Quality Evaluation of Computational Models for Movie Summarization

A. Zlatintsi*, P. Koutras*, N. Efthymiou*, P. Maragos*, A. Potamianos*, and K. Pastra†
* School of Electr. & Comp. Enginr., National Technical University of Athens, 15773, Athens, Greece
†Cognitive Systems Research Institute, Athens, Greece
Email: [nzlat,pkoutras,maragos]@cs.ntua.gr, [nefthymiou,potam]@central.ntua.gr, kpastra@csri.gr

***Abstract*—In this paper we present a movie summarization system and we investigate what composes high quality movie summaries in terms of user experience evaluation. We propose state-of-the-art audio, visual and text techniques for the detection of perceptually salient events from movies. The evaluation of such computational models is usually based on the comparison of the similarity between the system-detected events and some ground-truth data. For this reason, we have developed the MovSum movie database, which includes sensory and semantic saliency annotation as well as cross-media relations, for objective evaluations. The automatically produced movie summaries were qualitatively evaluated, in an extensive human evaluation, in terms of informativeness and enjoyability accomplishing very high ratings up to 80% and 90%, respectively, which verifies the appropriateness of the proposed methods.**

I. Introduction

Summarization refers to generating a shorter version of a video that includes as much as possible information required for context understanding without sacrificing much of the original informativeness and enjoyability. Automatic summaries can be generated either with key-frames, which correspond to the most important video frames and represent a static storyboard, or by video skims that include the most descriptive and informative video segments. Movie data are multimodal, containing visual, audio and textual streams, and many computational models have been proposed to estimate their multimodal saliency [1], [2], [3]. Besides their sensory cues, movies contain semantic events as well, whose modeling is difficult using only bottom-up and data-driven techniques, thus it is usually needed to incorporate high-level information.

There are many qualities that a movie has to include in order to give a pleasurable experience to the viewer. In exactly the same way, a movie summary, produced either by a human or automatically by a system, has to consist of features that will attract human attention, but also incorporate elements that assist the development of the plot. The features to be included and the techniques that are used for such a system are closely related to user experience. Hence, a computational summarization system could indeed benefit and get further improved through qualitative human evaluations of the automatically produced summaries. First, the developer needs to know what is conspicuous and attracts human attention as well as to have some ground-truth data for quality testing of his/her methods. Likewise, at the final stage he/she has to evaluate the system considering user responses and preferences in order to further improve it. The classical machine learning techniques can evidently assist such an evaluation, yet they cannot really account for the human factor. Nonetheless, human perspective is needed for the implementation of systems that takes into consideration user preferences, and produce "user-defined" summaries. In this paper, we present novel ways for the integration of user experience in movie summarization. Specifically, we propose a computational system for movie summarization and we introduce a movie database, enriched with salient event annotation in the sensory and semantic level The evaluation of the produced summaries is based both on a machine learning technique and on extensive qualitative user experience evaluations that verify the appropriateness of the proposed methods and the quality of the summaries.

This research work was supported by the project "COGNIMUSE" which is implemented under the "ARISTEIA" Action of the Operational Program Education and Lifelong Learning and is co-funded by the European Social Fund and Greek National Resources.

II. Database Description

Event detection and summarization algorithms can be significantly improved when there is adequate data for training, adaptation and evaluation of their parameters. The evaluation of the developed computational models is usually based on the comparison of the similarity or correlation between the system-detected observations and some ground-truth data (annotated reference event observations) selected by experienced/trained users. For this reason, we developed the *MovSum (Movie Summarization) Database*, which at this point is still under development, and part of an involving multimodal video oriented database annotated with saliency, semantic events and cross-media relations. The database at its current state has been used for objective evaluation of the system-detected salient events.

A. MovSum Database Annotated With Salient Events

Data collection: The process of creating the dataset includes data collection, data conversion to a suitable format and annotation. Specifically, the dataset consists of half-hour continuous segments from seven movies (three and a half hours in total), namely: "A Beautiful Mind" (BMI), "Chicago" (CHI), "Crash" (CRA), "The Departed" (DEP), "Gladiator" (GLA), "Lord of the Rings - the Return of the King" (LOR) and the animation movie "Finding Nemo" (FNE)[1]. Oscar-winning movies from

[1]Title, production year and production company of the seven movies: A Beautiful Mind 2001 (Universal & DreamWorks), Chicago 2002 (Miramax), Crash 2004 (Lions Gate), The Departed 2006 (Warner Bros.), Gladiator 2000 (Universal & DreamWorks), Lord of the Rings 2003 (New Line), Finding Nemo 2003 (Walt Disney Pictures, Pixar Animation Studios).

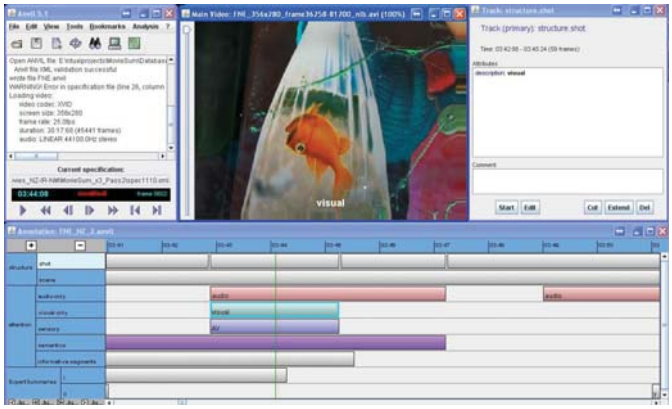


Fig. 1: MovSum Database annotation layers in "Finding Nemo".

various film genres (drama, musical, action, epic, animation) were selected to form a *systematic, genre-independent database* of acclaimed, high production quality videos. The sample database videos were chosen as continuous half-hour segments (with the final shot/scene included), and were taken from the official DVD releases. For reference purposes the exact time sequences were noted. The movie segments were ripped and saved in .avi format using particular technical specifications, in high resolution for summary visualization and rendering, and small resolution for processing and annotation. The database also includes the movie subtitles.

The specific movies were selected partly because of their popularity and partly because of their plot structure, which is made exclusively for the establishment of the emotional disposition themes of the characters. They all include three basic components: the main character/s, the desire and the conflict. Furthermore, Hollywood movies include some typical features such as music, vivid color variations, audio and visual effects, speed of action etc., which are used as a powerful tool for developing the plot. Such features and structure can thus lead to effective summaries.

The database has been annotated with respect to saliency (in a binary mode) and it consists of monomodal and multimodal saliency and semantic annotation including scene and shot segmentation. For this purpose Anvil (http://www.anvil-software.de/index.html) video annotation interface has been used. Although this kind of annotation is considered highly subjective – user preferences on what is salient cannot really be dictated – the three trained annotators could consult an instruction's manual consisting of definitions and guidelines (through examples) for each individual layer, in order to achieve a high degree of annotation uniformity.

Annotation Layers

(a) Sensory information: This is the pre-attentive layer of saliency, where the annotation process is done quickly, effortlessly, without any focused attention or thoughts and with little or no searching required. The annotation is based only on movie elements that capture the viewers attention instantaneously or in segments. It includes monomodal (audio, visual) and multimodal (AV) saliency of the sensory content, depending on the importance and the invoked attention they create to the annotator. The *audio-only saliency* (A) is annotated by only listening to the audio stream of the movie segment, and it includes acoustically interesting segments. Semantic information is not taken into account, thus, the annotator is instructed not to pay attention either to the type of the sound (e.g., speech, music) or its meaning (e.g., dialogue, genre of music). The *visual-only saliency* (V) is annotated by only watching the movie segment and consists of segments that are visually interesting, again without taking into consideration semantic information. The *audiovisual* (AV) annotation layer, is a common global saliency measure, that handles the two modalities (audio, video) as one multimodal cue. Examples of mono- and multimodal salient events include features such as loudness, pitch variations and sound effects either artificial or natural (aural cues); contrast, intensity, motion, color (visual cues) and combined audiovisual events, artificial or not.

(b) Semantic information: This layer includes segments that are conceptually important as stand-alone semantic events that have a beginning, a steady state and an ending, as a sequence of conceptual events not necessarily important just for the examined movie but generally, as an objective, direct or indirect meaning. Such semantically salient events include important names, plot elements, phrases, symbolic information, gestures, facial expressions (something that indicates a feeling), actions that boost the movie forward etc. This layer, for our objective evaluations, is used combined with the sensory AV layer, so as to include segments that are conceptually important as stand-alone sensory/semantic events, and henceforth referred to as audio-visual-semantic events (AVS).

(c) Informative-segments: It includes segments important for understanding the narration-plot of the specific half-hour movie clip. Informative-segments are usually a subset of the semantic information annotated in ***(b)***. This layer could be considered as a manually generated skim consisting of descriptive but not necessarily enjoyable segments, which we like because of their composition, as for instance: important statements and objects, repeated actions etc.

(d) Affective information: Both intended and experienced emotions have been annotated. Details on the followed scheme and the associated emotion tracking task are provided in [4].

Movie Structure: The movie clips were manually segmented into shots and scenes. A shot is the basic building block in a movie and can be defined as the interval between editing transitions (e.g., cut, fade etc). A shot or series of shots constitute a unit of continuous related actions. A scene, is defined as a complete, continuous chain of actions (shots) that occur at the same place and time. The average shot and scene duration in the movies are 2.5 sec. and 3.5 min., respectively.

Expert Summaries: For the seven movie clips, summaries (ca. five minutes long) created by an experienced user (professionally associated with film production and editing) are available. The expert user was instructed to create a summary in relation to the plot of the thirty minutes segment, according to his preferences, which could vary between 1–10 minutes. Since a summary and not a movie trailer was requested to be produced, he was urged to omit segments with strong audio/visual effects that usually attract the viewer, unless they contained important information for the development of the plot.

Table I shows the percentage of the annotated salient frames (labeled by at least two annotators) and the average (pairwise) correlation agreement between the annotators – overall satisfactory, considering the subjectivity of the task – for each movie and annotation layer. Note that the agreement

TABLE I: Statistics for MovSum Database annotated with salient events.

Percentage (%) of Salient Frames								
Layer	BMI	CHI	CRA	DEP	GLA	LOR	FNE	Mean
A	25.4	56.3	55.0	33.4	60.9	58.3	54.6	49.1
V	30.1	46.3	37.9	32.4	39.2	43.3	36.9	38.0
AV	27.4	47.7	43.1	37.8	49.6	50.7	39.7	42.3
AVS	63.2	76.6	64.8	71.8	68.5	72.7	67.6	69.3
Average (pair-wise) Correlation Between Annotators								
A	0.54	0.48	0.46	0.49	0.51	0.52	0.42	0.49
V	0.31	0.33	0.32	0.45	0.38	0.43	0.38	0.37
AV	0.45	0.45	0.41	0.54	0.44	0.50	0.44	0.46
AVS	0.29	0.24	0.27	0.29	0.31	0.33	0.23	0.28

is higher for the sensory (A, V, AV) layers compared to the sensory-semantic (AVS) layer. However, the ground-truth saliency indicator functions, used for subjective evaluations, consist of frames that have been labeled salient by at least two labelers. Thus, despite the lower agreement between annotators observed for certain movies, the final saliency ground-truth was formed on the basis of consistently-labeled salient frames only. Finally, a full movie (Gone with the Wind) is currently annotated with salient events, so as to be able to evaluate our computational system on a full-scale complete movie.

B. MovSum Database annotated with cross-media relations

Every day communication between people is a combination of different modalities. As a first step towards investigating the fusion of information between different modalities (audio, video, text) we label a subset of our movie database with crossmodal labels, using the COSMOROE cross-media framework [5]. Next we describe the annotation process of semantic interrelations between the different modalities and specifically language, image, body movements and acoustic events. Because of the demanding and time-consuming nature of the annotation process, the subset of movies used for COSMOROE-based annotation are: "Gone with the Wind" (GWTW) (1939) (total duration 1:44:15) and "Gladiator" (GLA) (2:28:36). For COSMOROE based annotation the use of full movies is essential since we want to study cross-media semantic interplay at a full-scale level, which will allow us to a) make valid observations on which relation types are more frequently used in a specific genre and b) explore potential interaction patterns among relations as the movie evolves.

In the COSMOROE framework, the three majors interaction relations are *Equivalence, Complementarity* and *Independence*; described next using examples from GWTW for better understanding. For the needs of such annotation, several visual and language units are segmented and annotated, including: a) Utterance (i.e., spoken language transcription, words or phrases), b) Graphic or Scene Text (shown on the video), c) Frame Sequence, which equals to shots and participate in cross-media relations, d) Key Frame Region, depicting a particular object of interest in a sequence of frames, e) Body Movements and Gestures (i.e., hand gestures, head movements or body movements that participate in a relation) and f) Acoustic Events based on five main categories, which are: animal sounds, human sounds, natural/environmental sounds, machine sounds, and general background sounds, including music or any other type of sound.

Equivalence (Literal or Figurative): Different modalities or media can express semantically equivalent information. Starting with **literal equivalence**, two cases have been distinguished; *Token-Token* in which the media refer exactly to the same entity, uniquely identified as such, and *Type-Token* in which one medium provides the class of the entity, action or feature expressed by the other. Examples that help us distinguish the two relations are: a token-token relation would have been if we had an acoustic event (e.g., barking) and the visual of the event (i.e., a dog barking) as in Fig. 2a; while a type-token relation would be annotated when someone said the word *"dog"* and an image of a dog was depicted in a sequence of frames. **Figurative equivalence** includes the relations of *Metonymy* and *Metaphor*; each modality refers to a different entity, but the viewer and the creator of the message considers these two entities as semantically equal. The most common metonymic pattern is the *metonymy "part for whole"*, as for instance when the image shows a part of an entity and the language refers to the whole (e.g., the word *"land"* and an image that depicts only a part of the land, see Fig. 2b). Finally, *metaphor* relations, which actually cannot be found that often, are annotated when a modality draws a similarity between two referents belonging to different domains.

Complementarity: The information expressed in one medium is complement to the information expressed in another medium. Four different sub-relations clustered into two groups can be found; those in which complementarity between the information expressed by each medium is *essential* for forming a coherent multimedia message and those in which complementarity is *non-essential*. Specifically, *Exophora* includes cases of "anaphora" in which one medium resolves the reference made by another (essential or non-essential). For example, the demonstrative word *"That"* does not express the specific object that is pointed at, however this information is provided by the image as in Fig. 2c, where the woman says *"That"* while pointing at a dress. In *Agent-Object* relations, one modality reveals the subject or the object of the other modality, e.g., in GWTW we hear: *"Scarlett, you look"*; the object, "a piece of paper", is omitted in the sentence, but it is depicted on the image, see Fig. 2d. *Defining Apposition* is called a relation when the extra information provided by one medium identifies or describes someone or something. When one medium reveals a generic property or characteristic of the very concrete entity mentioned by another, a *Non-Defining Apposition* relation is presented. *Adjunct* is a non-essential relation that denotes an adverbial-type modification.

Independence: Each medium carries an independent message and their combination creates a coherent multimedia message. This relation consists of three subtypes: *Contradiction*, when one medium refers to the exact opposite of another or to something semantically incompatible. *Symbiosis*, when different pieces of information are expressed by the media (i.e, when the actors discuss about something that is not depicted on the image), and *Meta-Information*, which is actually a type of relation that cannot be found in movies.

Table II shows statistics of the relations found in GWTW for the full duration of the movie. 500 relations have been found with average duration ca. 16 sec, referring to over one hour in the movie. The symbiosis relation is not annotated, and thus omitted from Table II, since it includes all events that do not belong in any other relation. Furthermore, it cannot be utilized, as it does not provide any useful information. The cross-media annotation will be used for the integration of top-down information, for the automatic object/action prediction given specific accompanying text as well as the prediction of

(a) Token-Token (b) Metonymy, part for whole (c) Complement., Exophora Ess. (d) Complement., Agent-object Ess.

Fig. 2: Examples of the most conspicuous relations of the COSMOROE cross-media relations in "Gone with the Wind".

TABLE II: Statistics of the COSMOROE cross-media relation in GWTW.

COSMOROE Relations	Subtypes	Percentage	
Equivalence	Token-Token	29.2	76.4
	Type-Token	27.4	
	Metonymy	19.0	
	Metaphor	0.8	
Complementarity	Exophora	14.4	23.4
	Agent-Object	1.2	
	Adjunct	2.8	
	Appostition	5.0	
Independence	Contradiction	0.2	0.2

the semantics of text segments given specific accompanying visual objects/actions. Additionally, we intent to investigate which of those relations are the most important to be included in a movie summary. For more details and examples regarding the COSMOROE annotation framework refer to [5], [6].

III. System Overview

A. Multimodal Analysis

Visual Analysis: For visual analysis an energy-based model for spatio-temporal visual saliency estimation is used, based on Itti et al. model [7], which is more relevant to the cognition-inspired saliency methods. It uses biologically plausible spatio-temporal filters, like oriented 3D Gabor filters, in order to extract visual features. In a first phase the initial RGB video volume is transformed into Lab space and split into two streams: luminance and color contrast. Then follows the core stage of our perception-inspired frontend for visual saliency [8], which is applied both on luminance and color contrast channels. This process can be divided into three individual steps. The first step consists of the Spatio-Temporal Gabor filtering [8], [9], while the others include postprocessing procedures like Quadrature Pair Energy computation and Dominant Energy selection followed by a temporal moving average applied on the resulting raw energies. In the last stage the produced energy maps can be mapped to a 1D map giving time-varying saliency features. We employed a simple 3D to 1D mapping by taking the mean value for each 2D frame slice of each 3D energy volume. The resulting temporal sequence of feature vectors, each corresponding to 4 different energies, along with its first and second time derivatives comprise the features set for the visual modality.

Audio Analysis: The issue of saliency computation in the audio stream is approached as a problem of assigning a measure of interest to audio frames, based on spectro-temporal cues. For the analysis and saliency-modeling of the audio stream an energy-based feature set was used, based on the nonlinear Teager-Kaiser differential energy operator [10], [11]. Since Teager energy is only meaningful in narrowband signals [11], the application of the operator is preceded by multi-band filtering with a filterbank of 25 linearly spaced Gabor filters, in order to isolate the signal's narrowband components. Then the energy operator is estimated at the outputs and the average for the frame duration gives a measure of each channel activity; the mean instantaneous energies (25 features in total). Moreover, we computed two additional perceptual features which are known to correlate to the functioning of the human auditory system. The first one is roughness proposed in [12] and reported to be associated with human attention; which is an estimation of the sensory dissonance of a sound. The second one is loudness, also associated with attention, corresponding to the perceived sound pressure level [13].

Text Analysis: In this work, we extend the text analysis of [3] – which is used for part-of-speech tagging, where each word is assigned a value of importance – and we include affective modeling of single words extracted from the subtitles information available with each movie distribution. A word w is characterized regarding its affective content in a continuous (within the $[-1, 1]$ interval) space consisting of three dimensions, namely, valence (v), arousal (a), and dominance (d) (affective features). For each dimension, the affective content of w is estimated as a linear combination of its semantic similarities to a set of K seed words and the corresponding affective ratings of seeds (for the corresponding dimension), as in [14]. The employed model is based on the assumption that *"semantic similarity can be translated to affective similarity"* [14]. The words' affective ratings were estimated using as seeds 600 entries selected from the ANEW lexicon [15]. More details about the corpus, seed selection, and the training of weights can be found in [14].

B. Machine Learning Approach

For the multimodal salient event detection we follow a non-parametric data-driven classification approach. The resulting temporal sequence of audio-visual (AV) features (27 audio and 4 visual features) along with their first and second temporal derivatives and the 4 text (T) features, comprise the feature set for the classification process. We employ a K-Nearest Neighbor Classifier (KNN) independently for the AV and T features, following similar framework as in [16], [3]. Specifically, we consider framewise saliency as a two-class classification problem, and a seven fold cross-validation is adopted by using the labeled frames from six movies (of the MovSum database) and tested on the seventh. In order to obtain results for various compression rates, a confidence score is defined for each classified frame.

C. Movie Summarization Algorithm

The new summarization algorithm extends our baseline algorithm presented in [3] and it includes features, which

make the summaries smoother (regarding audio and video transitions), while also enhancing the comprehension of the semantics. For the creation of the summaries we use the outputs of the classifier, which consist of the frames classified as salient. Thus, we use frames (chosen based on high confidence scores), as an indicator function curve that marks the most prominent audio-visual and text events. The pre-processing steps that we followed are: **1)** Median filtering of the audiovisual confidence scores C_{AV} followed by scene-based normalization (scenes' information is extracted from the database). **2)** Text confidence scores C_T that are only trained on speech segments are used, while frames where no speech existed are set to zero. **3)** Late fusion of the AV and T modalities is performed, where a fixed weight w for the text modality is chosen: $C_{AVT} = C_{AV} + w \cdot C_T$. In this paper we experimented using as weight $w = 0.10$ or $w = 0.20$.

In order to create summaries that do not include only salient events but semantically coherent as well, we perform correction of the boundaries of the selected segments. This is achieved using ideas from mathematical morphology and specifically, the reconstruction opening: $\rho^{-}(M|X) \triangleq$ connected components of X intersecting M [17], [16]. Hence, we can extract large-scale components by knowing only smaller markers inside them. Specifically, we use as marker M the raw salient events and as reference X the manually segmented shots and the automatically aligned single-word level boundaries [3]. The *reconstruction* performed is regarded significant for the performance of the summaries, as also shown on the results of the human evaluation, especially for the comprehension of the semantics and the creation of smoother transitions [16].

Concluding, the steps of the summarization algorithm (after the pre-processing) are: **a)** sorting of the confidence scores so as to define the raw salient segments to be included in the summary. **b)** Shot reconstruction and **c)** "speech reconstruction" (as described) ensuring that no words will be "clipped". **d)** The final step of the algorithm for the combination of frames/segments into the final continuous summary, is based on the final steps of the algorithm described in [3].

IV. Results and Discussion

A. Objective Machine Learning Evaluation

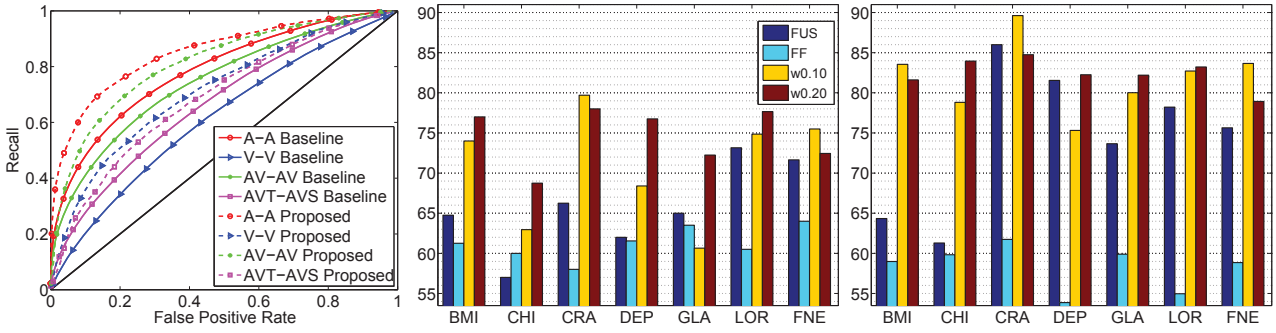
Figure 3a shows Receiver Operating Characteristic (ROC) curves for saliency classification, while changing the percentage of frames in summary (between 1–100%), for audio on audio (A-A), visual on visual (V-V), audiovisual on audiovisual (AV-AV) and audiovisual-text on audio-visual-semantics (AVT-AVS) annotation. The results for the proposed method (AV-AV and AVT-AVS) are produced using the movie summarization algorithm, while for the A-A and V-V results we use the sorted median filtered confidence scores. For the baseline method the results are shown for the sorted raw confidence scores (without any further processing). We note that the proposed system outperforms the baseline system [3] both when evaluating each modality individually as well as when two (AV) or three (AVT) modalities are combined. However, greater improvement can be seen for the monomodal salient event detection than the multimodal one; specifically, best performance is accomplished for the audio modality (A-A evaluation). Moreover, we observe that the audiovisual modality (AV-AV) manages to yield a quite as high score as well. We assume that the proposed system's improvement compared to [3] is due to the advanced monomodal frontends, in all modalities, that outperformed the baseline and the new and more carefully designed movie summarization algorithm, which accounts for smoother transitions of the selected segments but also corrects their boundaries when speech is present. We have to highlight however the fact that the classification approach used here is a framewise detection task, while the human annotators labeled the salient events as segments and not as single frames.

B. Subjective Qualitative Evaluation of Movie Summaries

Summaries obtained five times faster than real time were subjectively evaluated by 20 users in terms of informativeness and enjoyability on a 0–100% scale, similarly to [3]. In total, four summaries were evaluated, namely: two summaries based on the proposed method using different weights for the text modality, where $w = 0.1$ or 0.2, the best performing summary produced using the fusion methods (FUS) presented in [3] (the summaries were chosen based on the best enjoyability results), and a fourth fast-forward like summary (FF), which was created by subsampling 2 seconds every 10 seconds of the original clip. The subjects participating in the evaluation first viewed the original half-hour clip, for each of the movies, followed by the four summaries (ca. 6 min. each) in randomized order. To better normalize the ratings, the following scale was communicated to the subjects: poor between 0–40%, fair 40–60%, good 60–75%, very good 75–90% and excellent 90–100%, while they were also asked to give their scores in a ranking order.

In Fig. 3b and 3c we observe that the proposed method performs much better in terms of both metrics compared to the best performing summaries based on fusion and the fast-forward like summaries. Specifically, they achieved very high subjective ratings, up to 80% for informativeness and 90% for enjoyability. Regarding the proposed method we observed that the assignment of different weights in the text modality is important and it relates to the movie genre; usually a smaller weight is needed for a dialogue based movie than an action movie. This is because in action movies that usually include battles or scenes with high level audio and music (i.e., GLA, LOR, DEP, BMI and CHI), the algorithm tends to favor those high intensity events, probably because of the sharp scene changes, the high-intensity color motifs and the audio effects. So a higher text weight in such movies has as an immediate consequence the incorporation of many more textual salient events, such as dialogue segments, resulting also in summaries that cover more scenes from the original clip, instead of focusing only on those high intensity events. On the other hand, in movies such as Crash (CRA), which is a crime/drama movie with long dialogue scenes, smaller text weight is needed because otherwise the few action scenes (such as explosions, gun shootings) are omitted. User comments confirmed that a good summary has to be balanced regarding the different types of events. Additionally, the boundary correction of the selected segments, that was achieved through the reconstruction of shots and speech segments, contributed a lot to the enjoyability, since it resulted to smoother transitions, in both audio and video streams, and to semantically coherent events that aid the comprehension of the plot.

Concerning the summaries based on fusion, the subjects commented that they were enjoyable enough, however not as informative, which is actually reflected on the presented results



(a) ROC Curves (b) Informativeness (c) Enjoyability

Fig. 3: Objective and subjective evaluation results by 20 humans. (a) Saliency classification ROC curves for the different modalities. (b) Informativeness and (c) enjoyability results of AVT summaries at (×5) rate, were FUS denotes the best summary obtained using fusion methods as presented in [3], FF denotes "fast-forward", while the two newly produced summaries are differentiated by the weight used in the text modality.

(see Fig. 3b,3c). The specific summaries were created using scene (or shot) based fusion, hence they tended to keep longer and semantically complete segments only from scenes with salient characteristics; something that made the absence of important plot elements apparent to human evaluators.

Regarding the fast-forward like (FF) summaries only a few of the subjects realized that they were intentionally added for evaluation (as a naive approach indicating a lower bound for our metrics). In this way, we managed to prove that a uniform sampling of movie frames is not adequate in order to create an acceptable summary. However, whenever the summaries were assigned a high score, regarding informativeness, it was because they included visual information uniformly taken from the whole original clip; a significant observation telling us that a summary needs to include elements, more or less, from the full duration of the original clip. Yet, even in these cases the subjects judged them as "choppy", with too fast transitions and non-existing semantics.

Human quality evaluation of a movie summarization system, as shown here, is essential for improving the quality of the produced summaries. User comments are crucial in order to develop systems that include such features that will heighten the human experience, but also for the creation of summaries that consist of user-defined and preferred content.

V. Conclusions

In this work, we present a movie summarization system, that uses advanced techniques, and through human quality evaluations we investigate how such a system can benefit and be improved in order to heighten human experience. Moreover, we describe the MovSum database, an additional contribution of this paper, which consists of human saliency annotation as well as a crossmodal semantic analysis. Our experimental evaluation using human saliency annotation as ground-truth – denoting conspicuous events – confirms the adequacy of the proposed algorithms. The framework shows to be promising as it outperforms other state-of-the-art methods over the MovSum database. The human quality evaluation of the automatically produced movie summaries quantitatively verifies the appropriateness of both the proposed movie summarization algorithm and the multimodal saliency annotated database. For future work, we intend to extend the database with more movies as well as expert user annotations. Finally, based on the human evaluation we aspire to further refine our methods for movie summarization, in order to be able to produce *user-defined*, *high-quality* summaries.

Acknowledgment

The authors would like to thank E. Iosif for his contribution on affective text analysis and the students of NTUA for participating in the subjective evaluation and for their valuable comments regarding the summaries.

References

[1] Y. Wang, Z. Liu, and J.-C. Huang, "Multimedia content analysis using both audio and visual clues," *IEEE Signal Process. Mag.*, vol. 17, 2000.

[2] Y. Ma, X. Hua, L. Lu, and H. Zhang, "A generic framework of user attention model and its application in video summarization," *IEEE Trans. on Multimedia*, vol. 7(5), pp. 907–919, Oct. 2005.

[3] G. Evangelopoulos, A. Zlatintsi, A. Potamianos, P. Maragos, K. Rapantzikos, G. Skoumas, and Y. Avrithis, "Multimodal saliency and fusion for movie summarization based on aural, visual, textual attention," *IEEE Trans. on Multimedia*, vol. 15(7), pp. 1553–1568, 2013.

[4] N. Malandrakis, A. Potamianos, G. Evangelopoulos, and A. Zlatintsi, "A supervised approach to movie emotion tracking," in *ICASSP*, 2011.

[5] K. Pastra, "COSMOROE: a cross-media relations framework for modelling multimedia dialectics," *Multimedia Systems*, vol. 14(5), 2008.

[6] ——, "COSMOROE Annotation Guide," Cognitive Systems Research Institute, CSRI-TRS-150201, CSRI, Tech. Rep., 2015.

[7] L. Itti, C. Koch, and E. Niebur, "A model of saliency-based visual attention for rapid scene analysis," *IEEE Trans. Pattern Analysis and Machine Intelligence*, vol. 20(11), pp. 1254–1259, 1998.

[8] K. Maninis, P. Koutras, and P. Maragos, "Advances on action recognition in videos using and interest point detector based on multiband spatio-temporal energies," in *Proc. Int'l Conf. Image Processing*, 2014.

[9] D. J. Heeger, "Model for the extraction of image flow," *J. Opt. Soc. Amer.*, vol. 4, no. 8, pp. 1455–1471, 1987.

[10] J. Kaiser, "On a simple algorithm to calculate the energy of a signal," in *Proc. IEEE Int'l. Conf. Acoust., Speech, Signal Process.*, 1990.

[11] P. Maragos, J. Kaiser, and T. Quatieri, "Energy separation in signal modulations with application to speech analysis." *IEEE Trans. Signal Process*, vol. 41, p. 30243051, 1993.

[12] R. Plomp and W. Levelt, "Tonal consonance and critical bandwidth," *Jour. Acoust. Soc. of Am. (JASA)*, vol. 38, pp. 548–560, 1965.

[13] E. Zwicker and H. Fastl, *Psychoacoustics, Facts and Models*. Springer, 2nd edition, 1999.

[14] N. Malandrakis, A. Potamianos, E. Iosif, and S. Narayanan, "Distributional semantic models for affective text analysis," *IEEE Trans. Audio, Speech, and Language Process.*, vol. 21(11), pp. 2379–92, 2013.

[15] M. Bradley and P. Lang, "Affective norms for English words (ANEW): Stimuli, instruction manual and affective ratings. Tech. report C-1." The Center for Research in Psychophysiology, Univ. of Florida, 1999.

[16] A. Zlatintsi, P. Maragos, A. Potamianos, and G. Evangelopoulos, "A saliency-based approach to audio event detection and summarization," in *Proc. European Signal Process. Conf.*, 2012.

[17] P. Maragos, *The Image and Video Processing Handbook*, 2nd ed. Elsevier Acad. Press, 2005, ch. Morphological Filtering for Image Enhancement and Feature Detection, pp. 135–156.