must be able to support the annotation scheme. Second, the tool must be user friendly and, in many cases, compatible with other tools. Third, in some cases, it is desired that the tool can transcribe data from both audio and video files. Finally, the suitability of the tool for several tasks, such as annotation of speakers and addressees and several types of dialogue acts is important [3].

IBM – MPEG-7 Annotation Tool [4] assists annotation of MPEG files with MPEG-7 metadata (both video and audio. Each shot in the video sequence can be annotated with static scene descriptions, key object descriptions, event descriptions, and other lexicon sets. Audio segments can be delimited as well. The annotated descriptions are stored as

MPEG-7 descriptions in an XML file. The tool can also open MPEG-7 files.

ANVIL [5] 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 colour-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).

Ricoh – Movie Tool [6] supports hierarchical segmentation within a timeline-based representation of the video. The automatic shot boundary detection algorithm permits changes to threshold settings. The MovieTool is the most mature and complete of the systems, but has a complicated user interface, which is closely tied to the MPEG-7 specification.

ZGDV – VIDETO [7] hides the complexity of MPEG-7 by basing the description properties on a simple description template, which can then be mapped to MPEG-7 using XSLT. Domain-specific description templates together with their corresponding XSLT mappings are generated. The resulting flexibility, customisability and user-friendliness of this approach are VIDETO's biggest advantages. VIDETO was developed as a research tool to generate video (XML) metadata for testing a video server and retrieval module.

COALA – LogCreator [8] is a web based tool which supports video descriptions. It provides automatic shot detection and a good interface for hierarchical segmentation of videos that can be uploaded to the server, where it is saved in MPEG-7 format in a native XML database. However, it is a domain specific tool, developed specifically for TV news videos with a predefined structure. The descriptors that are used to annotate the different video segments are predefined as well.

3. ANTHROPOS7 EDITOR ANNOTATION TOOL

Anthropos7 Editor is a software package 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, like scenes or shots, information for a single frame, such as the Region of Interest (ROI) that encompasses a specific actor in this frame and information regarding the video itself, such as information regarding persons or actors appearing in the video. In order to visualize and manipulate duration/time 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 single 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 application. 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 an 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 their position and their 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 and track possible face detection and tracking errors.

4. 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 4.1 and 4.2, respectively. The movie database and the XML annotation files are available for downloading through the URL: http://poseidon.csd.auth.gr/EN/MUSCLE_moviedb.

4.1 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 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 presented 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.

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 back ground 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 and large intervals where no dialogue occurs might be included in the scene. The monologue types for video are labelled 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 labelled as CD or BD, including monologues, are considered to be non-dialogue (Non Dialogue - ND).

Table 1 – MUSCLE movie database description

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**

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 but no 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 (Section 5.1). 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 that no dialogue annotation and captions exist for the films The Prestige, and American Beauty.

4.2 Saliency annotation

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 colours (compared to the background colour), abrupt scene changes or movements, or sudden events attract the viewer’s attention [9].

Three movie clips of total duration 27 min and 14 sec 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. The smooth alternations between action and non action parts, as well as dialogue and non dialogue parts have also been taken under consideration while choosing the clips. All movie clips were annotated by two different annotators. Anvil was used as the annotation tool. 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.

Visual saliency is annotated using only the visual sense; audio saliency only the auditory sense while generic saliency is annotated using both modalities simultaneously. Visual saliency includes a description of the motion of the object seen in the scene. Changes of cast or pop-out events can be annotated too. Pop-out events are the salient events, and they may include outstanding colours (compared to the background), abrupt movements or changes. Saliency in a scene is measured as either high, mid, low or none. In Table 2, all visual saliency features are presented in detail.

Table 2: Visual Saliency Features

Visual Saliency
Motion: Start-Stop, Stop-Start, Impulsive event, Static, Moving, Other
Changes of cast

Pop-out event
Saliency Factor: None, Low, Mid, High

Audio saliency includes a description of the audio type heard in the scene. The categories that have been chosen to best fit all possible kinds of movie sounds are: voice/ dialogue, music, noise, sound effect, environmental sound, machine sound, background sound, unclassified sounds, and mix sound. There is the possibility to choose many different kinds of sounds since in a movie can be heard simultaneously up to 5 sounds or more. After choosing the type of sound a factor of high, mid, low or none is chosen to represent 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 3.

Table 3: Audio and Speech Saliency Features

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
Visibility: Visible, Non visible, Voice-Over visible, Voice-Over non visible
Saliency Factor: None, Reduced, Normal, Strong, Extra Strong

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

Table 4: Generic Saliency Features

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

The above selected audiovisual features have already been proven useful and promising in some of our ongoing experiments on comparing human vs. automatic annotations, as well as on human evaluations of video summaries.

5. VISUAL SPEECH DETECTION AND DIALOGUE DETECTION ALGORITHMS

A visual speech detection module and a visual dialogue detection module have been developed within the DIVA3D software package. DIVA3D is a software package for digital video processing and analysis. DIVA3D modules are application extensions of DIVA3D and they help to extend the usability and the operations that can be performed by it. For example, such modules are the visual speech and dialogue detection ones, which have been applied to the MUSCLE movie database in order to illustrate its usefulness as a test bed for performance evaluation.

The Visual Speech Detection module is used to detect speech on a video stream by applying signal detection algorithms on a simple and easily extracted feature from the mouth region. More specifically, the method utilizes the fact that the increased average value and standard deviation of the number of pixels with low intensities that the mouth region of a speaking person demonstrates can be used as visual cues for detecting visual speech. Results are saved in an XML MPEG-7 file that contains information about the mouth coordinates and the lip activity, according to the Anthropos7 structure. The method assigns to each mouth region detected in a frame one of the following two labels: “No Lip Activity Present” and “Lip Activity Present”.

The dialogue detection module is used to detect dialogue, monologue or declare that no speech occurs in a video stream based on the lip activity labels of each frame and the patterns of lip activity intervals for the actors in the scene, e.g. whether two actors exhibit lip activity in an overlapping or alternating manner. The results of the Dialogue Detection module are saved in XML MPEG-7 format. The dialogue information is saved as follows: each appearance of an actor is assigned the value “Dialogue” if dialogue is detected, “Monologue”, if the actor speaks but the other actors in its “temporal” neighbourhood are silent, and finally, “No_Dialogue” if there is no lip activity for this appearance of the actor.

6. 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. MUSCLE movie database was used to test visual speech detection and visual dialogue detection modules.

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

REFERENCES

[1] E. Benetos, S Siatras, C. Kotropoulos, N. Nikolaidis, and I. Pitas, ”Movie analysis with emphasis to dialogue and action scene detection”, in P. Maragos, A. Potamianos and P. Gros (Eds.), *Multimodal Processing and Interaction: Audio, Video, Text.* N.Y.: Springer, 2008.

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

[3] S. Garg, B. Martinovski, S. Robinson, J. Stephan, J. Tetreault, D. R. Traum, “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.

[4] IBM MPEG-7 Annotation Tool. www.alphaworks.ibm.com/tech/videoannex

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

[6] Ricoh MovieTool. www.ricoh.co.jp/src/multimedia/MovieTool/

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

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

[9] K. Rapantzikos, G. Evangelopoulos, P. Maragos, and Y. Avrithis, “An audio-visual saliency model for movie summarization”, in Proc. *IEEE Workshop Multimedia Signal Processing, 2007*, pp 320-323.