Zlatintsi *et al. EURASIP Journal on Image and Video Processing* (2017) 2017:54
DOI 10.1186/s13640-017-0194-1

EURASIP Journal on Image and Video Processing

RESEARCH Open Access

COGNIMUSE: a multimodal video database annotated with saliency, events, semantics and emotion with application to summarization

Athanasia Zlatintsi[1*], Petros Koutras[1], Georgios Evangelopoulos[2], Nikolaos Malandrakis[3], Niki Efthymiou[1], Katerina Pastra[4], Alexandros Potamianos[1] and Petros Maragos[1]

Abstract

Research related to computational modeling for machine-based understanding requires ground truth data for training, content analysis, and evaluation. In this paper, we present a multimodal video database, namely COGNIMUSE, annotated with sensory and semantic saliency, events, cross-media semantics, and emotion. The purpose of this database is manifold; it can be used for training and evaluation of event detection and summarization algorithms, for classification and recognition of audio-visual and cross-media events, as well as for emotion tracking. In order to enable comparisons with other computational models, we propose state-of-the-art algorithms, specifically a unified energy-based audio-visual framework and a method for text saliency computation, for the detection of perceptually salient events from videos. Additionally, a movie summarization system for the automatic production of summaries is presented. Two kinds of evaluation were performed, an objective based on the saliency annotation of the database and an extensive qualitative human evaluation of the automatically produced summaries, where we investigated what composes high-quality movie summaries, where both methods verified the appropriateness of the proposed methods. The annotation of the database and the code for the summarization system can be found at http://cognimuse.cs.ntua.gr/database.

Keywords: Video database, Saliency, Cross-media relations, Emotion annotation, Audio-visual events, Video summarization

1 Introduction

Videos of all kinds (i.e., movies, documentaries, home videos, music videos, etc.) have grown into an easily created and distributed media, and many hours are produced and uploaded to the internet every day. One of the main research challenges that arise, with this increasing amount of video data, is the automatic video understanding that will assist people with effective organization, retrieval, indexing, compression, or even summarization of the video content. People in order to parse, structure, and organize pieces of information use cognitive mechanisms such as attentional selection and information abstraction, which are grounded in conscious or non-conscious activities.

Attention may have two modes, a bottom-up stimulus-driven and a top-down expectation driven. Bottom-up saliency is mainly based on the sensory cues of a stimulus captured by its signal-level properties like spatial, temporal and spectral contrast, complexity, scale, etc. [1–4]. Attention, on the other hand, is a wider concept, including activities such as top-down cognitive information processing, object searching, action taking, and others [5, 6], usually constituting a multimodal process employing visual, audio, and semantic cues. Multimodal saliency as well consists of the fusion (both intra- and cross-modal) of the individual sensory modalities (aural, visual, linguistic, etc.) across time or perceptual scene. During the last

*Correspondence: nzlat@cs.ntua.gr
[1]School of Electr.& Comp. Enginr., National Technical University of Athens, 15773 Athens, Greece
Full list of author information is available at the end of the article

years, many computational frameworks have been proposed for attention and saliency modeling, since they play a significant role in various multimedia applications, such as action recognition [7–9], behavioral analysis, and movie summarization [10–12].

In order to perform tasks such as automatic indexing, retrieval, or classification of audiovisual data, such as videos, it is crucial to have ground truth data for training, content analysis, and evaluation. Further, there is the need of mechanisms that go beyond single-modality content analysis to multimedia and cross-media semantics [13], leading us towards the emerging demand of datasets, where all modalities are annotated with salient segments, actions, objects, or even the semantically important interrelations for message formation. The so far existing datasets include video clips and human annotations of the most significant visual actions, since they only deal with visual interestingness. Moreover, databases intended for audio event or visual action classification consist of independent and trimmed audio files or short video clips up to a few minutes containing only the specific event (rather than longer continuously annotated segments with a variety of events, such as the COGNIMUSE database).

Video summarization, which is the task examined in this work, addresses the problem of providing a short summary of a full-length video, including mainly information required for context understanding; however, without sacrificing the initial informativeness and enjoyability of the original material. Hence, an ideal video summary would contain all salient video segments, regarding both action and comprehensibility, yet being short in length. Video summarization constitutes an actual motivating challenge, mainly because we need prior knowledge of its main topic, that also acts as indicator of importance. Summaries that are well designed can actually improve many aspects of the users' experience, allowing them to glance through all the existing data quickly and thus take immediate decisions on how valuable the content is. In video production and particularly in the production of movies, there are empirical rules that are used so as to enhance the viewing experience or even attract the viewer's attention. In the same way, a summary produced either manually by a human or automatically by a computational system has to consist of those characteristics that will captivate human attention and embody elements, which will assist the development of the plot.

1.1 Contributions and overview

In this work, we propose the COGNIMUSE[1] database and we provide a computational framework for salient event detection and attention-based summarization. The COGNIMUSE database is a multimodal video oriented database, including movies and travel documentaries, annotated with audio-visual and semantic saliency, audio-visual events and actions, cross-media relations as well as emotion (Sec. 3), see also Table 1 for a brief description of the various annotation schemes. The advantages of this database are manifold. It can be used for training of event detection and summarization algorithms, as well as for evaluation of the automatically selected salient events and the produced summaries. The multiple and individually annotated streams/modalities also offer the possibility to evaluate different tasks, i.e., separately the three information streams (i.e., video, audio, and text). The additional annotation schemes, thus the audio-visual and cross-media events, and the emotion annotation can be used for audio-visual event classification, tracking, categorization of salient events regarding the above named schemes, or even for exploitation of the most relevant events in the automatically produced summaries (towards the end of creating user-defined summaries). Additionally, the fact that it contains long videos that are continuously annotated, renders it more relevant and useful, than existing databases for tasks such as recognition, tracking, and crossmodal longitudinal analysis. In order to enable comparisons with other computational models, we also present an extension of our baseline multimodal saliency-based movie summarization system, whose first version was described in TMM13 [12]. This extension includes a unified energy-based computational framework for visual and audio saliency estimation and a method for text saliency computation (Sec. 4). A machine learning approach (Sec. 5.1) is used so as to validate the efficiency of our models, and a new movie summarization algorithm is introduced (Sec. 5.2). Finally, we report on results (Sec. 5.3), both objective using the datasets saliency annotation as ground truth, as well as subjective, i.e., extensive qualitative user experience evaluation scores (for part of the videos). The evaluation verifies the appropriateness of the proposed methods and the quality of the produced summaries that consist of salient and semantically coherent events. Part of the experimental evaluation of this extended computational model was first presented in ICIP15 [14].

2 Background/related work

2.1 Video summarization

Video summarization has been the subject of many recent research works and various algorithms have been proposed in order to tackle the problem [12, 15–20]. Some of the methods used for summarization can relate to user attention or saliency models [12, 21, 22], be domain-specific [23], i.e., be based on a specific topic, such as sports, news, documentary, movies, etc., or relate to the video's story [20], be based on users' preferences, the query context [18], or even focus on dominant concepts [24]. Automatic summaries can be generated either

Table 1 Annotation schemes included in the COGNIMUSE database, providing also a brief description for each layer/category

Annotation Scheme	Annotated Content	Layers/Categories	Annotation Description
Saliency i.e., video elements that captured the viewer's attention instantaneously or in segments	Total: ca. 7 h including: Seven Hollywood movies (ca. 30 min/each) One full-length movie (ca. 100 min) Five travel series (ca. 25 min/each)	Audio	Acoustically interesting segments i.e., abrupt/loud sounds etc.
		Visual	Visually interesting segments i.e., motion, color variations etc.
		Audio-visual	Audio-visually interesting segments i.e., an explosion (that includes both visual and acoustic saliency)
		Semantics	Conceptually important as stand-alone semantic events i.e., names, plot elements, facial expressions etc.
		Informative Segments	Segments important for understanding the plot of the specific video clip, considered also as a manually generated summary.
		Expert Summaries	Summaries created by an "expert" related professionally with film production.
Audio Events	Total: ca. 5 hours including: Seven Hollywood movies One full-length movie	Human Nature Mechanical Music	Events regarding various human, nature, or mechanical sounds and music, i.e., voice, movement, animal sounds etc. For more info see Table 4.
Visual Actions		Facial actions Body movements Gestures	General facial actions or body movements incl. object manipulation or interaction, i.e., talk, smile, sitting down/up. For more info see Table 5.
Cross-media semantics	Total: ca. 100 min including: One full-length movie	Equivalence Complementarity Independence	Interaction relations between different modalities, i.e., images, language, body movements or acoustic events.
Emotion	Total: ca. 3.5 hours including: Seven Hollywood movies	Arousal Valence	Corresponding to viewer's excitement and describes emotional evaluation from negative to positive.

by using key frames, which correspond to the most important video frames, representing a static storyboard [16, 17, 19], or with video skims combining the most descriptive and informative video segments [12, 20]. For more general reviews about video summarization, we refer the reader to [11, 12, 21, 22, 25].

2.2 Evaluation methods

A common approach for the evaluation of summarization algorithms—closely related to quality of experience (QoE) methodology–is to perform qualitative user-based studies in order to compare various summaries of the same video [17, 19, 20, 26]. One way to accomplish this is the use of metrics such as *informativeness* and *enjoyability* [12, 22, 26], where the users review the summaries compared to the original clip and assign a score. The metric of *concept coverage* evaluates the number of relevant objects or actions included in a summary [17, 19]. Other methods for summary evaluation include automatic comparison to some reference summaries (a method inspired from text summarization literature [27]), where the comparison of the produced summary is performed towards a user-generated summary of the video [16, 19, 28]. Finally, data exploration tasks have also been used as a quality metric for evaluation [27].

2.3 Datasets for quantitative evaluation

Such *qualitative* user studies are imperative, since human perspective is necessary for implementing systems that take into account user preferences and thus produce "user-defined" summaries. However, it is also crucial to find metrics for automatic evaluation of summarization algorithms. For this reason, the existence of databases annotated with the most salient segments, actions, or objects, so as to avoid extensive and time-consuming user studies, has come to be a demand.

The TVSum50 dataset [29], developed at Yahoo Labs, contains 50 videos of various genres (news, how-to's, documentaries, and user-generated videos) and their shot-level content importance scores, including 20 annotations per video. For video collection, ten categories from the TRECVid Multimedia Event Detection (MED) task [30] were selected including five YouTube videos per category. Another related dataset is "MED Summaries" [23], consisting of videos from the MED 2011 challenge. It was developed for evaluation of dynamic video summaries, tailored to category-specific summarization, and it includes annotations of important and semantically consistent segments. In [31], a summarization dataset was introduced, containing ca. 600 videos (1430 min in total), from six different domains: "skating," "gymnastics," "dog," "parkour," "surfing," and "skiing," with ca. 60% of the videos having highlight annotations. The SumMe dataset for summarization of user-generated videos was introduced in [32], containing 25 videos, ranging from 1–6 min, with 15–18 ground truth annotations, i.e., summaries that contain the most important content. Finally, in [17], a dataset of egocentric videos, filmed with head-mounted cameras,

was produced, so as to imitate what is actually seen by the user. Four test videos are publicly available, with a duration of 17 h, captured by four people while performing daily activities, including also ca. 1660 spatial important object segmentations as well as a set of negative frames, thus not important for the summary.

2.4 Audio and visual events

During recent years, the automatic classification of environmental sounds has gained much attention with application to content-based multimedia indexing and retrieval [33–35]. Recognition of different sounds in a soundscape is important, since it can assist in making comparisons and separate or isolate various sounds [36]. Audio categorization has been extensively studied in the context of perceptual soundscape research in [37]. Schafer [37] was actually one of the first to introduce a taxonomy for different environmental sounds, proposing six different categories, taking into account the sounds' importance due to their individuality, their numerousness, or their domination regarding the soundscape. Brown et al. [38] proposed a rather detailed taxonomy of the acoustic environment for soundscape studies showing categories of places and sound sources (see also [39]). Salomon et al. [40], based on the work of [38], proposed an extensive taxonomy for urban sounds. For additional and more extensive review of previous works, we refer the reader to [41].

The field of video understanding, on the other hand, relates to research on object recognition and scene understanding, and many large databases with static images have been introduced for this reason [42, 43]. Action recognition is one of the most challenging fields in computer vision, and video data is needed in order to evaluate the various methods that have been proposed during the last decades [44–52]. Thus, the creation of large datasets with realistic video data is crucial for the development of efficient action recognition algorithms, and nowadays, many databases can be found including a variety of action categories [53–56].

2.5 Cross-media semantics

For the creation of automatic audiovisual presentations or summaries, single-modality content analysis is not enough; we need to go beyond to multimedia and cross-media semantics for better content selection [13], in order to preserve comprehensibility and cohesion in communication [57]. In [58], evidence has been provided that while watching TV series, brain activity is affected by both auditory and visual modalities. Semantics on the other hand are also important to understand what is presented. In [59], the effect of narration has been investigated (in form of auditory stimuli or captions) on eye-movement behavior. Although few would argue against the notion that when viewing multimedia data the users' main task is to follow the line of events (i.e., understand the plot), no systematic exploration has been made so far regarding the semantic information on viewing behavior and experience. COSMOROE [60], a corpus-based framework used in this work as an additional annotation scheme, provides such a mechanism for the description of how multimedia information interact to convey a meaning. Specifically, it defines a set of relations aimed to capture the crossmodal/semantic interrelations between images, language, and body movements, so as to capture the semantic aspects of message formation as well as to address questions related to how humans combine pieces of information.

2.6 Emotion datasets

Affective video content analysis aims to automatic emotion recognition with applications in mood-based personalized content delivery, video indexing, and summarization [61, 62], and ground truth data is needed both for training and benchmarking. The HUMAINE database [63], consisting of 50 clips from 1.5–3 min, is annotated with a wide range of labels, i.e., emotion-related states (intensity, arousal, valence, etc.), context labels, emotion words, and others, in a framewise manner; however, it is intended mainly for illustration of key principles of affective computing rather than applying it to machine learning. FilmStim [64] includes 70 film excerpts from 1 to 7 min, intended to elicit emotional states in experimental psychology experiments. Even though it is one of the largest databases, it uses unique global labels for emotional ranking (i.e., anger, sadness, fear, etc.), which is not sufficient to build ground truth data. DEAP [65] is an other database with ratings on arousal, valence, and dominance, containing 120 1-min music videos. MAHNOB-HCI [66], a multimodal database composed of 20 short (35–117 s) emotional excerpts derived from commercially produced movies and video websites, is also annotated with respect to arousal, valence, and dominance. Finally, the most recent LIRIS-ACCEDE database [67] contains 9,800 (ca. 26 h in total) video excerpts (8–12 s long), of diverse genres, with rankings in 2D valence-arousal space.

3 COGNIMUSE database

The *COGNIMUSE database* is a video-oriented database multimodally annotated with sensory and semantic saliency, audio and visual events, cross-media relations as well as emotion. Our aim is to introduce a *framework* that will assist in training and evaluation of event detection and summarization algorithms, regarding their accuracy in detecting salient events as well as for content analysis, with respect to the included annotation schemes.

3.1 Data collection and content

The process for the creation of the COGNIMUSE database included data collection, data conversion, and annotation in different phases.

Specifically, the dataset consists of half-hour continuous segments (with the final shot/scene included) from seven Hollywood movies (three and a half hours in total), which are "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)[2]. The specific Oscar-winning movies were selected to form a systematic database of acclaimed, high production quality videos from various genres (i.e., drama, musical, action, epic, animation). Further, this selection was based partly on the popularity and partly on the plot structure of the movies, which are made exclusively for the establishment of the emotional disposition themes of the characters. They include basic concepts, such as the main character/s, the desire, and the conflict as well as 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, leading therefore to effective summaries. The seven movie segments were annotated with sensory and semantic saliency, audio-visual events and emotion.

Five travel documentaries (ca. 20 min long each), including four episodes from "Alternate Routes" series, namely "London" (LON), "Tokyo" (TOK), "Sydney" (SYD), "Rio" (RIO), and one episode from "Get Outta Town" series, i.e., "London" (GLN), were also annotated with sensory and semantic saliency. These specific TV travel series usually involve one or more presenters visiting different places, being in contact with the locals, interviewing people, explaining the habits, traditions, and their way of life. They were selected due to the richness of the interacting modalities available in this genre, including a variety of language modalities (speech, text in the form of subtitles or in graphics etc.), image modalities (dynamic images) gestures, and other body movements.

Finally, a full-length movie, namely "Gone with the Wind" (GWW) (the first part with total duration 104 min) was selected for saliency, COSMOROE-based [60], and audio-visual event annotation. The annotation of cross-media relations in a full-length movie is essential, in order to study the cross-media semantic interplay at a full-scale level. This actually allows 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. Moreover, the selection of this movie is not only useful but also a necessity for the purpose of evaluating the developed movie summarization algorithm and the automatically produced summaries on a full-scale and semantically complete movie.

The Hollywood movies were taken from the official DVD releases, and for reference purposes, the exact time sequences were noted, while the rest of the data were downloaded from the web under a creative commons license. The movie segments were ripped and saved in .avi format in high resolution for summary visualization and rendering, and small resolution for processing and annotation. The full database also includes movie subtitles or transcripts, which are used for text processing and text saliency estimation.

3.2 Saliency annotation

All database videos have been annotated with respect to (mono- and multimodal) sensory and semantic saliency (in a binary mode), including scene and shot segmentation, by three annotators in separate runs for each individual saliency layer, starting with audio annotation, followed by visual, audiovisual, semantics, and the annotation of the informative segments. Figure 1 shows the annotation interface as well as key frames that were annotated as salient or not in the various modalities/layers. For annotation purposes of all the saliency layer, Anvil[3] a free video annotation tool has been used, which offers frame accurate, multi-layered annotation driven by user-defined annotation schemes [68]. Although this kind of annotation is considered highly subjective—user preferences on what is important cannot be dictated—the three trained annotators could consult an instruction's manual, created for this task, in order to achieve as high as possible degree of annotation uniformity. The various types of annotation were performed across multiple days, in order to avoid fatigue.

The movie clips were first manually segmented into shots (i.e., the interval between editing transitions, such as

Fig. 1 Interface for saliency annotation, showing the different annotation layers and the respective key frames (*red lines* show the time instances of the specific frames) that were labeled as either salient or not in each layer

cut or fade) and scenes (defined as a complete, continuous chain of actions (shots) that occur at the same place and time). The average shot and scene duration for the movies were 3.5 s and 2.3 min, while for the travel documentaries, the respective duration was 3 s and ca. 40 s, respectively. Next, sensory and semantic saliency content annotation was performed, i.e., segments that captured the viewer's attention, with respect to the following layers.

Sensory information: This is a pre-attentive layer of saliency, where the annotation has been performed quickly, effortlessly, without any focused attention, and with little or no searching required. The annotation was based on video elements that captured the viewers' attention instantaneously or in segments including monomodal, i.e., audio (A) and visual (V) saliency annotation, and multimodal audiovisual (AV) saliency annotation of the sensory content, hence, segments that are acoustically, visually, or audiovisually interesting, on separate annotation runs of each individual layer. The *audio-only saliency* (A) was annotated by only listening to the audio stream of the movie segment and the annotators were instructed not to take into account any semantic information such as the type of the sound (e.g., speech, music) or its meaning (e.g., dialogue, genre of music). The *visual-only saliency* (V) was annotated by only watching the movie segment, again without taking into consideration semantic information. For the *audiovisual* (AV) annotation, the annotators were instructed to handle the two streams (A,V) as one multimodal cue. Monomodal and multimodal salient events included features such as loudness, pitch variations, and sound effects (aural cues); contrast, intensity, motion, color (visual cues), and combined audiovisual events, artificial or not.

Semantic information: This layer includes segments that are conceptually important as stand-alone semantic events (S) that have a beginning, a steady state, and an ending, (e.g., important names, plot elements, phrases, actions, symbolic information, sounds, gestures, facial expressions indicating a feeling, etc.). The specific events are not considered important just for the examined movie but generally, as an objective, direct, or indirect meaning. For the purposes of our objective evaluations, this layer 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).

Informative segments: This layer consists of segments important for understanding the narration plot of the specific half-hour movie clip. They could be a subset of the semantically salient information, considered also as a manually generated skim consisting of descriptive but not necessarily the most enjoyable action segments.

Expert summaries: For the seven Hollywood movie clips, summaries (ca. 5 min 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 30-min segment, according to his preferences, which could vary between 1 and 10 min. Since the creation of a summary and not a movie trailer was requested, 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.

Statistical analysis: Tables 2 and 3 show the percentage of the annotated salient frames (labeled by at least two annotators), the average (pairwise) correlation agreement between the annotators—overall satisfactory, considering the subjectivity of the task—, and Chronbach α, measuring the internal consistency ("reliability") between the three annotators for all videos and annotation layers. The total annotation time for each individual layer was ca. 4 h for the audio (A) layer, 3.5 h for the visual (V), and the audio-visual (AV) saliency layers and ca. 6 h for the audio-visual-semantics annotation. Regarding the inter-annotator agreements for both metrics, note that it is higher for the sensory layers (with best agreement for A and AV) compared to the sensory-semantic (AVS) layer. Regarding Chronbach α, we notice that for the sensory A and AV layers, the agreement is acceptable/good for all videos (> 0.7) with the exception of the travel documentary "RIO." In order to overcome the lower agreement observed for certain videos, the final saliency ground truth was formed on the basis of consistently labeled salient frames only (considered as salient by two or all three annotators).

In order to check the validity of the annotations and thus see whether the reliability would increase if more annotators had been employed for the annotation task, we selected one movie, namely "DEP," which was annotated by five more users, thus eight in total. Figure 2 shows how Chronbach α changes while the number of annotators increases. As expected, α increases as the intercorrelations among the "test items" increase. For all layers, we observe that for up to five users, the increase is nominal (up to 0.07 for V) and many more users are needed in order to accomplish better reliability. However, the change while adding more users is rather small comparing to the increase observed from two to three annotators. Due to this observation, the constraints such as the fact that the annotation performed is time consuming (since all layers are annotated separately) and the subjectivity of the task,

Table 2 Statistics for COGNIMUSE database (Hollywood movies and GWW) annotated with salient events

Layer	BMI	CHI	CRA	DEP	GLA	LOR	FNE	GWW
Percentage (%) of salient frames								
A	25.4	56.3	55.0	33.4	60.9	58.3	54.6	69.2
V	30.1	46.3	37.9	32.4	39.2	43.3	36.9	71.5
AV	27.4	47.7	43.1	37.8	49.6	50.7	39.7	70.1
AVS	63.2	76.6	64.8	71.8	68.5	72.7	67.6	88.0
Average (pairwise) correlation between annotators								
A	0.54	0.48	0.46	0.49	0.51	0.52	0.42	0.55
V	0.31	0.33	0.32	0.45	0.38	0.43	0.38	0.36
AV	0.45	0.45	0.41	0.54	0.44	0.50	0.44	0.40
AVS	0.29	0.24	0.27	0.29	0.31	0.33	0.23	0.22
Chronbach α for three annotators								
A	0.78	0.73	0.72	0.74	0.76	0.76	0.69	0.79
V	0.58	0.60	0.58	0.71	0.65	0.69	0.65	0.62
AV	0.71	0.71	0.68	0.78	0.70	0.75	0.70	0.67
AVS	0.55	0.48	0.52	0.55	0.57	0.59	0.47	0.45

we can claim that the correlation between the annotators is acceptable.

Additional material: For all video data included in the COGNIMUSE database, the additional material includes (i) the number of total sentences included in each video with their start and end times, (ii) the number of words per sentence, and (iii) part-of-speech (POS) tag for each word with start and end times (or start and end frame index). The information provided for the words and sentences respectively has been obtained automatically by the analysis of the time-aligned transcripts, by performing forced segmentation on the audio stream using the text transcript and phone-based acoustic models. More specifically, the Sonic ASR toolkit [69] was used and general-purpose acoustic models, i.e., content-dependent tri-phone hidden Markov models trained on clean speech. Next, a shallow syntactic parser that performed POS tagging was used, specifically, a decision-tree-based probabilistic tagger [70]. For more information, we refer the reader to TMM13 [12].

Table 3 Statistics for COGNIMUSE database (travel documentaries) annotated with salient events

Layer	LON	RIO	TOK	SYD	GLN
Percentage (%) of salient frames					
A	58.7	43.8	60.0	55.1	45.6
V	46.5	48.5	46.6	48.8	40.5
AV	53.9	50.3	54.7	53.7	42.5
AVS	72.7	79.4	80.3	80.4	72.5
Average (pairwise) correlation between annotators					
A	0.62	0.25	0.57	0.68	0.52
V	0.50	0.31	0.33	0.51	0.40
AV	0.61	0.33	0.56	0.65	0.43
AVS	0.25	0.08	0.21	0.23	0.28
Chronbach α for three annotators					
A	0.83	0.50	0.86	0.80	0.77
V	0.75	0.57	0.76	0.60	0.67
AV	0.83	0.59	0.85	0.79	0.70
AVS	0.51	0.21	0.48	0.45	0.54

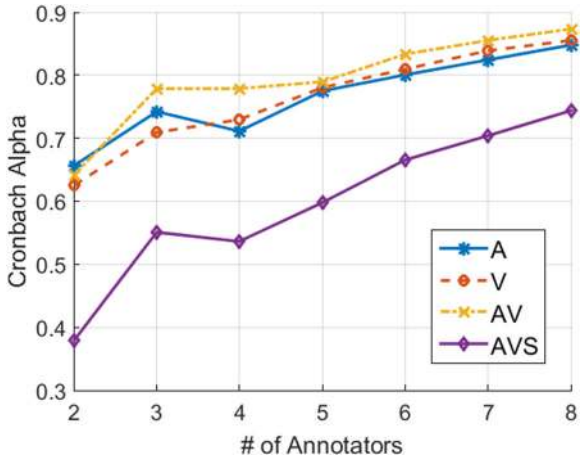

Fig. 2 Inter-annotator agreement showing Chronbach α for eight users on the movie "DEP"

3.3 Audio events and visual actions annotation

All movies included in the database have been annotated with audio events and visual actions. In total five annotators watched, the eight movies with both audio and visual streams simultaneously playing and marked the temporal boundaries of each predefined event. Since various of the events could have some temporal overlap, the annotation scheme included multiple layers for the same category, so as to enable the annotation of the same event more than once in the same time sequence. Such cases occur for instance when a dialogue evolves, in other words whenever two or more actors are present in a scene. Certain events, like talking, laughing, punching, etc. are multimodal, including both an acoustic event and a visual action (if visible) and thus they are annotated in both modalities. Such events are considered to be a projection to each individual modality.

One of the main contributions of this annotation scheme, compared to other audio [38, 40] or visual action

[54, 55] datasets, is that the employed movies are continuously annotated; thus, useful not only for classification but also for event recognition as well.

Audio events: For the audio event annotation, we consider the sound taxonomies proposed in [40]. The specific dataset includes 1302 recordings (27 h in total thereof 18.5 h are annotated) containing sounds from ten classes. For the needs of this work, four main categories were selected, i.e, (i) human sounds, (ii) nature sounds, (iii) mechanical sounds, and (iv) music, which are divided in two additional subcategories, as can be seen in Table 4. Subcategory two, in almost all categories, was enhanced with acoustic events that were assumed to be encountered often in the COGNIMUSE database, e.g., breathing sound in voice and horse galloping/neighing in the animal sounds. Moreover, in the music category, we included the annotation of the genre and the instruments (which were labeled according to the annotators' knowledge and judgment). Finally, some of the categories were further subdivided and made more distinct; as for instance, speech sounds were explicitly annotated as being male, female, child, or synthetic. The audio events that were added in the second subcategory (and thus are not included in the taxonomy of [40]) can be seen in italics in Table 4. Note that in square brackets, you can see the subcategories that no instances were found during the annotation.

Visual actions: For the visual action annotation, we selected categories specifically related to human actions. For this reason, the animation movie "Finding Nemo" was excluded, since the lead characters are non humans and the annotation scheme could not be applied. The selected categories were derived from popular and challenging databases for action recognition, i.e., Hollywood2 [54], HMDB [55], and the more recent work for action recognition [71]. Hollywood2 consists of 1707 clips from 12 categories selected from different movies with average length of 11.6 s. HMDB is one of the largest databases as it contains 6766 clips from 51 categories, where the clips were selected from movies and internet videos, having an average length of 3.15 s. In [71], a dataset annotated with sequences of actions is proposed. Clips from 69 movies were extracted based on Hollywood2 actions and were manually annotated with 16 classes (12 of which presented in Hollywood2). Recently, larger databases for action recognition have been proposed, i.e., UCF101 [56] with 101 categories and Sports1M [50] with 487 classes and one million clips; however, their categories are mainly related to sport activities and therefore not relevant for our task.

The categories used in our annotation scheme are a union of the actions employed in the three above mentioned datasets, i.e., (i) general facial actions, (ii) facial actions with object manipulation, (iii) general body movements, (iv) body movements with object interaction, and

Table 4 Categories for audio event annotation

Categories for Audio Event annotation		
Categories	Subcategory 1 (no. of layers)	Subcategory 2
Human	Voice (×3)	speech *male*, speech female, speech *child*, [speech synthetic], *crowd noise*, laughter, shouting, crying, coughing, [sneezing], *breathing*, *spitting*, *singing*, infant, other
	Movement (×3)	footsteps, *punching*, other
Nature	Elements (*x*2)	wind, water, waves, thunder, *fire*, *sand*, other
	Animals (×2)	dog bark, [dog howl], bird tweet, *bird sing*, *horse galloping*, *horse neighing*, [sheep], other
	Plants/Vegetation (×2)	[leaves rustling,] [other]
Mechanical	Construction (×2)	[jackhammer], hammering, drilling, [sawing], engine running, other
	Ventilation (×2)	[air-conditioner], other
	Non-motorized Transport (×2)	bicycle, skateboard, other
	Social Signals (×2)	bells, clock chimes, alarm/siren, [fireworks], gun shot, *explosion*, *glass breaking*, *door rusty*, *door opening/closing, swords*, other
	Motorized Transport (×2)	[marine], rail, road, [air], [other]
Music	Amplified (×1)	live, recorded
	Non-amplified (×1)	live
	Sound Source (×1)	*Diegetic: originated from the source within the film's world, Non-diegetic: mood music Background music: when music is not the basic element in the scene Foreground music: when music is basically the only thing you hear*
	Genre (×1)	*classical, symphonic, rock, pop*, [*punk*], *jazz, folk/country, blues*, [*metal*], *rock 'n roll, hiphop*, [*reggae*], *electronic, funk/soul/rnb, ethnic/world, other*
	Instrument (×1)	[*keyboard*], *string, wind, percussion, orchestra, electronic/amplified, mixed (e.g., rock band etc.),* other

(v) body movements for human interaction. Those were extended with a sixth category, specifically (vi) gestures that we assumed to be encountered in a regular basis. Gestures can be considered as a special case of general actions and are really important for the comprehension of the semantics. The list of the employed visual actions as well as their subcategories are presented in Table 5, where in italics the additionally included actions can be seen and in square brackets the actions that no instances were found.

Table 6 shows the total number of instances per annotated audio and visual category, plus the total duration in minutes (the overlapping instances are counted as well). We note that almost 19 h with 6262 instances of audio events are annotated in the various categories. From those, more than 4 h are annotated as voice, almost 4 h as music, ca. 16 min as social signals, ca. 20 min as animal sounds, etc. Additionally, ca. 4.5 h with 4847 instances are annotated as visual actions, whereof more than 2 h are annotated as general facial actions and more than 1 h as general body movements.

Table 7 shows subcategories with duration that exceeds 20 min in total. We notice that "male" and "female voice" has the longest duration and it is reasonable that the visually annotated "talk" action takes up almost the same duration. Regarding music genres, "symphonic" was found the most, while the action for "walk" was annotated 456 times with a duration of ca. 42 min. There were also events that had numerous instances; however, their total duration was quite small. Those events can be seen in Table 8.

3.4 Cross-media semantics annotation

Communication among people is primarily multimodal. People usually use different modalities, as for instance speech and body movements, in order to interact with each other. The information that is communicated by these modalities is fused and form coherent messages.

The cross-media semantics annotation of "Gone With the Wind" aims at characterizing the multimodal messages that are presented in the movie and thus contribute to the analysis of the semantic interrelations between the various modalities that are examined, i.e., images, language, body movements, and acoustic events. The annotation scheme was based on the COSMOROE framework [60]. Next, we describe the annotation process in GWW and we present the most relevant results and statistics. For the annotation, the ELAN[4] tool was used [72].

In the COSMOROE framework, three major types of interaction relations can be found: *Equivalence, Complementarity*, and *Independence*. Each type with its subtypes are described next, using examples from GWW for better understanding. A segment annotated with a COSMOROE relation consists of a label, indicating the type of the relation, and of various visual, audio and text elements (with start and end times). Specifically, 470 relations were annotated including the following elements: (a) utterance text (i.e., spoken language transcription), (b) graphic or scene text (text information that occurs on the screen), (c) frame sequences, thus part of shots that participate in a relation, (d) key frame regions, depicting a particular object of interest in a sequence of frames, (e) body movements and gestures (i.e., head movements, deictic gestures), and (f) acoustic events. The duration of the annotated segments range from 0.3–135 s.

Equivalence: Different modalities or media can express semantically equivalent information. Four sub-relations can be found clustered into two groups: *Literal Equivalence* and *Figurative Equivalence*. In *Literal Equivalence* we distinguish the relations of *Token-Token*, which appear when different modalities refer exactly to the same entity, uniquely identified as such, and *Type-Token*, including cases where one modality refers to a class of entities and the other refers to one or more members of the class.

Table 5 Categories for visual action/event annotation

Categories for visual action/event annotation	
Categories (no. of layers)	Subcategory 1
General facial actions (×2)	Smile, cry, laugh, chew, talk, other
Facial actions with object manipulation (×2)	Smoke, eat, drink, other
General body movements (×2)	Sitting down, sitting up, standing up, running, [cartwheel], clap hands, climb, climb stairs, [dive], fall on the floor, [backhand flip], [handstand], jump, pull up, push up, [somersault], turn, dance, walk, other
Gestures (×2)	*Wave hands, point at something, pantomime*, other
Body movements with object interaction (×2)	Answering phone, driving car, getting out of the car, open car door, open door, brush hair, catch, draw sword, [dribble], [golf], hit something, kick ball, pick , pour, push something, ride bike, ride horse, [shoot ball], shoot bow, shoot gun, [swing baseball bat], sword exercise, throw, other
Body movements for human interaction (×2)	Fighting, hugging, kissing, grab hand, threaten person, [fencing], kick someone, punch, shake hands, sword fight, other

Table 6 Statistics for COGNIMUSE database (Hollywood movies and GWW) annotated with audio-visual events per event category

Audio events		
No. of instances: 6262, total duration in hours: 19.24		
Category/Subcategory 1	Instances	Dur. (min)
Voice	3809	245.75
Movement	228	19.82
Elements	154	16.91
Animals	222	20.26
Plants	0	0.00
Construction	46	5.19
Ventilation	4	0.54
Non-motorized trans.	18	0.84
Social signals	444	15.66
Motorized trans.	48	3.86
Non-amp. music	12	5.16
Amplified	218	213.28
Sound Source	640	226.91
Genre	231	222.86
Instrument	200	162.80
Visual actions		
No. of instances: 4847, total duration in hours: 4.58		
Category	Instances	Dur. (min)
General facial actions	2233	129.67
Facial action with obj. manip.	90	4.08
General body mov.	1215	79.75
Gestures	284	9.09
Body mov. with object inter.	693	33.72
Body mov. for human inter.	332	18.79

For example, a Token-Token relation is defined when an acoustic event (e.g., the sound of a bell) is combined with the visual representation of the entity producing the sound (i.e., the image of somebody that tolls the bell) as in Fig. 3a. A Type-Token relation would be annotated when someone utters the word *"baby"* and a baby is presented in a sequence of frames. In *Figurative Equivalence* two types of relations are detected: *Metonymy* and *Metaphor*, meaning that the recipient of the multimodal message considers two entities as semantically equal despite that each media presents a different entity. The most common metonymic pattern is *"part for whole"*; the image presents a part of an entity while the language refers to the whole (e.g., the word *"land"* and an image showing only a part of the land, see Fig. 3b). Finally, Metaphor relations, which actually occur rarely, would be defined when a media draws a similarity between two referents belonging to different domains.

Table 7 Statistics for COGNIMUSE database (Hollywood movies and GWW) annotated with audio-visual events per event category. Subcategories that their annotated instances exceeded a duration of 20 min in total can be seen

Most frequent audio and visual events		
Category/subcategory	Instances	Dur. (min)
Voice: speech male	1874	102.39
Voice: speech female	1048	55.55
Voice: crowd noise	188	42.68
Sound source: background music	350	158.20
Sound source: foreground music	290	68.71
Genre: symphonic	119	118.61
Genre: other genre	70	42.14
Instrument: string	32	23.29
Instrument: percussion	102	91.86
Instrument: mixed	16	22.71
General facial actions: talk	1915	114.67
General body mov.: walk	456	41.72

Complementarity: The information expressed through the different modalities (i.e., audio, video, or text) complement each other. Complementarity relations are divided into two sub-relations, those in which the combination of the information expressed by different media is *essential* for the comprehension of the multimedia message and those where the information is *non-essential*. Both essential and non-essential complementarity include the relations of *Exophora*, *Agent-Object*, and *Apposition*. *Adjunct* relations are classified as non-essential.

Specifically, *Exophora* includes cases of "anaphora," where one modality resolves the reference made by another. For example, the phrase *"the darling thing"* does not express a specific object; however, this information

Table 8 Statistics for COGNIMUSE Database (Hollywood movies and GWW) annotated with Audio-Visual events per event category. Subcategories with numerous instances but with small duration

Frequent audio and visual events with small duration		
Category/subcategory	Instances	Dur. (min)
Voice: shouting	204	8
Voice: crying	101	7
Voice: breathing	102	18.55
Movement: footsteps	165	11.51
Social signals: door opening closing	114	2.18
General facial actions: smile	115	3.72
General body mov.: running	109	5.56
General body mov: turn	216	4.79
Gestures: wave hands	116	3.97

Fig. 3 Examples of COSMOROE cross-media relations in "Gone with the Wind." **a** Token-Token. **b** Metonymy, part for whole. **c** Complementarity, Exophora Essential **d** Complementarity, Agent-Object Essential

is provided by the image as in Fig. 3c, where a hat is depicted. *Agent-Object* relations are related to cases where an intentionally omitted subject or object is revealed in another modality, e.g., in the phrase *"Scarlet, you look,"* the omitted object is a piece of paper that Scarlet is urged to look at, see Fig. 3d. In *Defining Apposition* relations, the extra information provided by one medium identifies or describes someone or something, i.e., a woman that appears on the screen is called "hostess." However, we would not be able to identify her as such, unless the audio-textual information had revealed it. Cases of *Non-Defining Apposition* relations are presented when one modality reveals a generic property or characteristic of the very concrete entity mentioned by another. *Adjunct* is a non-essential relation that denotes an adverbial-type modification. For instance, a woman is shouting *"Get out of there!"* while a boy is running out of a house; in this case, the action of running (thus the body movement itself) reveals how the boy is getting out.

Independence: Each modality carries an independent message and their combination creates a coherent multimedia message. In this relation, three subtypes can be found: *Contradiction*, when the different modalities are semantically opposite or incompatible; *Symbiosis*, when different modalities expresses different information (i.e., spoken information that is not depicted on the visual modality); and *Meta-Information*, a relation that could not be found in movies.

Table 9 shows the percentage (%) of the relations that have been annotated in the movie. Contradiction is not presented, since only one instance was annotated. Moreover, the symbiosis relation, which actually does not provide any useful information, was omitted from being annotated, since it includes all segments that do not belong in another relation.

The elements that are annotated and participate in a COSMOROE relation are presented in Table 10 (Utterance-Text, Graphic and Scene Text are grouped as Utterances while Objects include Key Frame Regions and Frame Sequences). Two values can be seen; the first one concerns the total number of annotations, while the second one equals to the number of unique labels, when multiples of the same label are excluded. From the analysis of the COSMOROE annotations, we notice the following patterns. Firstly, in Token-Token relations, the majority of annotations concerns the combination of acoustic events with body movements, either of a human or the animal that produces the specific sound. Additionally, one in five of these cases include combination of labels: name of a person and his figure. Usually a noun, or a verb (less frequently), and its visual representation is classified as a Type-Token relation. The most frequent metonymic patterns are "part for whole," followed by "action for goal" and "action for cause," refer to [60] for their definitions. Pronouns and local adverbs are the majority of utterances that are contained in Exophora relations, while deictic gestures participate in 27% of the total cases of Exophora.

Correlation between the four saliency layers (i.e., audio, visual, audiovisual, and semantics) and the most frequent COSMOROE relations can be seen in Fig. 4; intending to show the salient crossmodal relations in each individual modality. We observe that audiovisual salient segments correlate mostly with Token-Token, Metaphor,

Table 9 Percentage (%) of the COSMOROE cross-media relations in GWW

COSMOROE relations	Subtypes	Percentage of no annotations	
Equivalence	Token-Token	28.9	73
	Type-Token	23.2	
	Metonymy	19.6	
	Metaphor	1.3	
Complementarity	Exophora	13.2	27
	Agent-Object	3.8	
	Appostition	7.9	
	Adjunct	2.1	
Independence	Symbiosis	37.5% of total duration	

Table 10 Total number of the elements participating in COSMOROE relations in GWW, with and without multiple labels

Elements	No. of annotations	No. of annotations (duplicates excluded)
Utterances	347	223
Objects	728	140
Body movements	153	73
Acoustic events	125	49
Gestures	34	13

and Metonymy relations, audio salient segments correlate mostly with Token-Token, Metonymy, and Type-Token, visual salient segments correlate with Metaphor and Type-Token, while semantic salient segments with Token-Token and Metaphor. Interestingly, we note that more than 40% and up to 70% (for AV) of Adjunct relations have also been annotated as salient, regardless modality; however, we have to mention that the total number of instances was ten; thus, we cannot be certain for the validity of this result. The presented statistical analysis is indicative, and more annotated movies (and movie genres) would be needed in order to draw more insightful conclusions.

3.5 Emotion annotation

We annotated the emotional content of the seven Hollywood movies, primarily to conduct experiments on recognition and tracking [73]. The emotion representation selected was that of the two factor dimensional model of arousal-valence, where the arousal value corresponds to the viewer's excitement, whereas valence describes the emotion evaluation, from very negative to very positive. The three dimensional variation of this representation, with the inclusion of dominance (the sense of control over one's situation) is also popular; however, the third dimension introduces a lot of complexity to the annotation process for very little added information on the emotional state of viewers [74]. The two-dimensional model used, along with some sample lexical emotion labels, is shown in Fig. 5.

For the purposes of this annotation, we distinguish three types of target emotions: *intended*, *expected* and *experienced*. The intended emotion annotations are meant to capture the emotional response that the movie tries to evoke in the viewer, without taking into account whether it is actually successful. Experienced emotion annotations represent the actual emotional experience of an individual while watching a movie. Finally, expected emotion is the normative emotional response, the expected emotional experience of a random viewer. This distinction is typically overlooked [75]; however, we feel it is necessary simply by virtue of how different these emotional experiences may be; a "bad" movie may have very different intended and expected outcomes, while a viewer's taste may lead to very different expected and experienced outcomes. While intended and expected outcomes are generally more useful, the experienced emotion annotations could be the basis of user-adapted predictions of emotion [76].

Annotating procedure: Two types of annotation were performed, with experienced emotion annotated by student volunteers and intended emotions annotated by experts; expected emotion was derived from the experienced emotion annotations. The annotations were performed using the FEELTRACE [77] emotion annotation tool. The annotators were presented with a two window interface, as shown in Fig. 6, where they could watch the movie in one window while registering their emotion annotations on a second window corresponding to the valence-arousal space, via moving the mouse cursor to corresponding position.

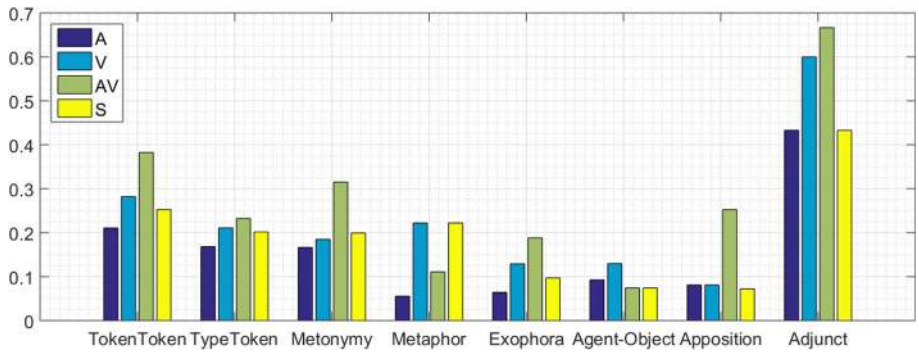

Fig. 4 Correlation between annotation of saliency and COSMOROE annotation. From *left* to *right*: Token-Token, Metonymy part for whole, Compl. Exophora Ess., and Compl. Agent-object Ess

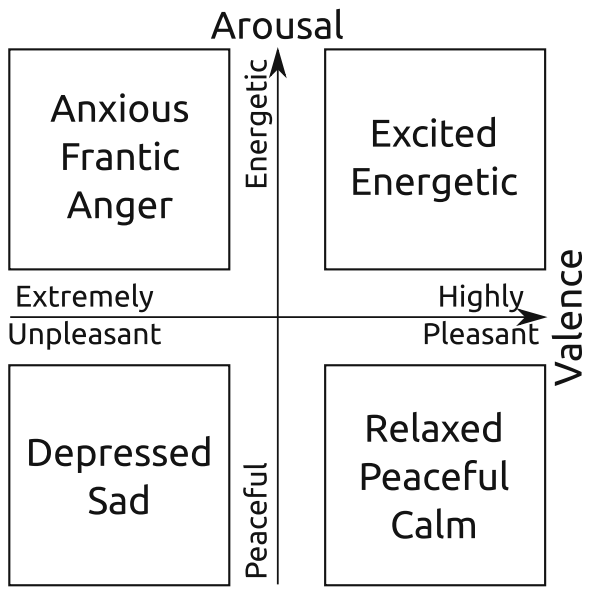

Fig. 5 The two-dimensional space of arousal-valence

Seven volunteers, 20–30 years old, two female and five male performed the annotation of experienced emotion. Each annotator was given a short training on the interpretation of the emotional space and the tool, before performing a sample annotation of a short clip from a different movie. After the training process was finished, they moved on to annotating the seven clips. Each subject annotated no more than two clips within the same day, accompanied by a questionnaire, to be filled after the annotation was finished, aiming to assess the subject's prior knowledge of the movie, as well as their overall evaluation of the movie. All annotators evaluated all clips, with five (out of seven) performing the entire process twice to allow the validation of intra-annotator agreement. Expected emotion was derived from the individual experienced emotion annotations using a correlation-based scheme similar to that in [78] with particularly uncorrelated annotations being rejected as outliers. Intended emotion annotations were performed three times by a single expert, expected to be very consistent.

Annotation results: Each annotation results in a pair of time series, one for arousal and one for valence, pseudo-continuous in time and values (see also Fig. 10). The values of each curve are in the range $[-1, 1]$. The sampling rate of FEELTRACE is irregular and very high, exceeding 1 kHz, resulting in very large outcome annotations, which were downsampled to match the video frame rate of 25 fps.

Figure 7 shows two-dimensional histograms of annotations for intended and expected emotion. The "V" shape exhibited in both graphs is very similar to that shown in [74, 75] for the response to emotional media, as expected. Figure 8 shows some sample frames taken from the extremes of the two emotional dimensions. Table 11 shows agreement statistics for the annotations of experienced emotion. As expected, the inter-annotator agreement is low; the individual emotional experience is highly subjective. It is worth noting the differences in agreement between arousal and valence. Arousal trends are more consistent, as shown by the higher Pearson correlation score, while valence absolute values are more consistent, as shown by the lower difference metrics. Finally, expected and intended emotion ended up being highly similar, with correlation coefficients of 0.74 for arousal and 0.70 for valence.

4 Multimodal saliency-based computational framework

Herein, we propose an extension of our baseline multimodal saliency-frontend (TMM13) [12], which is a unified energy-based framework for audio-visual saliency computation. Specifically, it is an improved synergistic approach to the problem of audio-visual salient event detection and movie summarization, also employing text saliency computation, initially presented in ICIP15 [14]. For the remainder of this paper, we refer to the two different frameworks as ICIP15 and TMM13. Next, we summarize the proposed framework and we present new objective results on all videos included in COGNIMUSE database. Figure 9 shows an overview of the proposed video summarization system.

4.1 Visual analysis

For the spatio-temporal visual saliency estimation, we use an energy-based model [14, 79], assumed to be more relevant to the cognition-inspired saliency methods [1, 80]. For the extraction of visual features, our visual saliency model uses biologically plausible spatio-temporal 3D Gabor filters. First, the original RGB video volume is transformed into (L,a,b) space and split into two streams, i.e., luminance and color. In the resulting video volume $\mathbf{I}_{Lab}(x, y, t)$, the L^* component expresses the perceptual response to luminance, while $a^*(x, y, t), b^*(x, y, t)$ describe the differences between red-green and yellow-blue colors, respectively. The double color opponent cells that exist in the primary visual cortex V1, which are used in color constancy applications, are modeled using the method in [81]. The resulting color stream that expresses both the color intensity and contrast is given by: $C_{ab}(x, y, t) = \sqrt{(a^*)^2 + (b^*)^2}$. Afterwards, a filtering process [79] follows, called *Spatio-Temporal Dominant Analysis (STDA)*, which is applied on both channels: luminance and color.

4.1.1 3D Gabor Filtering

For the filtering of the luminance channel, we use oriented Gabor filters, due to their biological plausibility and

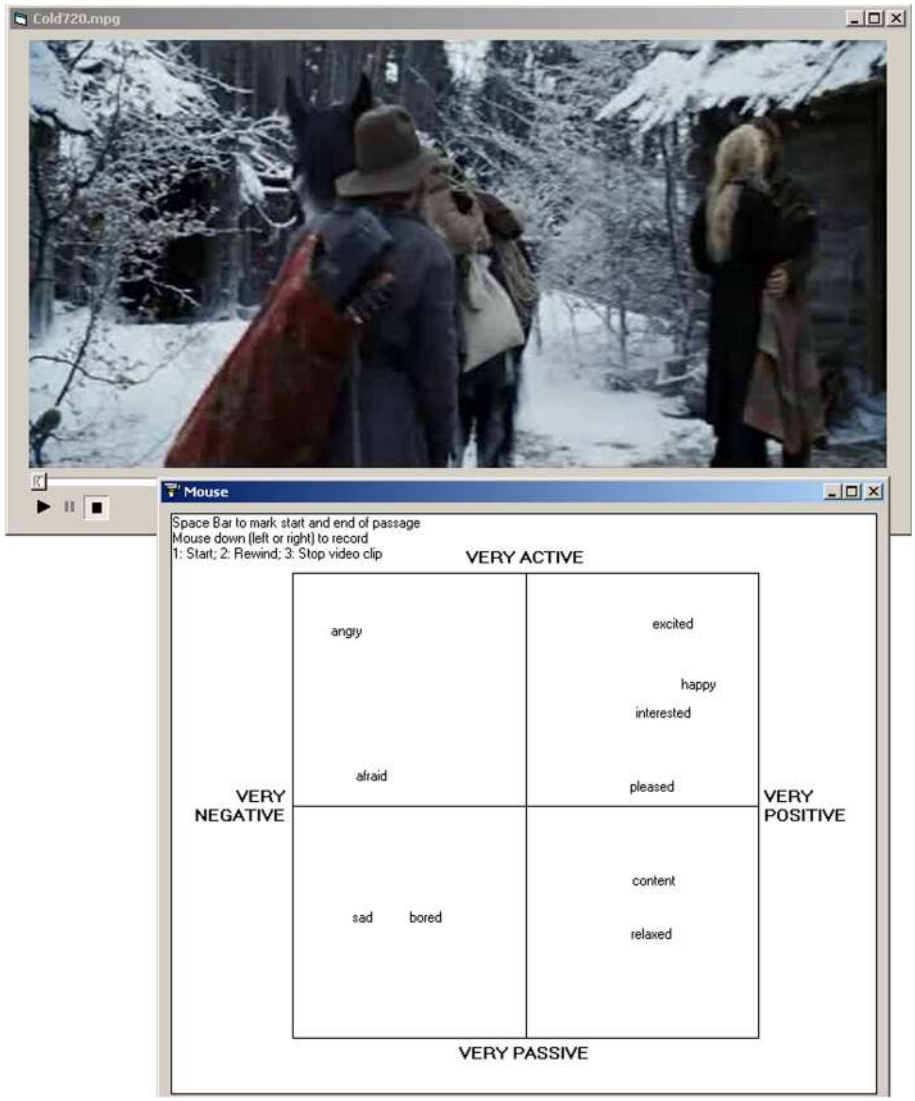

Fig. 6 The interface of Feeltrace

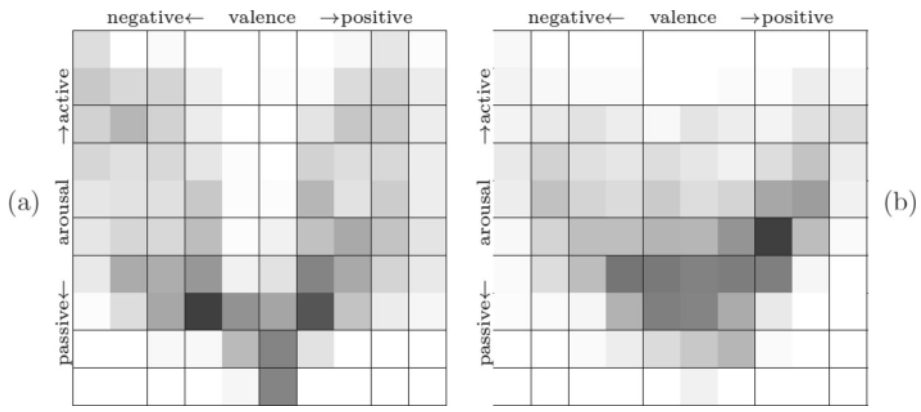

Fig. 7 Joint valence-arousal histograms for **a** intended and **b** expected emotion (*darker* signifies higher value)

Fig. 8 Sample frames for **a** low arousal, **b** high arousal, **c** very negative valence, and **d** very positive valence

their uncertainty-based optimality. Specifically, quadrature pairs of 3D (spatio-temporal) Gabor filters, which have identical central frequencies and bandwidth, are applied. Even though the 3D filtering is a time consuming process, due to the complexity of all required 3D convolutions, the Gabor filters used are separable [82], which means that we can filter each dimension on its own. Thus, we are able to apply only 1D convolutions instead of 3D [79]. The spatio-temporal filterbank used have $K_G = 400$ Gabor filters (isotropic in the spatial components), arranged in five spatial scales, eight spatial orientations and ten temporal frequencies. The spatial scales and orientations have been selected to cover a squared 2D frequency plane in a similar way as in [83]. Additionally, we have used ten temporal Gabor filters, five at positive and five at negative center frequencies, taking into account the 3D spectrum symmetries. For the static filterbank we have used the same spatial parameters with zero temporal frequency ($L_G = 40$ filters). The benefit of using both types of filterbanks, is that the spatio-temporal filterbank can detect motion activities, while the static one can find significant image regions that could attract human attention, e.g., specific textures or strong edges.

4.1.2 Postprocessing

After the filtering process, for each filter i, we obtain a quadrature pair output which corresponds to the even- and odd-phase 3D filter outputs. The total Gabor energy for each filter can be then computed by taking the sum of the squared energy of these two outputs, obtaining K_G *energy volumes* for the spatio-temporal part and L_G for the static part. The first step of *Dominant Component Analysis* [83, 84] is then performed to both energy volumes (spatio-temporal and static), in order to form one volume for each one of these independent filtering parts. In order to make our model more robust, instead of keeping only the dominant energy, we keep the $N_B = 6$ highest spatio-temporal energies for each voxel; followed by the computation of their minimum value, which was experimentally found to perform best. This way, we have obtained two raw energy volumes for each stream (luminance and color); the spatio-temporal dominant energy *STDE* and the static dominant energy *SDE*, which are further smoothed by applying a *temporal moving average* (TMA). The produced energy maps can be mapped to a 1D map giving time-varying saliency features. Concluding, 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 feature vectors, which correspond to the four different TMA energies, along with their first and second temporal derivatives constitute the feature set for the visual modality.

4.2 Audio analysis

The estimation of auditory saliency is assumed to be a problem of assigning a measure of interest to audio frames, based on spectro-temporal cues. The importance of amplitude and frequency changes for audio saliency has actually motivated various studies where subject responses were measured with respect to tones of modulated frequency or loudness [3, 4, 85]. Specifically, for the analysis of the audio stream, an energy-based feature set for saliency-modeling was adopted and approached using the nonlinear differential energy operator based on the Teager-Kaiser Energy Operator [86]. The Teager-Kaiser Energy Operator (TEO), which can track the instantaneous energy of a source, is given by: $\Psi[x] = \dot{x}^2 - x\ddot{x}$, where $\dot{x} = dx/dt$. Teager energy is actually only meaningful in narrowband signals, that is why the

Table 11 Inter-annotator agreement

Metric	Valence	Arousal
Correlation	0.293	0.409
Difference of means	0.288	0.411
Mean abs. difference	0.445	0.513
Krippendoff's α ordinal (7 levels)	0.308	0.152
Cohen's k (7 levels)	0.035	0.029

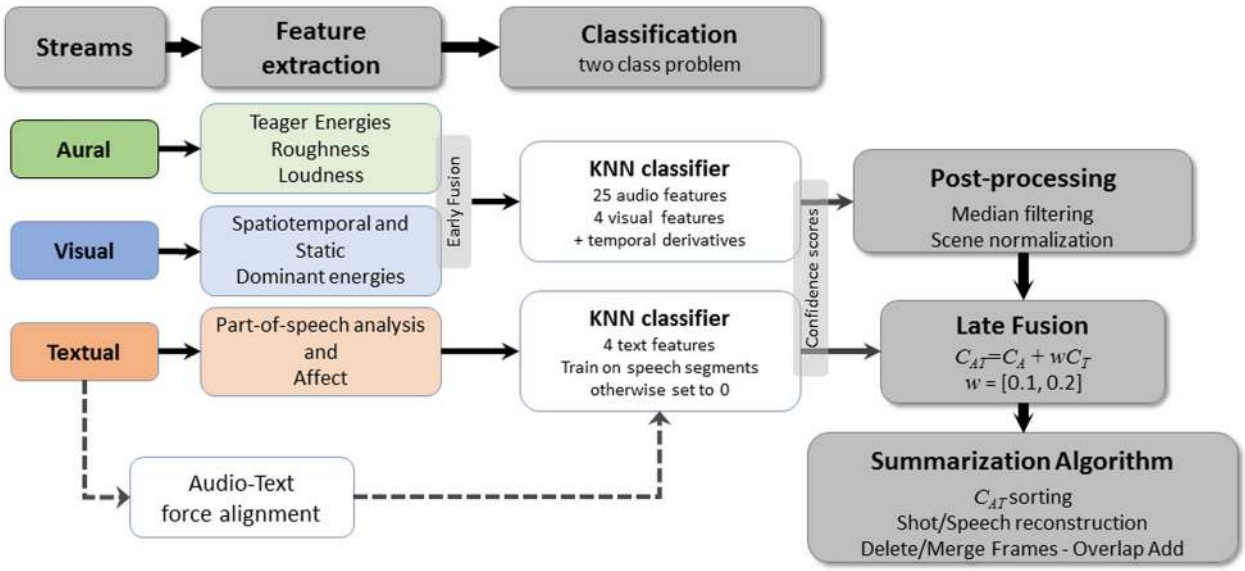

Fig. 9 System overview for multimodal saliency and summarization

application of the energy operator is preceded by band-pass filtering, using 25 linearly-spaced Gabor filters with 50% overlap, resulting in mean instantaneous energies derived from each filter.

In order to enhance the detection and the classification accuracy, two more perceptual features were computed, both assumed to correlate to the functioning of the human auditory system and to attention [87]. (a) Roughness, proposed in [88], is an estimation of the sensory dissonance of a sound, expressing a sense of roughness of a sound due to rapid fluctuations in its amplitude. A variant model [89], using a more complex weighting, has been used in this work. (b) Loudness corresponds to the perceived sound pressure level and for its computation the model proposed in [87] was used. The resulting temporal sequence of the 27 features, along with their first and second temporal derivatives constitute the feature set for the audio modality.

4.3 Text analysis

We have extended the text analysis of TMM13 [12], and we have included affective modeling of single words extracted from the subtitles and transcripts included in the database. Analysis of text to estimate affect is a relatively recent research topic that has attracted great interest with application to numerous domains spanning from tweet analysis [90] to dialogue systems [91]. Metrics such as high arousal and high absolute valence are expected to be good indicators for words related with salient events [30]. This is due to the fact that humans usually pick content (i.e., movies, music) based on its affective characteristics; hence, affective features draw particular interest to content delivery systems that provide personalized multimedia content, automatically extract highlights or summaries.

Text analysis in TMM13, was based on part-of-speech tagging and is summarized next: (a) extraction of the movie transcript from the English subtitle file, (b) part-of-speech tagging, (c) audio-text alignment, (d) assignment of a text saliency value $\{0.2, 0.5, 0.7, 1\}$ to each word, and finally (e) text saliency computation and assignment of a text saliency value to each frame.

ICIP15 extents this baseline method for text saliency computation using also affective modeling, based on the assumption that *"semantic similarity can be translated to affective similarity"* [92]. The semantic similarity metric can be computed within the framework of (corpus-based) distributional semantic models, relying on the hypothesis that *"similarity of context implies similarity of meaning"* [93].

Thus, a word w is characterized regarding its affective content in a continuous interval space $[-1, 1]$ consisting of three dimensions, namely valence (v), arousal (a), and dominance (d). The affective content of w, for each dimension, 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), for more details see [92]. A contextual window of size $2H+1$ words is centered on the word of interest w_i and lexical features are extracted. For every instance of w_i in the corpus, the H words left and right of w_i formulate a feature vector x_i. For a given

value of H, the semantic similarity between two words, w_i and w_j, is computed as the cosine of their feature vectors.

In this work, the context-based metric was applied with $H = 1$ over a web-harvested corpus, while the contextual features were weighted using a binary scheme. The word affective ratings were estimated using as seeds 600 entries selected from the ANEW lexicon [94]. More details about the corpus, seed selection, and the training of λ weights can be found in [92]. The three affective ratings plus the POS tagging values constitute the four features for the modeling of text saliency.

5 Experimental evaluation and movie summarization algorithm

5.1 Machine learning evaluation

For the multimodal saliency event detection task, a machine learning classification approach has been adopted, where a K-Nearest Neighbor Classifier (KNN) was employed, instead of experimenting with various fusion schemes as in TMM13 [12]. We used the combination of the 4 visual plus the 27 audio features, along with their first and second temporal derivatives (computed over 3 and 5 frames respectively). The 4 text features comprised the textual feature vector, where a second KNN model was built. The framework followed the same principles as explored in [12, 95] and being further refined in ICIP15 [14]. Specifically, we considered frame-wise saliency as a two-class classification problem, while a confidence score was also determined for every classification result (i.e., each frame), in order to obtain results for various compression rates and thus produce summaries of various lengths.

For the various data included in the COGNIMUSE database, different evaluation setups were adopted. For the Hollywood movies, a sevenfold cross-validation was applied, where the annotated frames from six movies were used for training and tested on the seventh. For the travel documentaries a five fold cross-validation was considered, where in the same manner four documentaries where used for training and the fifth for testing. For GWW, two types of evaluation were followed: (a) training on the seven movies or (b) training on all database data (i.e., seven movies plus five documentaries) and testing on GWW, so as to explore how the blending of the two data genres affects the classification accuracy of a movie.

5.2 Movie summarization algorithm

The movie summarization algorithm presented here is a modification of the TMM13 algorithm [12]. New features were included, which proved to be imperative, as detailed in the subjective evaluation (see Sec. 5.3). The initial purpose of those features was to make the automatically produced summaries smoother, regarding audio and video transitions, but also to enhance the comprehension concerning the semantics.

For the creation of the summaries, we have used the classifier's output, consisting of frames classified as salient; thus, segments or frames, chosen based on high confidence scores, form a binary indicator function curve, representing the most salient audio-visual-text events. The steps that have been followed are:

1. Median filtering of the audiovisual confidence scores C_{AV}, in order to obtain a smoother and coarse AV attention curve, followed by scene-based normalization (the boundaries of the scenes were extracted from the manual segmentation).
2. Text confidence scores C_T trained only on speech segments were used, while frames without speech were set to zero.
3. Late fusion of the audiovisual and text confidence scores, where a fixed weight w for the text stream was chosen: $C_{AVT} = C_{AV} + w \cdot C_T$. The text weight was experimentally set to be $w = 0.10$ or $w = 0.20$.
4. Confidence scores sorting so as to define the segments to be included in the summary; a five times faster summary than real time was created.
5. Boundary correction of the extracted events is performed, so as to produce summaries including meaningful events apart from salient only. Hence, shot and "speech" reconstruction is performed, where the boundaries of the manually segmented shots and the word boundaries are used to assure that no word "clipping" will occur; where ideas from mathematical morphology are used and specifically, the reconstruction opening: $\rho^{-}(M|X) \triangleq$ connected components of X intersecting M [96] (for more details see [95]). This reconstruction process is of high value for the produced summaries, for understanding of the semantics and the creation of smoother transitions.
6. The final step of the algorithm includes a process based on TMM13 [12] for the combination of frames into segments. Thus, (a) segments that are shorter than N frames are deleted from the summary, while neighboring segments selected for the summary are merged if they are less than K frames apart, where $N = 7$ and $K = 20$ (experimentally tuned). (b) The final rendering of segments into a summary is performed by using simple overlap-add to tailor together neighboring segments.

Figure 10 shows the monomodal saliency curves for the three modalities and the multimodal saliency curves (av vs avt), where the weight for the text modality was set to

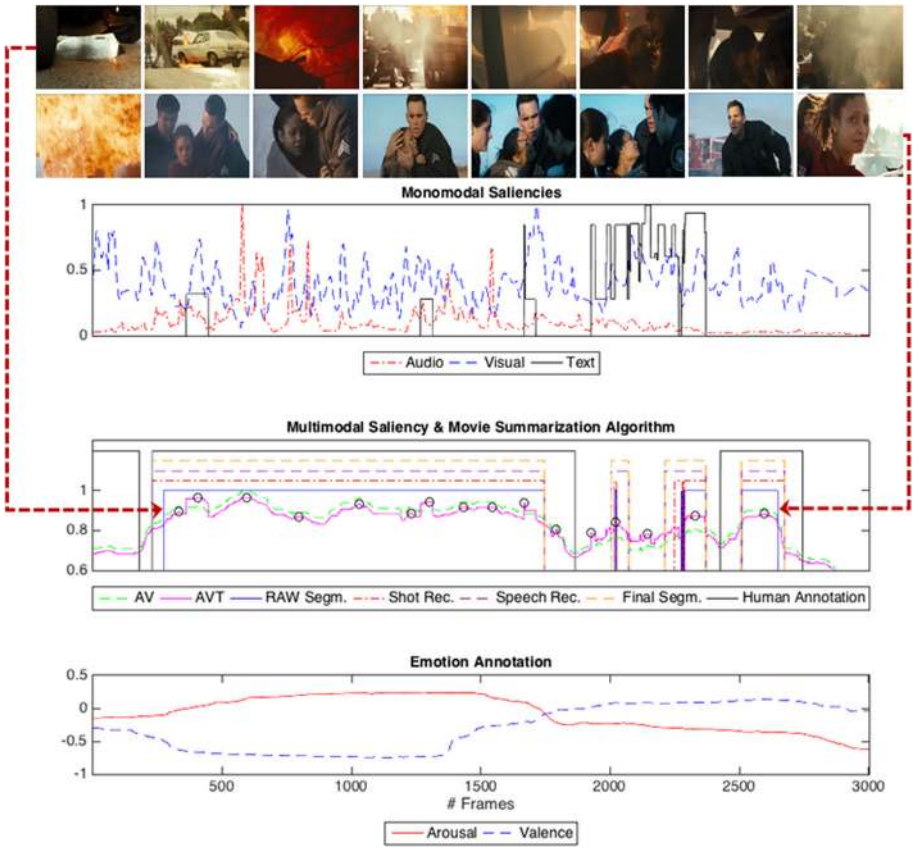

Fig. 10 Illustration of saliency curves, summarization algorithm, and emotion curves. Monomodal saliency curves for the three modalities (audio, visual, and text). Multimodal saliency, av vs. avt, where the weight $w = 0.1$ and summarization algorithm (illustrating the effect of reconstruction opening by showing the raw (automatically) selected segments, the shot reconstructed and the speech reconstructed segments) and the manual selected segments (by the human annotators). Emotion annotation illustrated by the two dimensions arousal and valence; for 3000 frames, scene from the movie "CRA". Keyframes (*top*) correspond to the indicated saliency peaks. Best viewed in color

$w = 0.1$. Moreover, the different steps of the movie summarization algorithm can be seen (illustrating in the figure the effect of reconstruction opening by showing the raw (automatically) selected segments, the shot and speech reconstructed segments as well as the final selected segments vs the human annotation. Additionally, for the specific segment from "CRA," emotion annotation is also illustrated by the two dimensions, i.e., arousal and valence. On top video frames associated with high saliency values marked with circles on the multimodal curve are shown.

5.3 Results and discussion

5.3.1 Objective machine learning evaluation

In Fig. 11, ROC curves for saliency classification can be seen for the seven Hollywood movies, while changing the percentage of frames in summary (from 1–100%, where 100% corresponds to perfect recall score), for audio on audio (A-A), visual on visual (V-V), audio-visual on audiovisual (AV-AV), and audiovisual-text on audio-visual-semantics (AVT-AVS) annotation for the TMM13 and ICIP15 methods. The results for AV-AV and AVT-AVS evaluation and the ICIP15 method are produced using the new summarization algorithm, while for the A-A and V-V evaluation the sorted median filtered confidence scores are used. For the TMM13 method, the results are shown for the sorted RAW confidence scores as presented in [12]. We have to emphasize that the employed classification approach is a frame-wise detection task, while the ground truth salient events are annotated as segments and not as single frames.

As shown in the figure, the method developed for ICIP15 outperforms the one in TMM13, both when evaluating each modality individually as well as when two (AV) or three (AVT) modalities are fused together. Greater improvement can be seen for the monomodal salient event detection than the multimodal one and specifically for the audio modality (A-A evaluation). However, the audiovisual modality (AV-AV) accomplishes a quite high score as well.

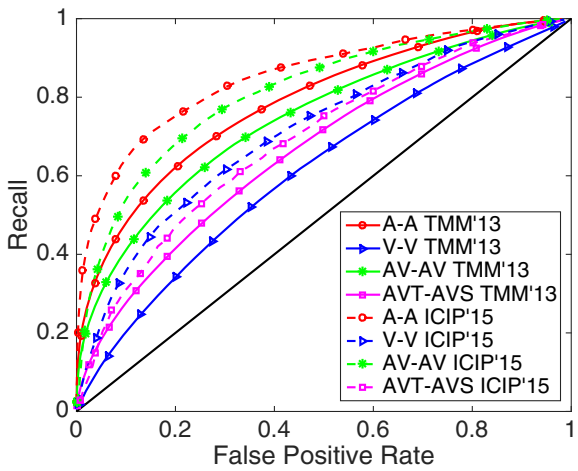

Fig. 11 Objective evaluation comparative results for TMM13 and ICIP15, where saliency classification ROC curves for the different modalities are presented

Figure 12 shows results for GWW and the ICIP15 method, where again we notice that best performance is accomplished for audio on audio (A-A) evaluation. Note that in this case, the improvement when the text modality is fused with the audiovisual modality is not obvious. Regarding the two different types of evaluation, where our aim was to examine the blending of the different data genres on the performance, we note that best performance is accomplished when the training is performed on the Hollywood movies only. This is probably due to the fact that the format of the specific documentaries is rather unstructured, not following an explicit scenario, using a more informal and everyday colloquial language. Further, even though there are components that resemble the structure of a movie (e.g., captioned frames, dynamic image sequences and graphics), the image capturing in the specific travel series resembles amateur film-making rather than the systematic and rule-based that is utilized in movies.

Regarding the travel documentary results, as seen in Fig. 13, the best performance is again accomplished for audio on audio evaluation (A-A) for shorter summaries, while the audiovisual (AV) fusion is observed to be slightly better for longer summaries. The text modality seems to worsen the AV case significantly, probably as previously mentioned due to the unstructured dialogues and the use of every day language. Moreover, we have to emphasize that even though the proposed algorithm is not domain-specific, different characteristics could be considered significant and salient for travel documentaries. For instance, in a travel documentary summary the user would be probably more interested in watching just the important places rather than the conversational parts.

Finally, Fig. 14 shows evaluation results using AUC (area under curve) as a metric comparing all modalities and evaluation setups (i.e., A-A, V-V, AV-AV, AVT-AVS). For the Hollywood movies, results are obtained both from the baseline TMM13 and the ICIP15 movie summarization systems, while for the five travel documentaries and GWW, we use the extended ICIP15 system; where GWW#1 shows AUC results obtained when GWW was trained on the movies and GWW#2 shows results when the training was performed on all data. Here, we can more clearly observe the superiority of the audio modality for all setups. Moreover, we notice that the fusion of the audio and visual modality (AV) accomplishes the second best result in all cases, while the visual modality (V) yields better results than the AVT when the movies and the travel documentaries are evaluated using the ICIP15 method.

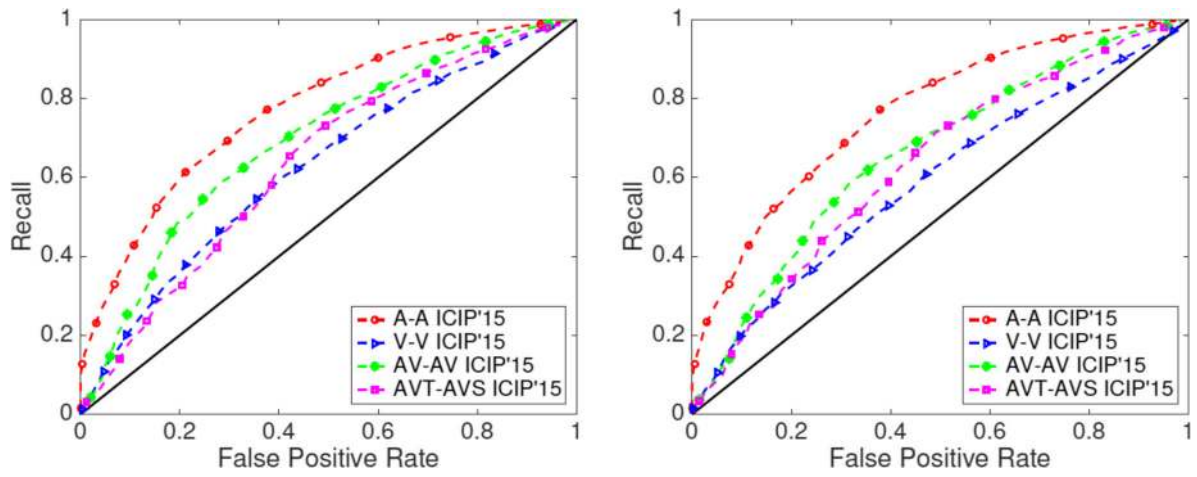

Fig. 12 Objective evaluation results for GWW. Saliency classification ROC curves for the different modalities for GWW trained in the Hollywood movies (*left figure*). Saliency classification ROC curves for the different modalities for GWW trained in all available data, thus seven movies and five documentaries

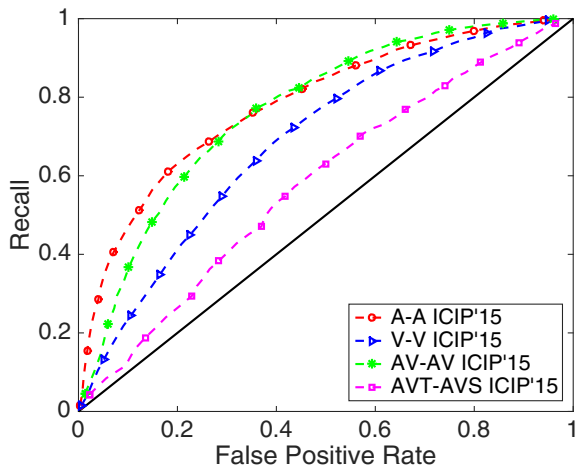

Fig. 13 Objective evaluation results for the five travel documentaries. Saliency classification ROC curves for the different modalities, using five fold cross-validation

On the other hand, for GWW the AVT outperforms the visual modality. Best overall performance can be seen for the evaluation of the seven movies evaluated with the newly developed method. Even though the text modality in the presented experiments does not show to improve the performance, we have to emphasize the fact that the text information can assist in many ways; i.e., segments that are not selected as salient by the other modalities could be highlighted by the text and most importantly it assists the summarization algorithm through the speech reconstruction process.

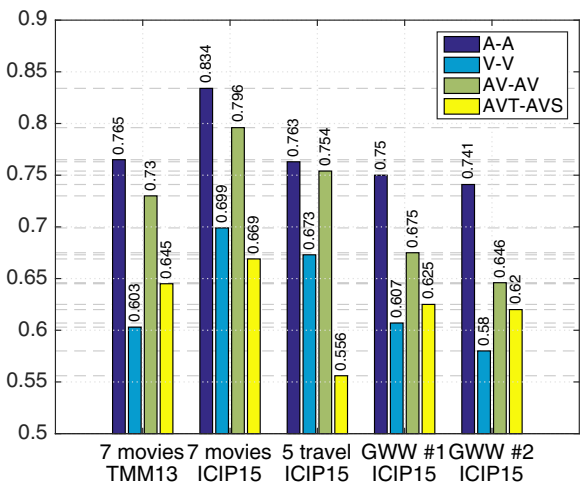

Fig. 14 Results overview over the COGNIMUSE database. Evaluation results using AUC (Area Under Curve) comparing all modalities and evaluation set-ups using the baseline TMM13 and the ICIP15 movie summarization system

5.3.2 *Subjective qualitative evaluation on the Hollywood movie summaries*

In addition to the subjective evaluation of the Hollywood movies, using as ground truth the saliency annotation of the database, a user qualitative study was conducted as well, where 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 [12] and as described in detail in [97]. Four summaries, in total, were evaluated: two summaries based on the ICIP15 method [14] 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 TMM13 [12] (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 s of the original clip. The subjects participating in the evaluation first viewed the original half-hour clip, followed by the four summaries (ca. 6 min each) in randomized order.

Figure 15 shows that the ICIP15 method performs much better in terms of both metrics compared to the best performing summaries based on fusion (FUS) and the fast-forward (FF) like summaries, where the subjective ratings were up to 80% for informativeness and 90% for enjoyability. Summarizing the main conclusions regarding the ICIP15 method, we could say 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 probably due to the fact that in action movies, the algorithm tends to favor high intensity events, such as battles; thus, using a higher text weight more salient textual events (e.g., dialogues) are included. In dialogue-based movies (e.g., CRA), smaller text weight is required so as to include the usually few existing action scenes. The users confirmed that a good summary has to be balanced with respect to the variability of events. Additionally, shot and speech reconstruction of the selected segments contributed a lot to enjoyability, since it resulted to smoother transitions and to semantically coherent events aiding the comprehension of the plot.

Regarding the fusion (FUS)-based summaries of TMM13, the users conceded that they were enjoyable, however not as informative, reflected also on the presented results. Finally, concerning the fast-forward (FF) like summaries only few of the users realized that they were intentionally added for evaluation (as a naive approach indicating a lower bound for our metrics). Those summaries actually helped us to prove that a uniform sampling of movie frames is not adequate in order to create acceptable summaries, since they were judged as "choppy," having fast transitions and non-existing semantics. Whenever these summaries were assigned a

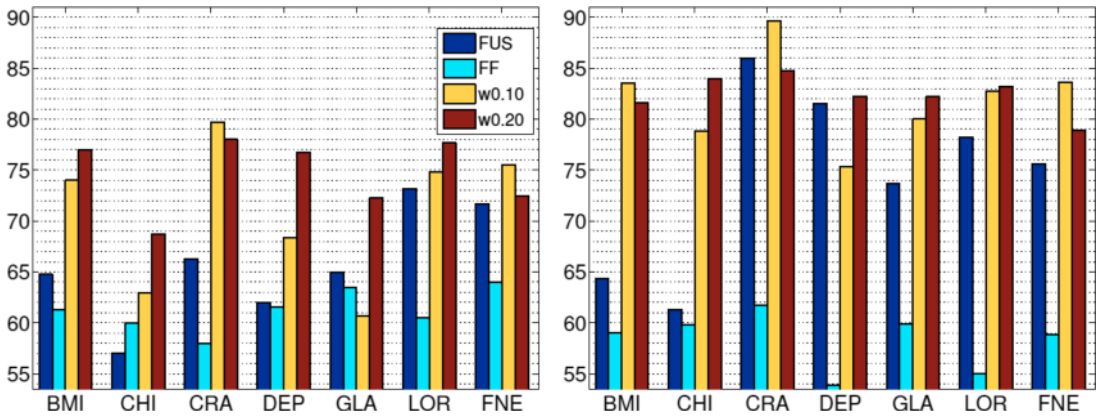

Fig. 15 Subjective evaluation results by 20 humans. Informativeness (*left figure*) and enjoyability (*right figure*) results of AVT summaries at (×5) rate were FUS that denotes the best summary obtained using fusion methods as presented in [12], FF denotes "fast-forward", while the two newly produced summaries are differentiated by the weight used in the text modality

high informativeness score, it was because they included visual information uniformly taken from the whole original clip; a significant observation, leading us to conclude that a summary needs to include elements from the full duration of the original clip.

Even though, the existence of datasets for evaluation of movie summarization algorithms is crucial, human quality evaluations, as shown here, are also essential for improving the quality of the produced summaries, by getting feedback that aids in the development of systems that further enhance the viewing experience.

6 Conclusions

In this work, a multimodal video-oriented database annotated with mono- and multimodal sensory and semantic saliency, audio and visual, cross-media semantics, and emotion is presented and proposed for training and evaluation of event detection algorithms and video summarization systems. The purpose of this database is to form some common ground and ground truth data—denoting conspicuous events—not only for robust benchmarking of the produced summaries but even for computational modeling intended for machine-based understanding. The multiple saliency annotations in all three modalities (i.e., audio, video, and text) offer the possibility to train and evaluate different task, i.e., each stream independently. Additionally, we presented a video summarization system and multimodal computational algorithms that employ advanced state-of-the-art methods for perceptually salient event detection. Our experimental evaluation using human saliency annotation as ground truth confirms that the framework is promising as it outperforms other methods over the COGNIMUSE database. Additionally, the qualitative user-based study of the automatically produced summaries verifies the appropriateness of both the proposed summarization algorithm and the database. For future work, we intend to further refine our methods and the movie summarization algorithm automating the weight selection for the text modality as well as the segmentation of shots and scenes. Moreover, we intend to incorporate in our experimental framework the data acquired from the audio-visual events, the cross-media relations, and the emotion annotation, so as to take advantage of this information, and produce summaries based on user-preferences. Finally, we envision to make the COGNIMUSE database a public state-of-the-art dataset for attention-related tasks and multimodal understanding applications.

Endnotes

[1] COGNIMUSE was a research project, where the multisensory and sensory-semantic information modeling was investigated, integrating all three modalities to detect salient events. For more information see http://cognimuse.cs.ntua.gr/

[2] 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).

[3] http://www.anvil-software.de/index.html

[4] http://tla.mpi.nl/tools/tla-tools/elan/

Acknowledgements

The majority of the work by G. Evangelopoulos was performed while he was affiliated with the National Technical University of Athens. The majority of the work by N. Malandrakis and A. Potamianos was performed while they were

affiliated with the Technical University of Crete. The authors would like to thank Elias Iosif for his contribution on affective text analysis, Nassos Katsamanis for his insightful ideas and discussions, Tolis Apostolidis for providing the expert movie summaries, Isidoros Rodomagoulakis, Monica Maragos, Valia Sfika, Kevis Maninis, Giorgos Anastasiou, Antigoni Tsiami, Tassos Tsiamis, Gio Panagiotaropoulou and Olivia Karathanou for their contribution in annotating different parts of the database. Finally, we would like to thank the students of NTUA for participating in the subjective evaluation and for their valuable comments regarding the summaries.

Funding
This research work was supported by the project "COGNIMUSE" which was implemented under the "ARISTEIA" Action of the Operational Program Education and Lifelong Learning and was co-funded by the European Social Fund and Greek National Resources. For more information please see: http://cognimuse.cs.ntua.gr/. It was also partially supported by the European Union under the projects MOBOT with grant FP7-600796, and BabyRobot with grant H2020-687831.

Authors' contributions
All authors read and approved the final manuscript.

Competing interests
The authors declare that they have no competing interests.

Publisher's Note

Author details
[1]School of Electr.& Comp. Enginr., National Technical University of Athens, 15773 Athens, Greece. [2]McGovern Institute for Brain Research at MIT MIT, MA 02139 Cambridge, USA. [3]Signal Analysis and Interpretation Laboratory (SAIL), USC, CA 90089 Los Angeles, USA. [4]Cognitive Systems Research Institute, Athens, Greece.

Received: 6 August 2016 Accepted: 21 June 2017
Published online: 07 August 2017

References

1. C Koch, S Ullman, Shifts in selective visual attention: towards the underlying neural circuitry. Hum. Neurobiol. **4**(4), 219–227 (1985)
2. L Itti, C Koch, Computational modelling of visual attention. Nat. Rev. Neurosci. **2**(3), 194–203 (2001)
3. C Kayser, CI Petkov, M Lippert, NK Logothetis, Mechanisms for allocating auditory attention: an auditory saliency map. Curr. Biol. **15**(21), 1943–1947 (2005)
4. M Elhilali, J Xiang, SA Shamma, JZ Simon, Interaction between attention and bottom-up saliency mediates the representation of foreground and background in an auditory scene. PLoS Biol. **7**(6) (2009)
5. MI Posner, SE Petersen, The attention system of the human brain. Ann. Rev. Neurosci. **13**(1), 25–42 (1990)
6. EI Knudsen, Fundamental components of attention. Ann. Rev. Neurosci. **30**, 57–58 (2007)
7. D Walther, C Koch, Modeling attention to salient proto-objects. J. Neural Netw. **19**(9), 1395–1407 (2006)
8. T Kadir, M Brady, Saliency, scale and image description. Int'l. J. Comput. Vis. **45**(2), 83–105 (2001)
9. K Rapantzikos, Y Avrithis, S Kollias, Spatiotemporal features for action recognition and salient event detection. Cogn. Comput. Special Issue Saliency Atten Visual Search Picture Scan. **3**(1), 167–184 (2011)
10. Y Ma, XS Hua, L Lu, H Zhang, A generic framework of user attention model and its application in video summarization. IEEE Trans. Multimedia. **7**(5), 907–919 (2005)
11. A Money, H Agius, Video summarization: a conceptual framework and survey of the state of the art. J. Visual Commun. Image Represent. **19**(2), 121–143 (2008)
12. G Evangelopoulos, A Zlatintsi, A Potamianos, P Maragos, K Rapantzikos, G Skoumas, Y Avrithis, Multimodal saliency and fusion for movie summarization based on aural, visual, textual attention. IEEE Trans. Multimedia. **15**(7), 1553–1568 (2013). doi:10.1109/TMM.2013.2267205
13. K Pastra, S Piperidis, Video search: new challenges in the pervasive digital video era. J. Virtual Reality Broadcast. **3**(11) (2006)
14. P Koutras, A Zlatintsi, E.Iosif, A Katsamanis, P Maragos, A Potamianos, in *Proc. Int'l Conf. on Image Process*. Predicting audio-visual salient events based on visual, audio and text modalities for movie summarization, (Quebec, 2015)
15. Y Liu, F Zhou, W Liu, F De la Torre, Y Liu, in *Proc. of the 18th ACM international conference on Multimedia*. Unsupervised summarization of rushes videos (ACM, 2010), pp. 751–754
16. SF de Avila, AB Lopes, A da Luz Jr, A de Albuquerque Araujo, VSUMM: A mechanism designed to produce static video summaries and a novel evaluation method. Pattern Recogn. Lett. **32**(1), 56–68 (2011)
17. YJ Lee, J Ghosh, K Grauman, in *Proc. Int'l. Conf. Computer Vision and Pattern Recognition*. Discovering important people and objects for egocentric video summarization, (2012)
18. M Wang, R Hong, G Li, Z-J Zha, S Yan, T-S Chua, Event driven web video summarization by tag localization and key-shot identification. IEEE Trans. Multimedia. **14**(4), 975–985 (2012)
19. A Khosla, R Hamid, C-J Lin, N Sundaresan, in *Proc. Int'l. Conf. Computer Vision and Pattern Recognition*. Large-scale video summarization using web-image priors, (2013)
20. Z Lu, K Grauman, in *Proc. Int'l. Conf. Computer Vision and Pattern Recognition*. Story-driven summarization for egocentric video, (2013)
21. Y Wang, Z Liu, J-C Huang, Multimedia content analysis using both audio and visual clues. IEEE Signal Process. Mag. **17**, 12–36 (2000)
22. Y-F Ma, X-S Hua, L Lu, H-J Zhang, A generic framework of user attention model and its application in video summarization. IEEE Trans. Multimedia. **7**(5), 907–919 (2005)
23. D Potapov, M Douze, Z Harchaoui, C Schmid, in *Proc. European Conference on Computer Vision*. Category-specific video summarization, (2014). http://hal.inria.fr/hal-01022967
24. P Over, AF Smeaton, G Awad, in *Proc. 2nd ACM TRECVID Video Summarization Workshop*. The Trecvid 2008 BBC rushes summarization evaluation, (2008)
25. BT Truong, S Venkatesh, Video abstraction: a systematic review and classification. ACM Trans. Multimedia Comput. Commun. Appl. **3**(1), 3 (2007)
26. C-W Ngo, Y-F Ma, H-J Zhang, Video summarization and scene detection by graph modeling. Circuits Syst. Video Technol. **15**(2) (2005)
27. C-Y Lin, in *Proc. Text Summarization Branches, ACL Workshop*. Rouge: a package for automatic evaluation of summaries, (Barcelona, Spain, 2004)
28. G Kim, L Sigal, EP Xing, in *Proc. Int'l. Conf. Computer Vision and Pattern Recognition*. Joint summarization of large-scale collections of web images and videos for storyline reconstruction, (2014)
29. Y Song, J Vallmitjana, A Stent, A Jaime, in *Proc. Int'l. Conf. Computer Vision and Pattern Recognition*. TVSum: summarizing web videos using titles, (2015)
30. AF Smeaton, P Over, W Kraaij, in *Proc. MIR-06*. Evaluation campaigns and TRECVid, (2006)
31. M Sun, A Farhadi, S Seitz, in *Proc. European Conf. on Computer Vision*. Ranking domain-specific highlights by analyzing edited videos, (Springer, Cham, 2014), pp. 787–802
32. M Gygli, H Grabner, H Riemenschneider, LV Gool, in *Proc. European Conf. on Computer Vision*. Creating summaries from user videos, (Springer, Cham, 2014)
33. R Radhakrishnan, A Divakaran, P Smaragdis, in *Proc. IEEE WASPAA*. Audio analysis for surveillance applications, (ACM, 2005)
34. M Xu, C Xu, L Duan, JS Jin, S Luox, Audio keywords generation for sports video analysis. ACM Trans. Multimedia Comput. Commun. Appl. **4**(2), 1–23 (2008)
35. T Heittola, A Mesaros, A Eronen, T Virtanen, in *Proc. 18th European Signal Processing Conf*. Audio context recognition using audio event histograms, (ACM, 2010)
36. E Parizet, V Koehl, in *Proc. Euronoise*. Categorisation: a useful tool for applied perceptive studies, (2006)
37. RM Schafer, *The soundscape: Our sonic environment and the tuning of the world*. (Simon and Schuster, 1993)
38. AL Brown, J Kang, T Gjestland, Towards standardization in soundscape preference assessment. Appl. Acoust. **72**(6), 387–392 (2011)

39. M Raimbault, D Dubois, Urban soundscapes: experiences and knowledge. Cities. **22**(5), 339–350 (2005)
40. J Salamon, C Jacoby, JP Bello, in *Proc. 22nd ACM Int'l. Conf. on Multimedia*. A dataset and taxonomy for urban sound research, (2014)
41. SR Payne, WJ Davies, MD Adams, Research into the practical and policy applications of soundscape concepts and techniques in urban areas. Technical report, DEFRA, HMSO, London, UK (2009)
42. BC Russell, A Torralba, KP Murphy, WT Freeman, Labelme: a database and web-based tool for image annotation. Int'l J. Comput. Vis. **77**(1-3), 157–173 (2008)
43. J Deng, W Dong, R Socher, L-J Li, K Li, L Fei-Fei, in *Proc. IEEE Conf. Computer Vision and Pattern Recognition*. Imagenet: a large-scale hierarchical image database, (2009)
44. R Poppe, A survey on vision-based human action recognition. Image Vis. Comput. **28**(6), 976–990 (2010)
45. S Sadanand, JJ Corso, in *Proc. IEEE Conf. Computer Vision and Pattern Recognition*. Action bank: a high-level representation of activity in video, (2012)
46. M Bregonzio, S Gong, T Xiang, in *Proc. IEEE Conf. Computer Vision and Pattern Recognition*. Recognising action as clouds of space-time interest points, (2009)
47. Z Zhang, D Tao, Slow feature analysis for human action recognition. IEEE Trans. PAMI. **34**(3), 436–450 (2012)
48. Y Yang, I Saleemi, M Shah, Discovering motion primitives for unsupervised grouping and one-shot learning of human actions, gestures, and expressions. IEEE Trans. PAMI. **35**(7), 1635–1648 (2013)
49. K Maninis, P Koutras, P Maragos, in *Proc. Int'l Conf. Image Processing*. Advances on action recognition in videos using and interest point detector based on multiband spatio-temporal energies, (IEEE, 2014)
50. A Karpathy, G Toderici, S Shetty, T Leung, R Sukthankar, L Fei-Fei, in *Proc. IEEE Conf. Computer Vision and Pattern Recognition*. Large-scale video classification with convolutional neural networks, (2014)
51. H Wang, MM Ullah, A Kläser, I Laptev, C Schmid, Evaluation of local spatio-temporal features for action recognition. in *Proc. BMVC*. BMVC 2009-British Machine Vision Conference (BMVA Press, 2009), pp. 124–1
52. H Wang, A Kläser, C Schmid, C Liu, Dense trajectories and motion boundary descriptors for action recognition. Int'l J. Comp. Vision. **103**(1), 60–79 (2013)
53. C Schüldt, I Laptev, B Caputo, in *Pattern Recognition, 2004. ICPR 2004. Proceedings of the 17th International Conference on*. Recognizing human actions: a local SVM approach, vol. 3 (IEEE, 2004), pp. 32–36
54. M Marszałek, I Laptev, C Schmid, in *Proc. IEEE Conference on Computer Vision & Pattern Recognition*. Actions in context, (2009)
55. H Kuehne, H Jhuang, E Garrote, T Poggio, T Serre, in *Proc. Int'l. Conf. on Computer Vision*. HMDB: a large video database for human motion recognition, (IEEE, 2011)
56. K Soomro, AR Zamir, M Shah, UCF101: A dataset of 101 human actions classes from videos in the wild (2012). arXiv preprint arXiv:1212.0402
57. M Bordegoni, G Faconti, S Feiner, M Maybury, T Rist, S Ruggieri, P Trahanias, M Wilson, A standard reference model for intelligent multimedia presentation systems. Comput. Standards Interfaces. **18**(6/7), 477–496 (1997)
58. C Bordier, F Puja, E Macaluso, Sensory processing during viewing of cinematographic material: computational modeling and functional neuroimaging. NeuroImage. **67**, 213–226 (2013). doi:10.1016/j.neuroimage.2012.11.031
59. NM Ross, E Kowler, Eye movements while viewing narrated, captioned, and silent videos. J. Vision. **13**(4), 1–17 (2013). doi:10.1167/13.4.1
60. K Pastra, COSMOROE: a cross-media relations framework for modelling multimedia dialectics. Multimedia Syst. **14**(5), 299–323 (2008)
61. S Arifin, PYK Cheung, Affective level video segmentation by utilizing the pleasure-arousal-dominance information. IEEE Trans. Multimedia. **10**(7), 1325–1341 (2008)
62. S Zhao, H Yao, X Sun, P Xu, X Liu, R Ji, in *Proc. 19th ACM Int'l. Conf. Multimedia*. Video indexing and recommendation based on affective analysis of viewers, (2011)
63. E Douglas-Cowie, R Cowie, I Sneddon, C Cox, O Lowry, M McRorie, J-C Martin, L Devillers, S Abrilian, A Batliner, N Amir, K Karpouzis, in *Proc. 2nd Int'l. Conf. Affective Comput. Intell. Interaction*. The HUMAINE database: addressing the collection and annotation of naturalistic and induced emotional data, (2007), pp. 488–500
64. A Schaefer, F Nils, X Sanchez, P Philippot, Assessing the effectiveness of a large database of emotion-eliciting films: A new tool for emotion researchers. Cogn. Emotion. **24**(7), 1153–1172 (2010)
65. S Koelstra, C Muhl, M Soleymani, J-S Lee, A Yazdani, T Ebrahimi, T Pun, A Nijholt, I Patras, DEAP: a database for emotion analysis using physiological signals. IEEE Trans. Affective Comput. **3**(1), 18–31 (2012)
66. M Soleymani, J Lichtenauer, T Pun, M Pantic, A multimodal database for affect recognition and implicit tagging. IEEE Trans. Affective Comput. **3**(1), 42–55 (2012)
67. Y Baveye, E Dellandrea, C Chamaret, L Chen, LIRIS-ACCEDE: a video database for affective content analysis. IEEE Trans. Affective Comput. **6**(1), 43–55 (2015)
68. M Kipp, in *Proc. Eurospeech-2001*. Anvil—a generic annotation tool for multimodal dialogue, (2001)
69. B Pellom, K Hacioglu, Sonic: the university of colorado continuous speech recognizer. Rep. tr-cslr-2001-01, University of Colorado, Boulder, Tech. (2001)
70. H Schmid, in *Proc. Int'l. Conf. New Methods in Language Processing*. Probabilistic part-of-speech tagging using decision trees, (1994)
71. P Bojanowski, R Lajugie, F Bach, I Laptev, J Ponce, C Schmid, J Sivic, in *Proc. IEEE European Conference on Computer Vision*. Weakly supervised action labeling in videos under ordering constraints, (2014)
72. P Wittenburg, H Brugman, A Russel, A Klassmann, H Sloetjes, in *Proc. 5th Int'l. Conf. on Language Resources and Evaluation*. ELAN: a professional framework for multimodality research, (2006)
73. N Malandrakis, A Potamianos, G Evangelopoulos, A Zlatintsi, in *Proc. Int'l. Conf. on Acoustics, Speech and Signal Process*. A supervised approach to movie emotion tracking, (2011), pp. 2376–2379
74. R Dietz, A Lang, in *Proc. Cognitive Technology Conf*. Affective agents: effects of agent affect on arousal, attention, liking and learning, (1999)
75. A Hanjalic, Extracting moods from pictures and sounds: Towards truly personalized TV. IEEE Signal Process. Mag. **23**(2), 90–100 (2006). IEEE
76. HL Wang, LF Cheong, Affective understanding in film. IEEE Trans. Circ. Syst. Video Technol. **16**(6), 689–704 (2006)
77. R Cowie, E Douglas-Cowie, S Savvidou, E McMahon, M Sawey, M Schröder, in *Proc. ISCA Workshop on Speech & Emotion*. 'FEELTRACE': an instrument for recording perceived emotion in real time, (2000), pp. 19–24
78. M Grimm, K Kroschel, in *Proc. IEEE Workshop on Automatic Speech Recognition and Understanding*. Evaluation of natural emotions using self assessment manikins, (2005), pp. 381–385
79. P Koutras, P Maragos, A perceptually based spatio-temporal computational framework for visual saliency estimation. Signal Process. Image Commun. **38**, 15–31 (2015)
80. L Itti, C Koch, E Niebur, A model of saliency-based visual attention for rapid scene analysis. IEEE Trans. Pattern Anal. Mach. Intell. **20**(11), 1254–1259 (1998)
81. S Gao, K Yang, C Li, Y Li, in *Proceedings of the IEEE international conference on computer vision*. A color constancy model with double-opponency mechanisms, (2013), pp. 929–936
82. DJ Heeger, Model for the extraction of image flow. J. Opt. Soc. Amer. **4**(8), 1455–1471 (1987)
83. JP Havlicek, DS Harding, AC Bovik, Multidimensional quasi-eigenfunction approximations and multicomponent am-fm models. IEEE Trans. Image Process. **9**(2), 227–242 (2000)
84. AC Bovik, N Gopal, T Emmoth, A Restrepo, Localized measurement of emergent image frequencies by Gabor Wavelets. IEEE Trans. Inf. Theory. **38**, 691–712 (1992)
85. JB Fritz, M Elhilali, SV David, SA Shamma, Auditory attention—focusing the searchlight on sound. Curr. Opin. Neurobiol. **17**(4), 437–455 (2007)
86. JF Kaiser, in *Proc. IEEE Int'l. Conf. Acoust., Speech, Signal Process*. On a simple algorithm to calculate the energy of a signal, (1990)
87. E Zwicker, H Fastl, *Psychoacoustics, Facts and Models*, 2nd edn. (Springer, Berlin Heidelberg, 1999)
88. R Plomp, WJM Levelt, Tonal consonance and critical bandwidth. Jour. Acoust. Soc. Am. (JASA). **38**, 548–560 (1965)
89. PN Vassilakis, Perceptual and physical properties of amplitude fluctuation and their musical significance. PhD thesis, Univ. of California (2001)
90. P Nakov, S Rosenthal, Z Kozareva, V Stoyanov, A Ritter, T Wilson, in *Proc. of 2nd Joint Conf. on Lexical and Computational Semantics (*SEM), 7th Int'l. Workshop on Semantic Evaluation*. Semeval 2013 task 2: Sentiment analysis in twitter, (2013), pp. 312–320

91. CM Lee, SS Narayanan, Toward detecting emotions in spoken dialogs. IEEE Trans. Speech Audio Process. **13**(2), 293–303 (2005)
92. N Malandrakis, A Potamianos, E Iosif, S Narayanan, Distributional semantic models for affective text analysis. IEEE Trans. Audio Speech Lang. Process. **21**(11), 2379–92 (2013)
93. Z Harris, Distributional structure. Word. **10**(23), 146–162 (1954)
94. M Bradley, 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)
95. A Zlatintsi, P Maragos, A Potamianos, G Evangelopoulos, in *Proc. European Signal Process. Conf.* A saliency-based approach to audio event detection and summarization, (2012), pp. 1294–1298
96. P Maragos. 2nd edn., in *The Image and Video Processing Handbook*, ed. by AC Bovik. Morphological filtering for image enhancement and feature detection (Academic Press, Inc, Orlando, 2005), pp. 135–156
97. A Zlatintsi, P Koutras, N Efthymiou, P Maragos, A Potamianos, K Pastra, in *Proc. 7th Int'l. Workshop on Quality of Multimedia Experience (QoMEX-2015), Costa Navarino, Messinia, Greece.* Quality evaluation of computational models for movie summarization, (2015), pp. 1–6