Affect in Multimedia: Benchmarking Violent Scenes Detection

Mihai Gabriel Constantin, Liviu-Daniel Ştefan, Bogdan Ionescu, Claire-Hélène Demarty, Mats Sjöberg, Markus Schedl, Guillaume Gravier

Abstract—In this paper, we report on the creation of a publicly available, common evaluation framework for Violent Scenes Detection (VSD) in Hollywood and YouTube videos. We propose a robust data set, the VSD96, with more than 96 hours of video of various genres, annotations at different levels of detail (e.g., shot-level, segment-level), annotations of mid-level concepts (e.g., blood, fire), various pre-computed multi-modal descriptors, and over 230 system output results as baselines. This is the most comprehensive data set available to this date tailored to the VSD task and was extensively validated during the MediaEval benchmarking campaigns. Furthermore, we provide an in-depth analysis of the crucial components of VSD algorithms, by reviewing the capabilities and the evolution of existing systems (e.g., overall trends and outliers, the influence of the employed features and fusion techniques, the influence of deep learning approaches). Finally, we discuss the possibility of going beyond state-of-the-art performance via an ad-hoc late fusion approach. Experimentation is carried out on the VSD96 data. We provide the most important lessons learned and gained insights. The increasing number of publications using the VSD96 data underline the importance of the topic. The presented and published resources are a practitioner's guide and also a strong baseline to overcome, which will help researchers for the coming years in analyzing aspects of audio-visual affect and violence detection in movies and videos.

Index Terms—violent scenes detection, multi-modal content description, VSD96 data set, benchmarking, literature review.

1 Introduction

Multimedia content analysis has a long history of concept detection in videos. In most cases, tangible concepts, e.g., "plane", "car", "fire", "rocket launch", "handshake", "hug", are targeted for several practical reasons, the most significant being the possibility to annotate data in an almost unambiguous manner, following strict annotation guidelines. The TRECVid evaluation series [1], [2], [3] is an emblematic illustration where concepts are defined mostly based on events and actions that can be identified by humans. In contrast, computing affect induced on viewers by videos do not follow the same path [4]. Emotions can hardly be annotated beforehand, being highly dependent on the viewer. Obviously, the same media material can trigger distinct feelings in two persons. In addition, the relation between cues in the videos and the viewers' feelings and experience is far less direct than in the case of tangible concepts. As a consequence, designing algorithms to predict the user's reactions to a video remains a mostly unsolved and highly challenging problem. A workaround is to rely on the *mid-level concepts* that, we expect, are strongly correlated with the viewers' feelings, and have cues in the video signal. These concepts are often not as clearly defined as classical tangible concepts, and they do not rely on measuring human perception.

Violence in movies and video materials, as dealt with in this paper, provides a perfect illustration of a mid-level concept. On the one hand, violent scenes in movies are obviously impactful on the emotional state of the viewer. On the other hand, there is an amount of evidence related to violence, e.g., "gunshots", "screams", and " explosions", with explicit cues in the video. In between the emotional state and the tangible cues there is the perception of violence, i.e., how violently the viewer perceives the movie. As for emotions and unlike tangible concepts, the perception of violence is highly dependent on the viewer. For instance, children are more likely than adults to be strongly impacted by violent scenes [5]. Yet, the perception of violence is more natural (though not easy) to annotate than affect.

In this sense, *violence detection* in movies provides an achievable basis for developing previewing tools targeting parental guidance, e.g., for video-on-demand (VOD) services. Traditional parental guidance recommendations issued by national agencies, in the countries where they exist, are defined globally on the movie and are highly dependent on time and culture. As discussed by Demarty et al. [6], these policies are poorly adapted to today's Internet diffusion of movies, where movies can be viewed only partially and where cultural and geographical frontiers are blurred. More fundamentally, global policies do not reflect well the fact that the notion of offending material is primarily a personal and cultural matter. Hence, the ultimate idea of providing previewing services targeting several mid-level concepts related to affect, such as violent or sexually explicit scenes, used for the fast, interactive selection of appropriate video material, is more than needed.

Similar to any concept involving human perception, the most difficult issue when studying violence detection in movies is to establish a *reference of what is considered as violent*, both for evaluation purposes and machine learning tasks. Due to the multiple facets of violence, no common and generic enough definition for violent events has ever been proposed, even when restricting ourselves to physical violence. The World Health Organization defines violence as "*the intentional use of physical force or power, threatened*

Manuscript received January 7, 2019; final revision January 31, 2020.

or actual, against oneself, another person, or against a group or community, that either results in or has a high likelihood of resulting in injury, death, psychological harm, maldevelopment, or deprivation" [7]. This definition is sound but too broad to be used as guidelines for annotation. More focused definitions are provided with systems designed to detect violence in various scenarios, e.g., "*a series of human actions accompanied with bleeding*" [8], "*scenes containing fights, regardless of context and number of people involved*" [9], "*behavior by persons against persons that intentionally threatens, attempts, or actually inflicts physical harm*" [10], "*fast paced scenes which contain explosions, gunshots and person-on-person fighting*" [11]. Inherently, these definitions are highly correlated to intentional human actions and physical contact, but they overlook some relevant situations like accidents, which may result in disturbing content, or verbal violence.

In this paper, we report on the creation of a publicly available, *common evaluation framework* for Violent Scenes Detection (VSD). We focus, in particular, on the techniques for automatic violence detection in Hollywood movies and YouTube videos. These resources have been developed under the framework of the MediaEval benchmarking initiative for multimedia evaluation[1]. To deal with the diversity of violent content, we approach the definition of violence from two perspectives: (i) an *objective* formulation of violence defined as "*physical violence or accident resulting in human injury or pain*", that implies the segments annotated as violent must contain both the violent actions and the results of these actions; (ii) a *subjective* formulation of violence: something "*one would not let an 8-year old child see in a movie because it contains physical violence*", with the objective to account for a broader use case scenario.

The remainder of the article is organized as follows. Section 2 presents an overview of the existing literature and positions our contributions. Section 3 introduces the released, publicly available data sets and discusses their annotations. Section 4 proposes the evaluation methodology. Section 5 deals with the experimental validation of the data: an overview of the MediaEval benchmarking results, and comparison to the state-of-the-art methods from the literature. Section 6 introduces the perspective that goes beyond the individual systems, proposing the design of a super system for violence detection. Section 7 concludes the paper and provides some general perspectives.

2 Previous Work

We focus our review of the literature on the existing initiatives for creating annotated data sets for VSD. For a thorough analysis of the existing violence detection techniques, we refer the reader to Section 5.

A wide variety of human actions are recorded in many action recognition data sets, and a few of these have some particular actions that are considered violent. For example, the data sets of Weinland et al. [12] and Liu et al. [13] contain scenes depicting punching and kicking. In contrast, others have a broader range of violent human interactions: some samples of crowd fighting and behavior in the BEHAVE data set [14], instances of two people fighting in the CAVIAR data set[2], instances of boxing, punching and sumo wrestling in the UCF101 data set [15], or car crashes and explosions in the TRECVid data sets [2]. However, these data sets were not specifically designed for violence detection, and therefore sometimes only small parts of them are suited for developing violence detection algorithms, which fall short of covering a broad spectrum of violent actions.

In more recent developments, the data set proposed by Fu et al. [16] is composed of YouTube videos depicting real-world surveillance scenarios that contain fighting scenes in different environments, for example, bar fights, street fights, prison fights. These clips are annotated as fight or non-fight at a sub-clip level. The individual video clips are 10 seconds long on average, and 119 clips are non-violent, while 147 clips contain fighting scenes. Marsden et al. [17] developed the Multi Task Crowd data set consisting of 100 individual images, captured in different settings, serving 3 purposes: (i) crowd counting, (ii) crowd density estimation, and finally, (iii) crowd violent behavior detection. The subset of 50 violent images was selected from the WWW Crowd data set [18], representing images extracted from videos that are tagged as violent or mob, while the non-violent subset consists of the other cases.

As one can observe, the existing attempts to provide a common evaluation framework are either tangential to the main purpose, e.g., action recognition which also uses violent actions, or too narrow to address a more general goal for the detection of the violent scenes, e.g., only a few violent situations such as fights, accidents, etc. The size of the data also tends to be very restrained, therefore providing reduced capabilities for training data greedy systems employing, for instance, deep learning. There is also a limitation in what concerns the target detection, most of the data being annotated at sequence level (short sequences). There are too few approaches providing annotations for unconstrained segments, i.e., variable length segments, which are, in fact, closer to real-world scenarios.

In this paper, we propose a *comprehensive evaluation framework* that is robust, both in terms of the size and variability of the corpus, as well as in terms of the definition of violence employed for the annotations. These publicly available data were conceived and validated during the yearly benchmark campaign run at MediaEval between 2011 and 2015. Although the data was released a few years ago, we stress on its current importance and impact in the community. In this respect, Figure 1 provides some statistics about the interest in violent related detection tasks worldwide. It is quantified in terms of the number of published research papers, which we identified using Google Scholar and Clarivate Analytics Web of Science[3], searching for "violence detection", "violent events", and "violent scenes", and retaining the articles that use these terms in their body text. Although the search is not exhaustive, i.e., there are other keywords covering this topic, such as "violence concepts", it is a good indicator of the trend. We have also added to the plot the number of times external users requested the VSD96 data.

As can be seen from the values in 2009 and 2010, violence

1. http://www.multimediaeval.org/
2. http://groups.inf.ed.ac.uk/vision/CAVIAR/CAVIARDATA1/
3. https://clarivate.com/products/web-of-science/

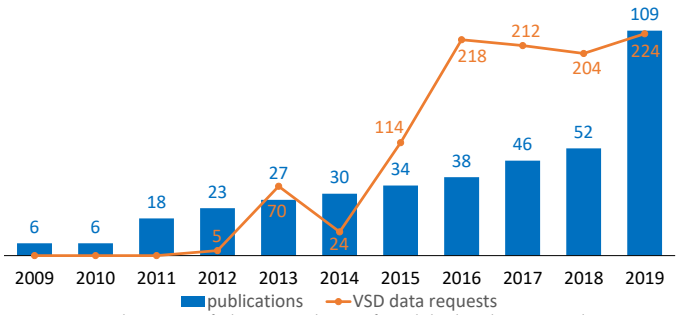

Fig. 1: Evolution of the number of published research papers related to violence detection and of the number of VSD96 data download requests (search made via Google Scholar and Clarivate Analytics Web of Science using "violence detection", "violent scenes", and "violent events").

detection in movies has received very little attention prior to the establishment of our VSD benchmark data in 2011. Hereafter, in contrast, the number of publications shows a steady upward trend. The same trend is observed for the number of downloading requests, which reached more than 200 starting with 2016.

In this context, the added contributions of this work can be summarized as follows: (i) We introduce a publicly available, *common evaluation framework* for VSD with more than 96 hours of video and a high variability of content, e.g., various genres, both Hollywood and YouTube videos, annotations for the objective and subjective definitions of violence as well as different levels of granularity (shot-level, segment-level, and clip-level), annotations of mid-level concepts (e.g., "blood", "screams"), various pre-computed multimodal content descriptors and various system output results as baselines. This is the most comprehensive VSD data set available to this date; (ii) We provide an *in-depth analysis* of the crucial aspects of VSD algorithms, by investigating the capabilities and evolution of existing systems (analysis of relevant approaches from the MediaEval benchmark and from the literature, influence of the employed features and fusion techniques, influence of deep learning approaches). This is again the first comprehensive study covering all these core aspects. It is a practitioner's guide for best practice in this field and also a strong baseline to overcome; (iii) We introduce a *super system design* based on late fusion to discuss the possibility of going beyond the state-of-the-art performance by combining existing systems. The provided data and evaluation resources will further help researchers to analyze these aspects, and push forward the field.

Relation to previous work. The reader should note that some preliminary contributions to the points above have been already published by the authors and can refer to those for more detailed information: Demarty et al. [19], an overview of the 2011 MediaEval campaign and submitted systems; Demarty et al. [6], a detailed comparison between the 2011 and 2012 MediaEval campaigns; Demarty et al. [20], an overview of the 2013 MediaEval campaign and submitted systems; Demarty et al. [21], a state of the art in multimodal violence detection, an overview of the 2011 and 2012 MediaEval campaigns and description of two of the top systems submitted to the task; Schedl et al. [22], release of the 2014 MediaEval data together with the baselines.

Abbreviations. Throughout the paper, we employ the following abbreviations: BER — band energy ratio, BoW — bag-of-words, CM — color moments, CNN — Convolutional Neural Networks, CNH — color naming histogram, DoG — Difference of Gaussian, DT — dense trajectories, EoH — Edge Orientation Histograms, FBW — frequency bandwidth, GMM — Gaussian mixture models, HMM — hidden Markov models, HoF — histograms of optical flow, HoG — histograms of oriented gradients, kNN — k-Nearest Neighbours, LBP — local binary patterns, LDA — Linear Discriminant Analysis, LPC — linear predictive coefficients, LSF — line spectral frequency, LSP — line spectral pairs, MBH — motion boundary histograms, MFCC — mel-frequency cepstral coefficients, MKL — multiple kernel learning, OBSI — octave band signal intensity, PCA — principal component analysis, QDA — Quadratic Discriminant Analysis, RMS — root-mean-square energy, SC — spectral centroid, SE — spectral entropy, SF — spectral flux, SIFT — scale invariant feature transform, STIP — space-time interest points, SVM — support vector machines, ZCR — zero-crossing rate, MLP — multi-layer perceptron.

3 Proposed Data Set

We present the proposed VSD data and their evolution over the years. The section is organized as follows: (i) we discuss the composition of the data and its evolution through years (see Section 3.1), (ii) we discuss the annotation protocol and the quality of the labels (see Section 3.2), (iii) we present the provided, precomputed, audio and visual features (see Section 3.3). Please note that all annotations, metadata, audio and video features, and where possible the video footage, are available for download[4,5].

3.1 Composition and evolution

As VSD is mainly intended to be a machine learning task, all VSD96 data are split into a training set (*Dev.*) for the development stage (with released annotations for the training and validation data), and a test set (*Test*) for the evaluation of the participating systems. The former is used to elaborate algorithms and allow classifiers to learn, and the latter is used for the actual evaluation.

To provide a sufficient amount of training data for both classes (violent and non-violent), we selected *popular Hollywood movies* that span a wide range of genres, subjects and amount of violence, ranging from very violent ones, e.g., "Saving Private Ryan" with $34\,\%$ violent content, to movies with no violence at all, e.g., "Legally Blond". One should note that due to copyright issues, the Hollywood videos are not provided with the data set. However, they are popular movies that can be easily purchased.

Accounting for the rising importance of user-generated content shared on the web, we additionally provide *YouTube videos*. These data allow us to assess the generalization capabilities of algorithms to different types of footage, in particular to those characterized by low video and audio

4. VSD data sets for 2011, 2012, 2013 and 2014 are available for download here: https://www.interdigital.com/data_sets/violent-scenes-dataset.

5. VSD data set for 2015 is available for download here: http://liris-accede.ec-lyon.fr/ (look for MediaEval 2015 data).

TABLE 1: Basic statistics of the VSD96 data. Columns indicate usage over years, movie name or source, total playtime in minutes (Dur.), total number of shots, segments, or clips (#Segm.), amount of violence in percent (V), and average duration of a violent scene in seconds. The latter two are reported for the subjective definition of violence. The figures in brackets indicate the number of movies in the corresponding category for each year, where ^{o} stands for use of the objective definition of violence, ^{s} for the subjective definition, and Gen. stands for generalization task.

2015	*2014*	*2013*	*2012*	*2011*	*Movie*	*Dur.* (m)	*#Segm.*	*V* (%)	$\overline{Dur.}$ (s)
	Dev. (24^s)	Dev. ($18^{o,s}$)	Dev. (15^o)	Dev. (12^o)	Armageddon	145	3,562	7.78	25.01
					Billy Elliot	106	1,236	2.46	8.68
					Eragon	100	1,663	13.26	39.69
					Harry Potter 5	133	1,891	5.44	17.30
					I am Legend	96	1,547	15.64	75.36
					Leon	106	1,547	16.36	41.52
					Midnight Express	116	1,677	7.12	24.80
					Pirates of Caribbean	137	2,534	18.15	49.85
					Reservoir Dogs	95	856	30.41	115.82
					Saving Private Ryan	162	2,494	33.95	367.92
					The Sixth Sense	103	963	2.00	12.40
					The Wicker Man	98	1,638	6.44	31.55
				Test (3^o)	Kill Bill	106	1,597	17.47	23.98
					The Bourne Identity	114	1,995	7.18	27.21
					The Wizard of Oz	98	908	1.02	8.56
			Test (3^o)		Dead Poets Society	124	1,583	0.58	14.44
					Fight Club	133	2,335	15.83	32.51
					Independence Day	147	2,652	13.13	68.23
		Test ($7^{o,s}$)			Fantastic Four 1	102	2,002	20.53	62.57
					Fargo	94	1,061	15.04	65.32
					Forrest Gump	136	1,418	8.29	75.33
					Legally Blond	92	1,340	0.00	0.00
					Pulp Fiction	148	1,686	25.05	202.43
					The God Father	170	1,893	5.73	44.99
					The Pianist	143	1,845	15.44	69.64
	Test (7^s)				8 Mile	106	17	4.70	37.40
					Braveheart	170	87	21.45	51.01
					Desperado	100	35	31.94	113.00
					Ghost in the Shell	83	23	9.85	44.47
					Jumanji	100	29	6.75	28.90
					Terminator 2	147	83	24.89	53.62
					V for Vendetta	127	85	14.27	25.91
	Test (86^s Gen.)				YouTube videos	157	86	44.47	109.75
Dev. (100^s)					Hollywood-like movie clips	1,014	6,144	4.42	9.90
Test (99^s)					Hollywood-like movie clips	784	4,756	4.90	9.88
						5,792			

quality (e.g., very low resolution, high compression, noisy audio), as well as short duration. To compose this set, we defined a number of target queries, such as "brutal accident" or "killing video games". We only considered the videos shared under Creative Commons[6] Attribution 3.0 Unported license allowing redistribution. We then selected an approximately uniform number of violent and non-violent videos, for which we additionally retrieved metadata offered by YouTube (e.g., publishing date, category, title, number of likes/dislikes).

The final part of the collection is also addressing user-generated data and represents an extension of the Discrete LIRIS-ACCEDE data set [23]. LIRIS-ACCEDE was originally annotated for its emotional content on the valence-arousal dimensions. Given the related use case, we imported some of the data which was re-annotated for violence. It is composed of short clips extracted from Creative Commons *Hollywood-like movies* (i.e., movies replicating Hollywood's specific composition and editing style) of various genres: "action", "adventure", "animation", "comedy", "documentary", "drama", "horror", "romance", and "thriller".

Table 1 presents a global overview of the VSD96 data together with some basic statistics. Overall there are 31 annotated full movies, 86 annotated YouTube short videos, and 10,900 annotated clips extracted from 199 movies. The total duration of the video data is over 96 hours.

6. https://creativecommons.org/

3.2 Annotations

To deal with the diversity of the 'violence' content, we approach the definition of violence from two perspectives: (i) we provide an *objective* formulation of violence, i.e., "*physical violence or accident resulting in human injury or pain*", and (ii) a *subjective* formulation of violence, i.e., something "*one would not let an 8-year old child see in a movie because it contains physical violence*". For more details, we refer the reader to our previous publications [6], [22].

3.2.1 Annotation protocol

The 2011, 2012 and 2013 collections provide video shot segmentation, obtained via automatic shot boundary detection. Therefore, the localization of violence is first at *shot level*. For the 2012, 2013 and 2014 data, the localization of violence is also provided at *frame level*, allowing variable length segment detection. For the 2015 collection, the localization of violence is provided at *clip level*.

On general principle, the violence annotations consisted of marking the violent shots, segments, or clips, using both

TABLE 2: Annotation statistics for concepts: duration represented as percentage/Fleiss' κ agreement between the concept and the overall violence annotation.

Movies	*violence (%)*	*blood*	*car chase*	*cold arms*	*explosions*	*fights*	*fire*	*fire arms*	*gore*	*gun shots*	*screams*
Armageddon	9.33	0.8/-0.03	0.2/-0.04	0.0/-0.04	5.8/0.41	2.9/0.19	9.3/0.40	3.9/0.02	0.0/-0.04	0.4/0.03	4.5/0.15
Billy Elliot	4.77	0.2/0.02	0.0/-0.01	1.8/-0.02	0.0/-0.01	1.9/0.32	1.0/-0.02	0.0/-0.01	0.0/-0.01	0.0/-0.01	5.0/0.22
Eragon	16.04	5.1/-0.06	0.0/-0.07	13.4/0.16	0.4/-0.01	10.5/0.85	21.2/0.18	0.0/-0.07	2.2/0.11	0.0/-0.07	6.4/0.29
Harry Potter 5	8.93	4.4/0.09	0.0/-0.03	2.4/-0.04	1.8/0.22	4.8/0.58	14.2/0.01	0.0/-0.03	0.0/-0.03	0.0/-0.03	2.9/0.21
I am Legend	17.71	8.8/0.12	1.1/-0.09	3.3/-0.05	0.5/-0.04	5.6/0.38	2.0/0.11	12.9/0.01	9.4/0.33	0.7/-0.01	8.2/0.44
Leon	18.51	12.0/0.07	0.0/-0.09	1.7/-0.09	0.2/-0.06	3.4/0.26	0.8/-0.03	20.2/0.33	0.0/-0.09	1.4/0.07	1.2/-0.02
Midnight Express	8.17	2.0/0.21	0.0/-0.04	0.4/-0.01	0.0/-0.04	5.1/0.67	3.6/-0.06	6.7/0.10	0.1/-0.01	0.2/-0.04	10.4/0.20
Pirates Carib. 1	19.89	0.6/-0.05	0.0/-0.10	25.5/0.09	0.8/-0.03	9.4/0.44	17.9/0.05	20.0/0.03	4.8/0.21	2.1/0.03	10.4/0.36
Reservoir Dogs	34.37	36.8/0.59	0.0/-0.18	1.9/-0.13	0.0/-0.18	4.1/0.07	0.2/-0.18	19.1/0.23	21.6/0.62	0.8/-0.13	4.7/0.05
Saving Private Ryan	34.54	21.5/0.11	0.0/-0.20	18.7/-0.07	12.9/0.04	10.8/0.28	11.7/-0.06	53.9/0.08	8.1/0.13	26.0/0.22	9.4/0.06
The Bourne Identity	9.50	3.3/0.31	2.9/0.06	2.3/0.08	0.1/-0.01	2.6/0.40	0.4/0.00	6.1/0.20	0.0/-0.04	0.4/0.05	2.3/0.10
The Sixth Sense	2.49	1.0/0.23	0.0/-0.01	4.3/-0.02	0.0/-0.01	0.1/0.03	1.9/-0.02	0.8/0.11	0.1/0.09	0.0/0.01	1.4/0.11
The Wicker Man	11.74	0.7/-0.03	0.0/-0.03	1.8/-0.03	0.2/0.03	0.4/0.09	4.8/0.31	6.2/-0.00	0.0/-0.03	0.3/-0.03	7.1/0.28
The Wizard of Oz	1.14	0.0/-0.01	0.0/-0.01	33.3/-0.16	1.1/-0.00	1.2/0.20	6.7/0.14	7.4/-0.04	0.0/-0.01	0.0/-0.01	5.1/0.08
Dead Poets Society	0.72	0.3/0.14	0.0/-0.00	0.8/0.00	0.0/-0.00	0.3/0.42	3.3/-0.02	0.5/-0.01	0.0/-0.00	0.0/-0.00	3.3/0.10
Fight Club	19.24	7.8/0.20	0.0/-0.09	1.1/-0.04	0.2/-0.09	4.4/0.34	2.6/-0.06	5.1/0.06	0.4/-0.04	0.1/-0.08	5.0/0.24
Independence Day	14.62	0.3/-0.06	0.0/-0.07	0.6/-0.07	2.6/0.20	4.9/0.44	7.9/0.34	5.5/0.17	0.0/-0.07	1.4/0.11	4.4/0.12

TABLE 3: Annotation statistics for the evaluation data (subjective definition of violence): the amount of violence as annotated by each assessor, overlap percentage, final amount of violence, and inter-annotator assessment.

Videos	*Annotator1 dur. (%)*	*Annotator2 dur. (%)*	*An.1&An.2 Overlap dur. (%)*	*Final dur. (%)*	*Percent agreement*	*Fleiss' κ*	*Randolph's κ*
8 Mile	7.27	8.70	3.97	4.70	0.9197	0.4538	0.8394
Brave Heart	15.72	20.25	11.44	21.45	0.8692	0.5565	0.7383
Desperado	24.34	28.96	17.55	31.94	0.8180	0.5346	0.6361
Ghost in the Shell	11.82	15.45	6.99	9.85	0.8672	0.4360	0.7343
Jumanji	7.95	8.41	4.37	6.72	0.9238	0.4929	0.8476
Terminator 2	26.35	27.92	20.70	24.89	0.8713	0.6744	0.7425
V for Vendetta	8.83	14.33	8.12	14.27	0.9309	0.6623	0.8617
average values	*14.61*	*17.72*	*10.45*	*16.26*	*0.8857*	*0.5444*	*0.7714*
YouTube videos (2014)	37.80	28.11	21.18	44.47	0.7646	0.4672	0.5291
Hollywood-like movie clips (2015)	3.55	8.18	3.45	4.90	0.8931	0.7095	0.9369

visual and audio information. The annotations are binary. They were carried out by three expert annotator groups. Specifically, two groups, first, conducted all the annotations independently. No discussions were held between the two groups during this stage. Then, the third master annotator group merged the two sets of annotations and made decisions for the inconsistent cases. The subjective formulation of violence, in particular, required panel discussions for borderline cases, held by people with different cultural backgrounds. Annotators provided a text description for each segment, motivating their choice. This helped the final decision of the panel.

In terms of demographics, the annotators were, across all the iterations, moderately to highly educated, with age spans varying from 25 to 50 years and based in Europe and Asia. The initial annotators were mostly graduate students, with no children, while the master annotators were senior researchers, married with children. The number of annotators varied for each year, starting with 7 in 2011, 9 in 2012, 25 in 2013, 11 in 2014, and finally 17 in 2015.

Overall, the consistency of the annotations is ensured via: (i) using expert annotators instead of crowd-workers, (ii) providing clear definitions and use cases for violence, (iii) ensuring at least two annotations per each sample, (iv) employing master annotators and panel discussions for settling inconsistencies and borderline cases, (v) selecting the annotators from different countries and continents, to reduce cultural bias.

Apart from the binary violence annotation, part of the data is provided with annotations for several concepts that are related to violence. The idea is to enforce the learning by providing mid-level concept annotations. The 17 movies presented in Table 2 are annotated for 7 visual concepts: "blood", "car chase", "cold arms", "fights", "fire", "firearms", "gore", and 3 audio concepts: "explosions", "gunshots", "screams".

3.2.2 Quality assessment

To better understand the quality of the annotations, we further investigate the consistency of the labels using three indicators: (i) the percent agreement, (ii) Fleiss' kappa [24], and (iii) Randolph's kappa [25]. Percent agreement is a widely used criterion, as stated by McHugh [26], which calculates the degree to which two or more independent annotators agree on what they observe when watching the same events. This measure does not adjust for the agreement expected by chance. Fleiss' kappa takes into account the agreement occurring by chance, therefore it is more robust than the simple percent agreement. However, Fleiss' kappa is a fixed-marginal rater, that assumes that the distribution of labels (in our case violent vs. non-violent) is known in advance. This shortcoming would be accentuated in our binary annotation since the duration of non-violent scenes is much longer than that of violent scenes, and annotators have the liberty of annotating as many shots, segments or clips as violent as they see fit. To overcome this, Brennan and Prediger [27] suggest using a free-marginal rater. Therefore, we also apply Randolph's multirater kappa, which is a free-marginal rater and does not limit the distribution of labels. Values range from 0 to 1 for percent agreement, and from -1

to 1 for the other two. For the kappa, the positive/negative values indicate the agreement/disagreement between annotators. Larger values indicate higher reliability or consistency. The resulted statistics are presented in Table 3. For brevity reasons, we provide here only the numbers for the evaluation data. They extrapolate to the remainder of the data.

For the 2014 test data, the two annotator groups provided labels accounting for similar percentages of violence, i.e., 14.61 %, and 17.72 %, respectively. For the 2014 YouTube data, the percentages are different: 37.80 % vs. 28.11 %, accounting for a higher variability of content. The overlap percentage between the annotations is 10.45 % for the movies and 21.18 % for the YouTube videos. For the 2015 test data, the two annotator groups yielded a percentage of violence of 3.55 %, and 8.18 %, respectively. This shows again the higher variability of clip level annotations. The overlapping between annotations is 3.45 %.

As for what concerns the agreement, for the 2014 test data, we obtain a percent agreement of 88.57 % for movies and 76.46 % for YouTube videos. The average Fleiss' kappa and Randolph's kappa are 0.5444 and 0.7714, respectively, for movies and 0.4672 and 0.5291 for the YouTube videos. The values show that the annotators were fairly reliable with consistent annotations, despite the high subjectivity of the task (according to Landis et al. [28], a score between 0.41 and 0.60 means moderate agreement, while a score between 0.61 to 0.80 means substantial agreement). The agreement values for the 2015 test data are 89.31% percent agreement, Fleiss' kappa 0.7095, and Randolph's kappa 0.9369. These are again good indicators for reliable annotations.

As regards the agreement for the annotated concepts, Fleiss' kappa is presented in Table 2. The agreement is computed between each concept and the overall violence annotation showing the correlation between the two. Although some of the concepts are weakly correlated to the final violence decision, most of them are useful clues (see Section 5.1.3).

3.3 Audio and visual features

The data comes with several pre-computed common audio and visual descriptors, to address a broader community.

Audio descriptors. For each video frame (of 40 ms length at 25 fps), we provide the following descriptors: amplitude envelop, RMS, ZCR, BER, SC, FBW, SF, and MFCC. While the first three features describe the audio signal in the time domain (time vs. amplitude representation), the remaining ones are computed in the frequency domain (frequency vs. magnitude representation). As the audio exhibits a sampling rate of 44,100 Hz, and the videos are encoded with 25 fps, we consider windows of 1,764 audio samples in length. We compute 22 MFCCs for each window, while all the other features are 1-dimensional. For a detailed discussion of the audio features, please refer to [29].

Visual descriptors. For the visual information we provide: CNH [30], CM [31], LBP [32] and HoG [33]. The CNH features are 99-dimensional, the CM and HoG features 81-dimensional, and the LBP 144-dimensional. The CNHs are computed on 3-by-3 image regions and map colors to 11 universal color names: "black", "blue", "brown", "gray", "green", "orange", "pink", "purple", "red", "white", and "yellow". The global CM in the hue-saturation-value (HSV) color space (9 values) contains the first three central moments of an image color distribution: mean, standard deviation, and skewness, which are computed on 3-by-3 image regions. Also, the global LBP (16 values) and the global HoG are computed using a 3-by-3 spatial division. The global HoG contains the average of the HoG features (9 values) that exploit the local object appearance and shape within an image via the distribution of edge orientations.

4 Evaluation Methodology

To benchmark violent scenes detection, as done in the MediaEval campaigns, one has to produce a violence prediction for the provided test set data. The systems should output a violence judgment for each movie segment indicating whether it represents a violent or non-violent scene. The judgment could be accompanied by a confidence score or probability estimate (a higher value indicates a higher probability that the segment is violent; typical values are between 0 and 1). For the 2011 to 2013 collections, the prediction is at the video shot level. For the 2012 and 2013 collections, the prediction is also at arbitrary length segment level. For the 2014 collection, the prediction is only at segment level, while for the 2015 collection, the prediction is at clip level.

To assess performance, we provide a number of metrics that were validated during the MediaEval 2011-2015 benchmarks. However, the data is not limited to a specific metric, and other measures can be implemented as well.

MediaEval cost. With the 2011 data we introduced what we called the "MediaEval cost" which is a cost function weighing false alarms (FA, or false positives) and missed detections (MI, or false negatives). It is defined as: $C = C_{\text{FA}} \cdot P_{\text{FA}} + C_{\text{MI}} \cdot P_{\text{MI}}$, where $C_{\text{FA}} = 1$ and $C_{\text{MI}} = 10$ are selected to reflect the higher cost of missing a violent scene than making a false positive judgement [19]. The idea is that when protecting children from seeing violence, it is worse to miss a violent scene than to erroneously mark a non-violent scene as violent. P_{FA} and P_{MI} are, respectively, the FA and MI rates of the system's output. For the segment-based runs, FA and MI are computed on a per second basis.

Mean Average Precision. The average precision has been shown to be a stable and discriminating measure for general-purpose retrieval tasks [34]. MAP is calculated by taking the arithmetic mean over the uninterpolated average precision scores for all the test set movies [35]. We use two MAP-based metrics: (i) mean average precision over the 100 highest ranked shots (MAP@100); (ii) the full MAP. For the free segment annotations, the ground truth segments are of variable length and have no a priori positions in the video like the video shots. Therefore, computing MAP needs to handle two special situations properly: (i) a single wide hypothesis segment (system output) matching several reference segments (ground truth), and (ii) several small hypothesis segments matching a single reference one. The first case is taken care of by counting only a single correct match per hypothesis segment and using the number of ground truth segments as the divisor for calculating the

average precision. The second case is covered by counting a maximum of one match per ground truth segment by picking the one with the highest confidence score. We avoid a complex combinatorial problem by keeping track of a single "best" ground truth match for each hypothesis segment. "Best" is here defined as the largest overlap percentage. The other possible matches are discarded, i.e., they are not counted as false positives either. We called this adaptation MAP2014. This is an alternative to segment matching with the Hungarian method [36], i.e., finding the optimal solution to the bipartition graph matching problem, where the ground truth and hypothesis segments are seen as the two sets of nodes. This algorithm is slower as it has a cubic running time compared with MAP2014, which iterates through all the segments once and match them with the largest overlapping one. We have decided to remain with the MAP approach for consistency reasons.

Other metrics. To facilitate further studies and comparisons to previous years and other benchmarks, we also compute FA and MI rates, precision and recall, and the detection error trade-off (DET) curve. The DET curve is formed by plotting P_{FA} as a function of P_{MI} given a confidence score for each segment [37].

All the evaluation measures are implemented in the `trec_eval` tool[7]. The tool is available with the data.

5 Experimental Results

To serve as a baseline and reference for future developments, we analyze the performance of various systems tested on the VSD96 data, namely: (i) systems submitted to the MediaEval benchmark (see Section 5.1), and (ii) state-of-the-art approaches reported in the literature (see Section 5.2). The objective is to investigate the crucial aspects such as the capabilities and the evolution of the systems, the employed features, and the underlying approaches.

5.1 Benchmarking of MediaEval systems

This section provides an in-depth analysis of the performance of the systems reported at the MediaEval Violent Scenes Detection Task (VSD), namely: 2011 — 28 runs, 2012 — 36 runs, 2013 — 57 runs, 2014 — 67 runs, and 2015 — 48 runs (a total of 236 runs).

The section is structured as follows: (i) we overview the most relevant methods (see Section 5.1.1), (ii) we analyze the overall performance trends (see Section 5.1.2), (iii) we study the influence of the employed content descriptors (see Section 5.1.3), (iv) we investigate the influence of the prediction methods (see Section 5.1.4), and (v) we analyze the reliability of the rankings and thus of our conclusions (see Section 5.1.5).

All the comparisons of the results are mainly carried out on a common basis, i.e., the official development-test set split. The systems are grouped according to the version of the data set they were benchmarked on (2011 to 2015), definition of violence (objective or subjective), granularity level (shot level, segment level or clip level detection), and type of videos (Hollywood or YouTube content).

7. http://trec.nist.gov/trec_eval/

5.1.1 Overview of the methods

We first provide a global description of the most relevant MediaEval systems. The analysis of their performance is provided in the following sections.

MediaEval 2011: Penet et al. [38] propose several methods based on K2 and Naive Bayesian Networks that use visual (e.g., shot duration, average number of blood pixels, average activity, number of flashes) and audio (e.g., energy, centroid, asymmetry, ZCR, flatness) features. Temporal aggregation of these features is achieved through the creation of contextual representations with the help of: (i) decision maximum voting, and (ii) probability averaging, that aggregated violence scores across consecutive segments. Both early and late fusion implementations are provided.

Glotin et al. [39] introduce the only unsupervised approach which employs entropic visual and audio confidence scores. For the visual modality, the authors implement a histogram of multi-scale LBP features over the whole image, as described in [40], while for the audio modality they use MFCC. Visual confidence is computed using the entropy of a probability mass function, while audio confidence is computed using the average of the normalized entropy of each MFCC probability distribution.

Safadi et al. [41] develop a visual violence detection system based on 3 types of traditional visual features: RGB histograms, Gabor Transforms, and bag of SIFT features. These descriptors are optimized in a two-step approach that involve power transformation in normalizing the distribution and PCA for decorrelation. The classification is achieved via kNN, employed for each feature vector, and a weighted late fusion scheme that aggregates the decision.

Gninkoun et al. [42] employ LDA and QDA for classification. The information is represented with audio (e.g., energy entropy, signal amplitude, short time energy, ZCR, SF, spectral rolloff), visual (e.g., shot length, shot motion, shot motion content, skewness of motion vector), and conceptual (based on swear words) features.

MediaEval 2012: Schlüter et al. [43] propose a MLP system built on mid-level features. Firstly, visual (e.g., HoG, CNH, visual activity) and audio (e.g., LPC, LSP, MFCC, ZCR, SC, SF, rolloff, kurtosis) features are used to train a MLP concept detector, targeting the provided set of concepts (see Section 3.2). Secondly, the final violence prediction is achieved via the fusion of the concept detectors. It uses the thresholding of the output of the networks, whereas the cutoff values are individually determined via a cross-validation process in the training phase.

Penet et al. [44] develop a hybrid learning method that uses Naive Bayesian Networks and K2, similar to the one in [38]. The information is represented with visual and audio features, e.g., color coherence, color harmonization, shot duration, average number of blood-like pixels determined in the HSV space, average activity, number of flashes, and energy, flatness, centroid, asymmetry, ZCR, respectively. Confidence scores from individual systems are averaged via a late fusion scheme.

Jiang et al. [45] propose a system that uses low-level audio and visual features, including motion (e.g., HoG, HoF, MBH), sparse keypoint detectors (e.g., DoG, Hessian), STIP features, and MFCC-based features. Two different methods for temporal aggregation are experimented: (i) feature

smoothing, consisting of averaging features, and (ii) score smoothing, consisting of averaging prediction scores. The classification is carried out via SVM, and several early and late fusion combinations are tested.

Lam et al. [46] introduce a visual concept detector system. Five keyframes are extracted from each shot, and several traditional visual features are extracted, e.g., CM, color histograms, EoH, and LBP. Features are aggregated at shot-level using max, min, and average pooling. A visual concept detector for each of the 7 visual violence concepts (see Section 3.2) is trained using an RBF kernel SVM. Late fusion is applied to the output of the 7 detectors to generate the final violence score.

MediaEval 2013: Tan et al. [47] employ multiple SVMs to predict mid-level violent concepts. They exploit low-level audio (e.g., MFCC, LSF, OBSI, linear predictor coefficients), and visual (e.g., HoG, HoF, MBH describing dense trajectories, the DoG and Hessian features via BoW) features. The authors use the 10 concepts provided with the data (see Section 3.2) and infer 42 more, using the external data, i.e., Youtube videos with automatic ground truth. Based on the ontology from ConceptNet [48], the authors construct a Conditional Random Fields [49] model that understands the relationships and co-occurrences of the 52 concepts. The SVM classifiers are combined in a late fusion approach, and score smoothing is applied to generate the final prediction.

Dai et al. [50] use trajectory features (i.e., a combination of HoG, HoF, and MBH features quantized via visual codebooks and TrajMF [51]), STIP, MFCC and part-level attributes based on the work in [52]. PCA is employed to reduce dimensionality. Each feature group is trained with a χ^2 or linear SVM, and the final violence prediction is computed via late fusion and temporal score smoothing.

Goto and Aoki [53] propose a system based on visual and temporal dense trajectory features [54], MBH, RGB histograms, and MFCC, MFCC with delta components and audio energy features. All the features are converted into BoW representations. The authors apply a MKL strategy for optimizing the weights of multiple SVM systems. The final system is based on a voting approach that uses a set of binary decisions extracted from each SVM to predict each segment, followed by an integration step to restore the continuity of the segments.

Derbas et al. [55] employ a set of 6 visual and audio features (e.g., RGB histograms, texture information via Gabor transforms, SIFT, MFCC, STIP, audio-visual BoW of MFCC and HoFs). The information is reduced using the PCA decorrelation. The classification is carried out independently for each descriptor, and late fusion with optimized weights is applied in the final stage. A combination of SVM and kNN is used for the classification of violence.

Penet et al. [56] introduce an audio concept detection system. It uses MFCCs, energy, and flatness coefficients. The authors train different contextual Bayesian Networks for each feature and aggregate them via a weighted late fusion scheme. The network classifies each sample according to its context. The final violence scores are produced using a simple chunk aggregation that groups contiguous segments that have the same label.

MediaEval 2014: Dai et al. [57] use visual and audio features, e.g., HoG, HoF, MBH encoded with Fisher Vectors, TrajMF [51] trajectory shape features encoded with Fisher Vectors, STIP, MFCC. Dimensionality reduction via expectation-maximization PCA is used for the TrajMF features. Two classifiers compose the best-performing systems: (i) an SVM classifier that uses Fisher Vectors of HoG, HoF, and MBH features, and (ii) a DNN-based classifier [58], [59] that uses the rest of the features. The final violence prediction is achieved via a late fusion scheme. Temporal score smoothing and clip merging are employed for determining segment-level predictions.

Sjöberg et al. [60] employ visual (e.g., CNH, CM, LBP, HoG, Color Structure Descriptor, Grey Level Run Length Matrix) and audio (e.g., MFCC, amplitude envelope, ZCR, SC and flux, RMS, band-energy ratio) features with a MLP which is trained for mid-level violent concept prediction. The input data is normalized with mean and standard deviation. The low-level audio-visual features, together with the mid-level conceptual predictions, are fed to the same MLP classifier for predicting the final violence scores. Temporal smoothing via median filtering and thresholding is used.

Zhang et al. [61] propose a salient keypoint trajectory-based detector. The authors extract an accurate detection of human motion via the method in [62] and then extract HoG, HoF, and MBH features around these points. These motion features are normalized with a square root approach, and the dimensionality is reduced via PCA. The video-level aggregation is then computed with Fisher Vector representations. Dense SIFT descriptors are computed at different scales and grids. MFCC features with delta and double-delta components, encoded via GMM, represent the audio description of the samples. SVM classification is processed separately for the visual and audio features, and the final result is computed using a non-weighted late fusion that sums the outputs of each of the SVMs.

MediaEval 2015: Dai et al. [63] use traditional visual and motion features (e.g., MBH, HoG, HoF, TrajShape encoded with Fisher Vectors, STIP encoded with BoW), audio features (e.g., MFCC) and a series of features extracted from DNN layers. The authors fine-tune AlexNet [64] using a subset of ImageNet composed of 2,614 manually selected classes representing categories of scenes, weapons, people, etc., that are semantically linked with violence, and extract features from the FC6, FC7 and FC8 layers of the network. Another DNN feature is obtained via the two-stream CNN from [65], which is composed of a spatial component, pre-trained on the full ImageNet data set, and a temporal stream that takes stacked optical flow information as input. The last 3 layers of the spatial component and the last layer of the temporal component are then used as input for an LSTM network, pre-trained on the UCF-101 data set, that models the dynamic information. The average output of the last LSTM layer is used as a feature. The final classification is performed with linear and χ^2 SVMs, and kernel-level fusion is adopted in generating the final violence scores.

Lam et al. [66] employ an SVM based system that uses motion features (e.g., a combination of HoG, HoF and MBH encoded with Fisher Vectors), audio features (e.g., MFCC encoded with GMM), and deep learning features (e.g., the FC6, FC7 and FC8 layers of the pre-trained VGG16 [67] model). The authors use a linear kernel for the motion and audio features, while a χ^2 kernel is used for training

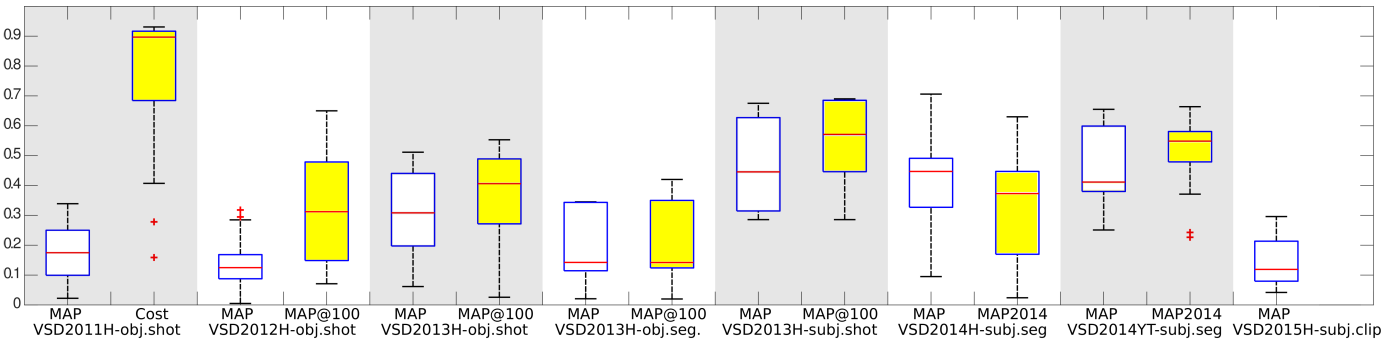

Fig. 2: Boxplot representation of the overall performance (interquartile range IQR of 50%, median values, lower and upper adjacent values calculated as 1.5·IQR, and outlier values marked with red crosses). Two metrics are presented for each data set: standard MAP and the official metric on which the submitted methods were optimized (H - Hollywood movies, YT - YouTube movies, obj - objective definition of violence, subj - subjective definition of violence, shot - video shot level, seg - video segment level, clip - video sequence level). No representation is provided for the VSD2012H-obj.seg and VSD2013H-subj.seg data sets due to the very reduced number of submitted systems, i.e., 1 and 2, respectively.

the deep features. The final results are generated using an average weighting approach.

Seddati et al. [68] also use a deep feature approach. The authors extract the TV-L1 optical flow map features using the implementation from [69], and use a 24-layer deep CNN as classifier for predicting violence.

Yi et al. [70] extract traditional visual features (e.g., IDT, dense SIFT, Hue-Saturation histograms), audio features (e.g., MFCC) and deep features (e.g., based on the CNN_M_2048 model from [71], pre-trained on ImageNet). The features are aggregated with Fisher Vectors, and classification is carried out via the use of SVM models for each feature. A linear late fusion approach generates the final violence scores. Several combinations of features are tested.

5.1.2 *Analysis of the overall performance*

We provide a global analysis of the results achieved during the MediaEval VSD campaigns. Figure 2 presents a boxplot representation of the results in terms of general mean average precision (MAP) and the original metrics of each year's campaign, namely: "MediaEval cost" for the 2011 data normalized between 0 and 1, mean average precision at a cutoff of 100 highest ranked shots (MAP@100) for the 2012 and 2013 data, mean average precision with special constraints for the 2014 data (see Section 4), and standard mean average precision for the 2015 data. It should be noted that a direct comparison between the general MAP score and the official metrics should be treated cautiously, as systems were not optimized in the same way. However, both metrics account for the same principle, i.e., capturing the prediction performance, and an overall statistical interpretation is valid.

For VSD 2011 to 2013, each movie is pre-segmented into video shots via shot boundary detection, and the algorithms are expected to conduct the prediction at shot level. Analyzing the data for the objective definition of violence, one can observe a significant increase of MAP in 2013, reaching up to 0.51, an increase of 18 percentage points over 2011. This can be explained by the increase in the number of training movies (see Table 1) and by the increasing diversity of runs submitted by the participants, including the use of more multimodal features and more advanced algorithms (see Section 5.1.3 and 5.1.4). Compared to the results achieved for the segment-level prediction in 2013, i.e., the systems are expected to predict the exact (variable length) violent parts of a movie, the best MAP is visibly lower, 0.34 vs. 0.51. This is expected, given the additional difficulty of localizing the segments precisely.

Looking at the subjective definition of violence, the results in 2013 for shot-level prediction show a significantly higher MAP compared to the objective definition, i.e., 0.67. Though this may seem surprising, an analysis of the annotations from VSD 2013 shows that more shots are labeled as violent under the subjective definition (20.24%) than under the objective definition (10.49%). This creates a lower class imbalance, and systems may be able to train better on the subjective task. Segment-level prediction, is also solved better for 2014, i.e., MAP of 0.7, compared to 0.35 in 2013. The robustness of the methods is sustained with the experimentation on the YouTube videos in 2014. Although the systems were trained on Hollywood movies, they were able to generalize well when evaluated on user-generated content, generally achieving slightly higher performance than the systems tested on Hollywood movies. The better performance for the YouTube videos may have also been the result of a lower class imbalance, i.e., on average, 44.47% violence duration for YouTube videos versus 16.26% for the movies in the testset.

In 2015, we noticed a significant drop in performance for the subjective definition of violence and clip-level prediction, i.e., the best MAP is 0.29. There is a higher number of individual clips, and the variety of content is a bit lower compared to Hollywood and YouTube data. Then, the systems are required to be more general, as the 2015 data was also shared for the prediction of induced affect [4]. Participants are asked to develop systems that can solve both violence detection and affect classification. This is not a regression of the methods, but on the contrary, the systems became more general, validated in-the-wild.

Finally, some systems stand up as outliers compared to the others. We review the positive ones. These systems achieved notable results in the 2012 campaign. Schlüter et

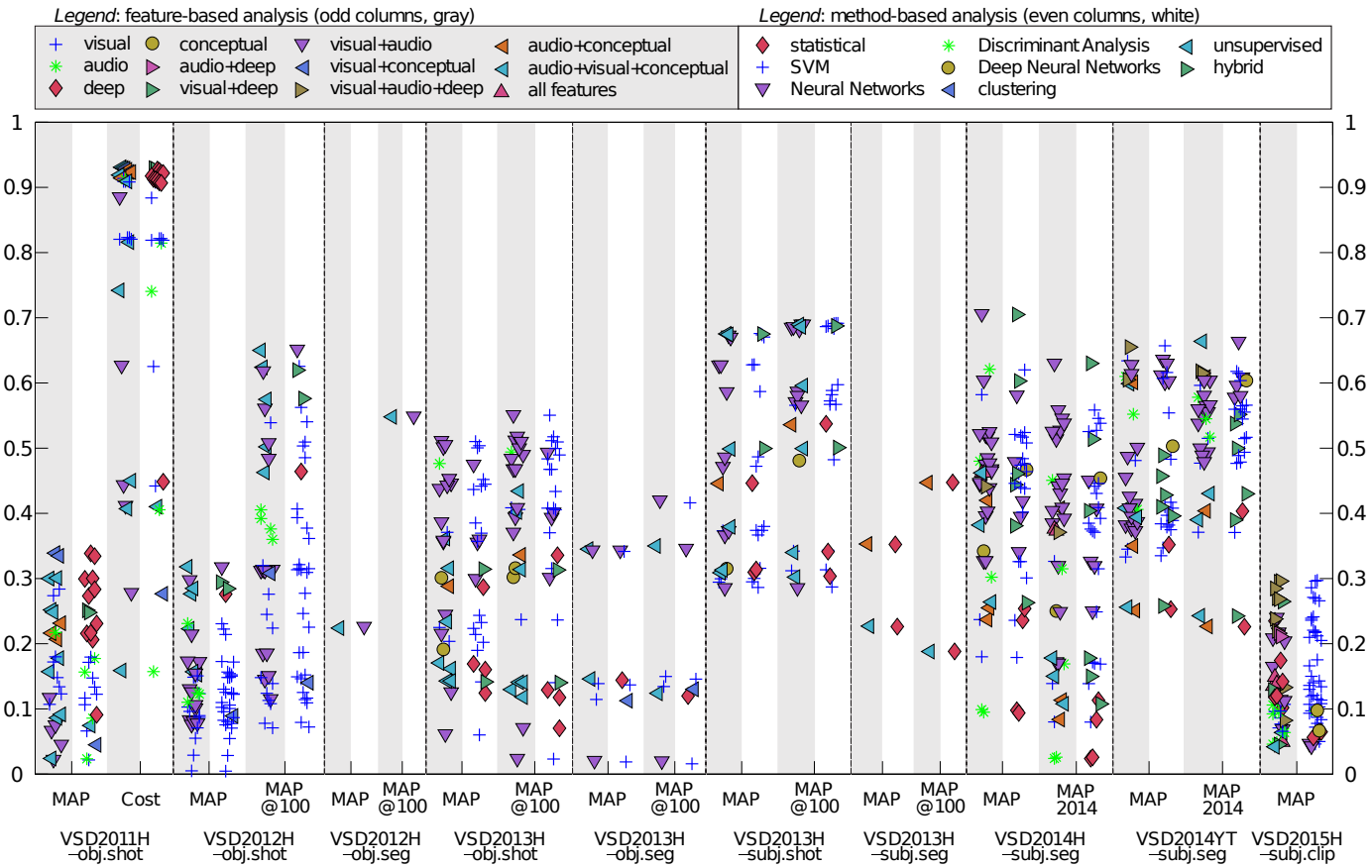

Fig. 3: Performance of the employed features and methods (gray and white columns, respectively). Two metrics are presented for each data set: standard MAP and the official metric on which the submitted methods were optimized. Different shapes and colors (generated using color blindness rules for maximal perceptual discrimination via the tool from https://medialab.github.io/iwanthue/) account for different modalities/techniques used and their combinations (H - Hollywood movies, YT - YouTube movies, obj - objective definition of violence, subj - subjective definition of violence, shot - video shot level, seg - video segment level, clip - video sequence level).

al. [43] employ low-level visual and auditory features (e.g., HoG, LPC, MFCC) and train a MLP via mid-level concepts, achieving a MAP@100 of 0.65 and a MAP of 0.318, which is the best result. Penet et al. [44] employ traditional visual and audio features and use a hybrid classification system (via K2 for audio and Bayesian Networks for visual information) achieving a MAP@100 of 0.6182 and a MAP of 0.2947 (see Section 5.1.1). One noteworthy mention is that both these systems use a multimodal approach, exploiting visual and audio features and concepts, therefore analyzing violence from different perspectives and improving on more simple unimodal approaches.

5.1.3 Analysis of the employed features

We provide an in-depth analysis of the employed features for the various submitted systems. The results are presented in Figure 3. The main modalities used are: (i) *visual content* (e.g., DT, dense SIFT, GIST features), (ii) *audio content* (e.g., ZCR, MFCC, SC), (iii) *conceptual features*, representing the mid-level concepts presented in Section 3.2 (e.g., "gunshots", "blood"), and (iv) *deep features*, representing the feature vectors extracted from the output of different layers of different DNNs (e.g., AlexNet, VGGNet). The different feature combinations are also represented. However, for practical reasons, not all the existing combinations are reported here. Investigating the best performing feature combinations on a per data set basis, we observed that several types of combinations consistently achieve the best results. In particular, 4 multi-modal feature combinations stand out: (i) *visual + audio*, (ii) *audio + conceptual*, (iii) *visual + audio + conceptual*, and (iv) *visual + audio + deep*. While some systems use only one type of concepts, e.g., only audio concepts, in our analysis, we consider them a different category than audio features, representing a higher-level symbolic description.

Multimodal. A first analysis reveals that out of the total of 236 runs, 69 % (163 runs) employ multimodal features, as combinations of audio, visual, conceptual, or deep features. Multimodal approaches are able to provide superior performance and have constantly been the best performers for each of the data sets. Single modal systems achieve, on average, across all the data sets, a MAP of 0.208, while the multimodal systems reach 0.3138. For instance, Dai et al. [57] use shape, motion and temporal visual features and MFCC audio features [58], [59], achieving a MAP score of 0.706 (VSD2014H-subj.seg). Other impressive performances are also multimodal, with MAP results of 0.675 and 0.674 (VSD2013H-subj.shot). Tan et al. [47] extend the concepts

list to up to 52 violence-related concepts based on external Youtube videos and ontology models extracted from ConceptNet. Motion and SIFT visual features, as well as MFCC, LSF, OBSI and linear predictor coefficients audio features are extracted from the frames/shots, corresponding to the detected concepts.

Traditional visual and audio features achieve the best results on 4 data sets: VSD2013H-obj.shot, VSD2013H-obj.seg, VSD2013H-subj.shot, and VSD2014H-subj.seg. For instance, Dai et al. [50] use a combination of features such as trajectory-based motion features, STIP, MFCC, obtaining a MAP@100 of 0.553 (VSD2013H-obj.shot). Goto and Aoki [53] achieve a MAP@100 of 0.42 using visual and temporal dense trajectory features [54], MBH, RGB histograms, MFCC and audio energy features (VSD2013H-obj.seg). All these features are concatenated and converted to a BoW representation. Two runs from Derbas et al. [55] reach a MAP@100 of 0.69 on VSD2013H-subj.shot. They employ color, texture, SIFT, STIP, MFCC, and joint audio-visual BoW features and PCA for decorrelation.

Audio and conceptual features provide the best results on VSD2013H-subj.seg, with a MAP@100 of 0.447. Penet et al. [56] use an audio concept detector based on MFCC, energy, and flatness.

The visual, audio and conceptual features are the best performers for the VSD2011H-obj.shot, VSD2012H-obj.shot, VSD2012H-obj.seg, and VSD2014YT-subj.seg data sets. Penet et al. [38] attain a MediaEval cost value of 0.761 (0.931 normalized) by using 5 audio features: energy, centroid, asymmetry, ZCR, and flatness, together with 4 visual features: shot duration, average number of blood pixels, average activity, and number of flashes (VSD2011H-obj.shot). The authors also use contextual representations to improve results. Schlüter et al. [43] achieve a MAP@100 of 0.65 (VSD2012H-obj.shot) using visual features based on color, shape and visual activity and 8 types of spectral and temporal audio features for predicting visual, audio, and audio-visual mid-level violent concepts. Sjöberg et al. [60] obtain a MAP2014 of 0.663 (VSD2014YT-subj.seg) using mid-level concepts based on color, texture, spectral and temporal audio features.

Concepts. Concept features are a particular case of descriptors and account for higher-level information. The audio and visual concepts annotated in the data sets are presented in Section 3.2, while the correlation between the concepts and violence is presented in Table 2. Many of the teams use the 10 concepts as classes for an intermediary system, creating machine learning methods to predict the presence of violent events, and then use the output predictions as features for their systems. For instance, Schlüter et al. [43] achieve the highest concept prediction recall for the detection of "blood" and "coldarms", while the highest precision and F-score, 0.24 and 0.3, respectively, are achieved for the detection of "fire". The authors use low-level visual features (HoG, CNH, and visual activity) and audio features (LPC, LSP, MFCC, ZCR, SC, SF, rolloff, and kurtosis). The results prove that some concepts such as "firearms" and "fire" show a good performance, while others, such as "carchase", perform badly.

Deep features. Deep features are now state of the art for many classification systems, in various domains like image recognition [72], generating artificial data [73] and action detection [74]. In the context of the VSD data, some participants choose to fine-tune existing DNNs and extract some of the CNN layers [63], others use pre-trained networks [66], while some create new models for this particular task [68]. Dai et al. [63] base one of their features on the FC6, FC7, and FC8 layers of a fine-tuned AlexNet model and manage to achieve the highest score on VSD2015H-subj.clip, MAP of 0.296. The tuning process consists of manually picking 2,614 classes from the ImageNet data set that are related to violence, and a retraing of the network. Lam et al. [66] use the pre-trained VGG [67] model, and it is worth noting that the addition of a feature set containing, among other descriptors, the feature vectors extracted from the FC6 and FC7 layers of the VGG network improves their results significantly, from a MAP of 0.22 to 0.268 (VSD2015H-subj.clip). Finally, the 2D architecture created by Seddati et al. [68] uses optical flow maps as inputs for the network. However, MAP was not as high as the one achieved by other approaches, reaching 0.09 (VSD2015H-subj.clip) with methods that used an adapted deep CNN with 5 convolutional layers (ConvNet).

Deep features are much more effective in combination with other modalities. For instance, Dai et al. [63], the best performers on VSD2015H-subj.clip, use visual, audio, and deep features, achieving a MAP of 0.296. In particular, they employ an AlexNet [64] based violence CNN descriptor, spatio-temporal CNN features aggregated with an LSTM model [75], conventional trajectory features encoded using Fisher Vectors, STIP features, and MFCC audio features. It is worth noting that overall, only 12% (29 runs) of the total number of runs employ deep features, with the first appearance of such features in 2014. In the following year, 2015, up to 50% (24 runs) of the runs used deep features, which shows an increasing interest in these approaches. A boost in performance is also visible if we compare the average MAP obtained with the runs employing deep features, 0.179, and the average MAP of the other type of features, 0.142, for the 2015 data.

Fusion. In terms of fusion techniques, the most employed approach is early fusion. Overall, 80% (188 runs) of the total number of runs use some kind of fusion technique; 163 runs use early fusion (by ensembling unimodal features before classification) and 97 runs use late fusion (by reducing unimodal features to separately learned model scores, and then ensemble to generate new predictions). These include also methods that use both, early and late fusion combined. However, an estimate of the average MAP across these runs shows that early fusion achieves 0.294, which is lower than late fusion with 0.343. Surprisingly, a higher performance is achieved when mixing both fusion techniques, an average MAP of 0.407.

5.1.4 Analysis of the detection methods

We analyze the performance of the deployed methods for VSD. The following categories were prominent: (i) SVM, (ii) statistical approaches (e.g., HMM, GMM, Bayesian approaches), (iii) Neural Networks (e.g., MLP with one hidden layer), (iv) Discriminant Analysis (e.g., QDA, Probabilistic LDA), (v) Deep Neural Networks (e.g., CNNs with multiple

hidden layers), (vi) clustering (e.g., kNN), (vii) unsupervised learning (e.g., entropic confidence), and (viii) hybrid approaches combining more than one type of methods[8]. An overview of the results is presented in Figure 3.

Support Vector Machines. The use of SVMs is consistently predominant with 65 % of the runs using one of its variants. Three top runs use it on the VSD2013H-obj.shot, VSD2014YT-subj.seg and VSD2015H-subj.clip data sets, and three top runs on the VSD2013H-obj.seg, VSD2013H-obj.shot, and VSD2013H-subj.shot data sets. For instance, Dai et al. [50] use a standard χ^2 SVM to classify the 4 types of audio-visual features. This approach achieves the top performance with a MAP@100 of 0.553 (VSD2013H-obj.shot). Tan et al. [47] employ multiple SVM classifiers to predict mid-level violent concepts using low-level audio-visual features. Furthermore, the authors use the provided audio and visual concepts (see Section 3.2) and inferred additional concepts by training over the external data gathered from YouTube. This approach generates a more diverse set of mid-level violent concepts and achieves the best results in terms of MAP, namely 0.675 (VSD2013H-subj.shot). For the VSD2013H-obj.seg data, Goto and Aoki [53] obtain the best result in terms of MAP@100, 0.420, using multiple SVM kernels, one for each feature type. The goal was to find the optimized weights when multiple SVM kernels are employed. The final system is based on a voting approach that uses a set of binary SVMs to predict each segment, followed by an integration technique to include the continuity of the segments. A multi-SVM approach is also employed for the VSD2015H-subj.clip data, where the best run in terms of MAP is obtained by Dai et al. [63], namely 0.296. The authors use two variants of SVM, namely a linear kernel and a χ^2 kernel, one for each different category of features, i.e., neural network-based and handcrafted.

Statistical approaches. Statistical methods account for 13% of the total number of runs, being the second most frequently used approach. For the VSD2013h-subj.seg data set, Penet et al. [56] achieve the best results in terms of MAP@100, 0.447, using a late fusion approach based on multiple Bayesian Networks for each audio feature type. The final scores are produced using a simple chunk aggregation technique via grouping the contiguous segments which yield the same label. The decision is set by a voting system where a segment inherits the probability of being violent from the highest probability of the segments that lie within the chunks. The K2 greedy methods are only used for VSD2011H-obj.shot data with good results. Penet et al. [38] obtain a MAP of 0.33, which is the top run in the aforementioned data set using the K2 algorithm to learn from the features that are extracted only from the visual modality. The authors refine the decision by exploiting the temporal structure of the movies, employing a temporal window, and taking the maximum decision over samples. The GMM/HMM-based methods do not produce notable results.

8. Please note that some of the categories might seem to include each other, e.g., deep networks and neural networks. In our analysis, "neural networks" refer to shallow networks with a single hidden layer, while "deep networks" refer to deep networks with multiple hidden layers. The scope of this analysis is to identify the performance of some subclasses of techniques and not to propose a categorization.

Neural networks. Standard neural networks represent 8% of the total number of runs. Schlüter et al. [43] achieve the best results in terms of MAP@100 for the VSD2012Hobj.shot and VSD2012H-obj.seg data sets, namely 0.650 and 0.548, respectively, by training a frame-wise violence predictor based on a MLP network. The system is built on a set of visual and auditory features, to predict violence from mid-level concepts (e.g., "blood" or "fire"). The final violence prediction score is obtained by thresholding the output of each network. The cut-off points are determined by maximizing the official metric in the training phase, using cross-validation. A similar approach is employed for the VSD2013H-obj.seg data, where the top run is obtained by Sjöberg et al. [76], with a MAP@100 of 0.350, and for the VSD2014YT-subj.seg data, where Sjöberg et al. [60] achieve the best MAP2014 of 0.663.

Deep neural networks. Deep network-based methods account for almost 2% of the total runs. While, in general, these methods are less effective than the other approaches, two runs stand out: one for the VSD2014H-subj.seg data and one for VSD2014YT-subj data, being situated over the third quartile. Dai et al. [57] use audio-visual features to train a DNN that captures the relationships between distinct features. This approach achieves a MAP2014 of 0.45 on VSD2014H-subj.seg, and a MAP2014 of 0.6 on VSD2014YT-subj.seg.

Discriminant analysis. Discriminant analysis methods account for 2 % of the total runs. The majority of the runs are concentrated on the VSD2011H-obj.shot data. These approaches do not achieve notable results, as the best run of Gninkoun et al. [42] is placed in the second quartile for both MAP and cost, 0.17 and 0.81, respectively. The authors train an LDA classifier based on audio-visual-textual features.

Clustering. Clustering-based techniques, such as the kNN, are employed in around 1% of the total number of runs and do not achieve notable results. They are first introduced on the VSD2011H-obj.shot data by Safadi et al. [41], achieving a MAP of 0.04 and a cost of 0.27.

Unsupervised approaches. It is worth mentioning here the only such approach, which is suggested by Glotin et al. [39], and tested on the VSD2011H-obj.shot data. The authors use entropic audio-visual confidences computed as the average of the entropies of visual and acoustic features achieving a MAP of 0.07 and a cost of 0.41. Although the results are low, it is a first notable attempt towards unsupervised approaches.

Hybrid approaches. Hybrid approaches consist of using more than one type of classifier, e.g., using SVM and kNN, and they account for 8 % of runs. Four top runs use this class of approaches. For instance, on the VSD2013H-subj.shot data, the top run, with a MAP@100 of 0.690, is obtained by Derbas et al. [55]. The authors use a hierarchical fusion of outputs based on a temporal re-ranking method using two different classifiers, one based on multiple SVMs for better handling of the class imbalance problem and one based on kNNs. Dai et al. [57] obtain the top run over the VSD2014H-subj.seg data, with a MAP2014 of 0.63, using a late fusion approach based on two classifiers: a DNN that models both feature correlation and feature diversity, and a χ^2 SVM.

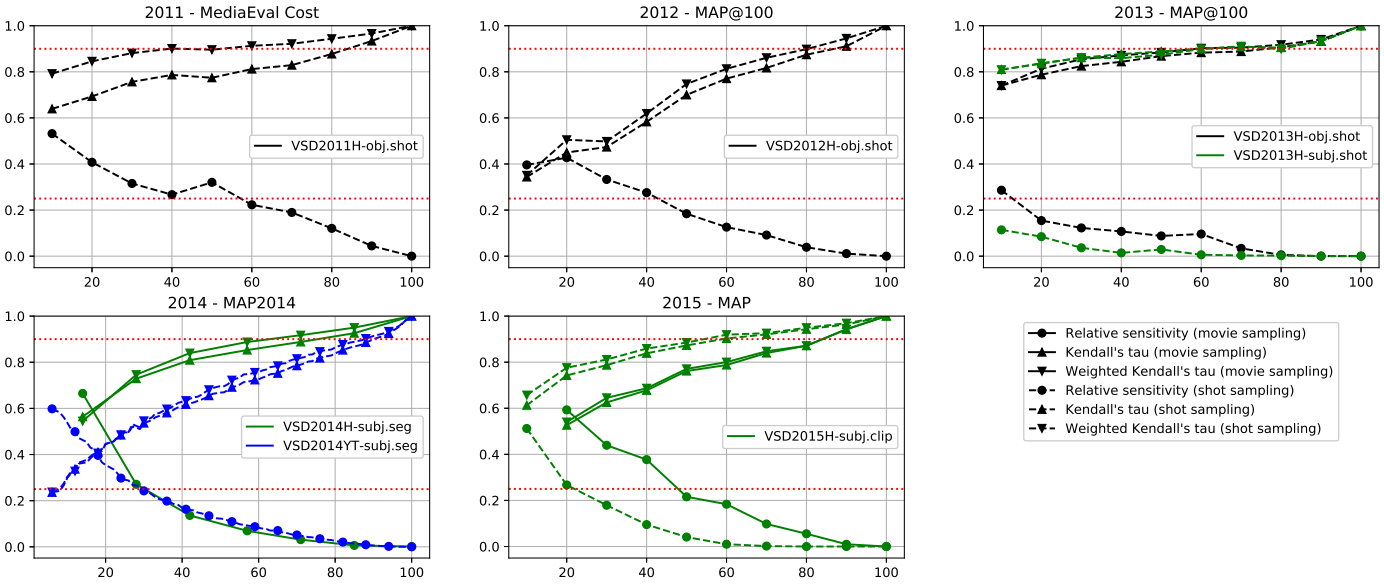

Fig. 4: Reliability scores of the system rankings (along the x-axis is the subsampling percentage and along the y-axis is the reliability score value). Relative sensitivity scores are marked with •, Kendall's tau with ▲, and weighted Kendall's with ▼. Shot-based sampling is marked with dashed lines. Use of objective definition of violence is marked with black, of subjective definition with green, and the YouTube clips from 2014 with blue. Finally, the reliability limits for the scores $\tau = 0.9$ and $\delta_r = 0.25$ are indicated with horizontal red dotted lines.

5.1.5 *Reliability analysis of performance scores*

We analyze the reliability of the previous method rankings, and therefore the reliability of our conclusions about the effectiveness of the systems. This analysis tests whether the same ranking results are obtained with different data configurations, and analyzes the stability of the results across different trials.

Systems are ranked using an evaluation metric based on comparing their responses to the ground truth for a set of queries $q \in \mathcal{Q}$. If we denote the score achieved by system A with $\lambda_{\mathcal{Q},\mathsf{A}}$, and the score received by a different system B with $\lambda_{\mathcal{Q},\mathsf{B}}$, we say that system A is better than system B if $\lambda_{\mathcal{Q},\mathsf{A}} > \lambda_{\mathcal{Q},\mathsf{B}}$. As demonstrated by Urbano et al. [77], if this ranking is reliable, it could be replicated with another set of queries $\mathcal{Q}'$, so that $\lambda_{\mathcal{Q}',\mathsf{A}} > \lambda_{\mathcal{Q}',\mathsf{B}}$ still holds.

We studied the reliability of rankings by randomly sampling pairs $(\mathcal{Q}', \mathcal{Q}'')$ of query subsets with the same size from the set of all queries used, $\mathcal{Q}$. We can then compare the system evaluations achieved with $\mathcal{Q}'$ with the ones achieved with $\mathcal{Q}''$. Urbano et al. [77] suggested several reliability indicators for performing this comparison. These measures can be grouped into two types: (i) score-based and (ii) ranking-based. As most of the measures were found to be relevant and highly correlated to each other, we have chosen one of each type for this study, namely the relative sensitivity and Kendall's rank correlation.

The relative sensitivity δ_r is defined as the minimum difference $(\lambda_{\mathcal{Q}',\mathsf{A}} - \lambda_{\mathcal{Q}',\mathsf{B}}) / \max(\lambda_{\mathcal{Q}',\mathsf{A}}, \lambda_{\mathcal{Q}',\mathsf{B}})$ that needs to be observed with $\mathcal{Q}'$ such that the differences with $\mathcal{Q}''$ have the same sign, at least 95 % of the time. Relative sensitivity should tend to 0 and $\delta_r = 0.25$ is given as the limit for reliability [78].

Kendall's rank correlation τ depends on the ranks of the systems only and does not consider the specific scores received by each system [79]. It counts the number of inversion of pairs of objects that would be needed to transform the ranking induced by $\mathcal{Q}'$ with the one by $\mathcal{Q}''$. The rank correlation ranges from 1 (identical rankings) to -1 (inverse ranking). Voorhees [80] establishes $\tau = 0.9$ as a limit for reliable ranking. In addition, we also compute a weighted version of Kendall's rank correlation τ_w, where exchanges of highly ranked objects are considered more influential than exchanges of low ranked objects [81]. We believe this is well-motivated in this case as the worst systems are performing essentially randomly, and their ranking can thus be considered somewhat arbitrary. We have used the additive hyperbolic weighting scheme, as suggested by Vigna [81].

For the different data sets, depending on the granularity of the annotations, we employ two different subsampling schemes: (i) movie-based and (ii) shot-based. In most cases, the data set consists of a small number of movies, from which several shots were extracted. Shots sampled from the same movie cannot be considered to be statistically independent, and thus, in theory, sampling on the movie level is more appropriate. Unfortunately, for the years 2011 to 2013, the number of movies was too small for movie-based subsampling to make sense. For those years, we have relied only on shot-based sampling. For movie-based sampling, we have subsampled in decrements of one, so that if the total number of movies is N, we have proceeded to randomly generate pairs of $N - 1$ movies, $N - 2$, and so on. For shot-based sampling, we have subsampled in decrements of 10 percentage points, i.e., 90 %, 80 %, and so forth. For each subsample size, the reported numbers are averages calculated across at least 50 randomly generated pairs of that size.

Figure 4 shows the reliability scores for each data set based on the official metrics for that particular data (see

also Figure 2). For all plots, the horizontal axis shows the subsampling percentage, while the vertical axis indicates the average reliability score for the pairs sampled at that level. Shot-based lines are dashed, while movie-based ones are full lines. The reliability limits for the scores $\tau = 0.9$ and $\delta_r = 0.25$ are indicated by horizontal red dotted lines.

Analyzing the numbers, one can observe that $\tau \geq 0.9$ is reached in all cases at 90 % sampling, or even earlier. In all cases, τ_w gives the same or a better correlation, in most cases reaching the reliability limit already at 60 %. This indicates that the ranking is often more reliable for the top ranks, which is expected, as often the poorest results can be essentially random. The limit for the relative sensitivity was reached in all years at 60 % or earlier, indicating high reliability. For 2015, where both movie-based and shot-based sampling was employed for the same videos, we can see that the scores are indeed better for the shot-based one, as one would expect. In 2014, the YouTube clips were marked as shot-based sampling, but they are, in fact, independent shots and do not have common movies. In fact, in 2014, we can see that the YouTube runs are less reliable than the regular movie runs. In 2013, the only year to have both objective and subjective definitions of violence, we can see that the subjective rankings were more reliable indicators of system performance. Finally, and most importantly, both Kendall's scores tend to 1, and the relative sensitivity tends to 0 as the number of queries that are evaluated increases. With weighted Kendall's and relative sensitivity scores, we cross the reliability limit in most cases already at 60 % subsampling.

5.2 Benckmarking of the state-of-the-art methods

To have a complete analysis of the existing systems' performance, we provide an in-depth analysis of the representative methods from the literature, trained and tested on the VSD96 data. These were not submitted to the MediaEval benchmark, and most importantly, they were developed without any time constraints. Figure 5 gives an overview of the results (note that some of the methods provide different runs with different parameters). To be able to compare them with the best results from the MediaEval campaign (see Section 5.1.4), we present, where available, the same official metrics. However, in some cases, these are not reported in the publication. In those cases, we present the metric reported by the authors (e.g., accuracy, AP@100).

For the VSD2012H-obj.shot data, 2 methods were particularly interesting due to their multimodal approach, the ones proposed by Eyben et al. [82] and Acar et al. [83]. Compared to the results achieved at MediaEval, all methods perform worse than the best method (see the green MediaEval runs in Figure 5).

For the VSD2013H-obj.shot, VSD2013H-subj.seg, and VSD2013H-subj.shot data sets, we have selected 8 approaches that stood out due to their diverse classification methods and features. They were proposed by Goto and Aoki [84], [85], Moreira et al. [86], Tan et al. [87], Lam et al. [88], [89], Derbas et al. [90], and Mironica et al. [91]. The majority of them achieved a score above the MediaEval average, 3 of them outperforming the best MediaEval run. Goto et al. [84], [85] used visual (HoG, MBHx, MBHy) and auditory (first derivative of MFCC and its energy) features in a multimodal approach. Starting from the premises that the features might be largely different depending on the characteristics of violence, the authors used a clustering process where each cluster is trained via MKL. The obtained scores are integrated, and the final decision of violence is obtained by thresholding. It achieves a MAP@100 of 0.55 for the objective definition of violence. Tan et al. [87] exploited a set of relations between the concepts from an ontology, such as spatial, temporal, social, physical, and psychological relations for synthesizing sentences with implied meaning. The approach uses three types of features, including visual (DoG, Hessian Affine and SIFT), audio (LSF, OBSI, LPC, MFCC, and their first and second-order derivatives), and action-oriented (HoG, HoF, and MBH). Furthermore, an SVM classifier is trained to detect the occurrence of an extended list of concepts crawled from YouTube, achieving a MAP@100 of 0.62 for the objective definition and a MAP@100 of 0.74 for the subjective one. Mironica et al. [91] creates a dictionary of frame words, based on a Random Forest approach, by computing a Fisher Kernel representation for each descriptor type, namely: visual (HoG and CNH), motion (HoF) and audio (LPC, LSP, MFCC, ZCR, SC, SF, roll-off, and kurtosis). The classification is performed via SVM, achieving a MAP@100 of 0.76 for the objective definition and a MAP@100 of 0.72 for the subjective one. This approach represents the state of the art for both data sets (objective and subjective definitions of violence).

For the VSD2014H-subj.seg and VSD2014YT-subj.seg data sets, we have selected 5 approaches that stood out due to their diversity. They were proposed by Khokher et al. [92], Ali et al. [93], Lam et al [94], [95], Sarman et al. [96], and Acar et al. [97]. From the aforementioned, the approach proposed by Khokher et al. [92] is the only one to surpass the best run at MediaEval. The authors use MFCC auditory features and DT visual features. To retain the structure of interactions between features, the vectors are further arranged in the form of tensors that undergo a tensor decomposition process. For the classification, robust features are selected using Fisher ranking, and then they are used to train a linear SVM. This approach achieved a MAP2014 of 0.6 for the Hollywood movies and a MAP2014 of 0.68 for the YouTube clips.

For the VSD2015H-subj.clip data set, we have selected 2 approaches that showcase two different methods to infer sub-concepts for violence detection. The methods were proposed by Acar et al. [98] and Li et al. [99]. The approach proposed by Li et al. [99] overcomes the best MediaEval run of Dai et al. [63] by a narrow margin. The authors extended the training annotations by manually labeling a set of sub-concepts to help the system better interpret the cross-data set divergence and consequently reduce the difficulty of generalizing the learned features to unseen data. The system learns from a handful of features such as the CNN features extracted from three pre-trained popular models, i.e., the last fully-connected layer of VGGNet trained on ImageNet, and the pool5 layer of GoogleNet [100] and GoogletNet4k [101], trained on ImageNet and a bottom-up reorganization of the ImageNet hierarchy, respectively; motion features based on Improved Dense Trajectory, and the auditory features based on MFCC. Finally, a set of linear

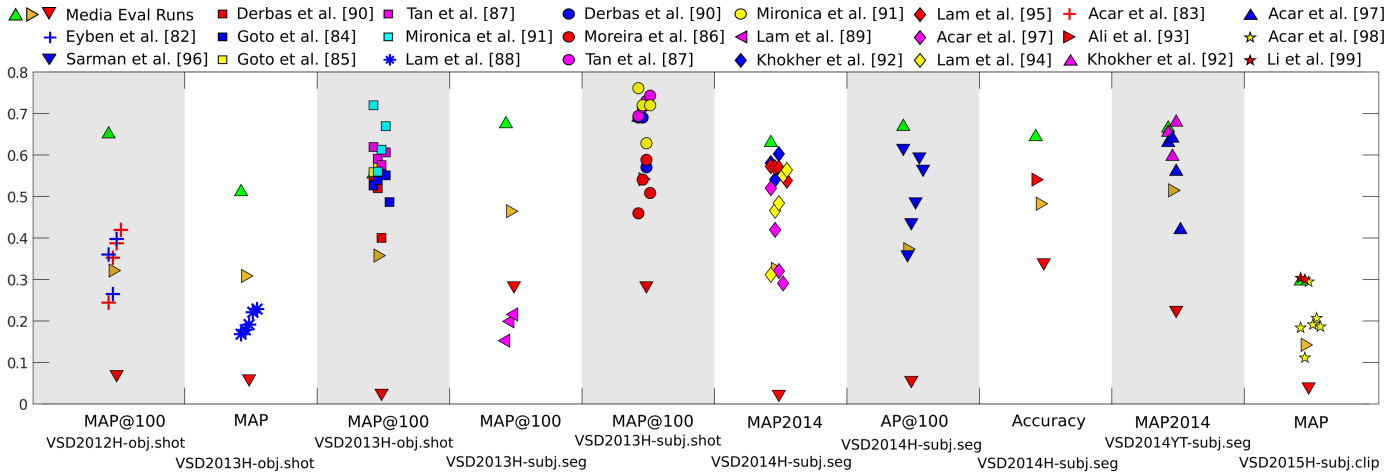

Fig. 5: Performance of the state-of-the-art methods. Different shapes and colors account for different approaches. In particular: green upward-pointing triangle accounts for the best MediaEval run on that data, orange right-pointing triangle accounts for the overall average score, and red downward-pointing triangle accounts for the lowest MediaEval run on that data (H - Hollywood movies, YT - YouTube movies, obj - objective definition of violence, subj - subjective definition of violence, shot - video shot level, seg - video segment level, clip - video sequence level).

SVMs is trained, and the late fusion is used for the final decision. It achieves a MAP of 0.3. This is again a good example of the effectiveness of mid-level-based systems that are able to capture more discriminant information compared to the use of other types of features.

6 Super System Design

In this final experiment, we take a holistic approach by investigating the possibility of building a *super system* on top of existing systems. To do so, we investigate some ad-hoc and standard late fusion schemes that are described below. Our goal is to create a late fusion scheme that can achieve significantly better results than the individual systems composing the scheme. While a drawback for this kind of approach is the high computational complexity necessary for running such a system, a significant improvement of the results would be hard to ignore and, given the ever-growing GPU-based parallel processing power, the implementation of such a system could be feasible. We do not aim to introduce a novel fusion scheme but to prove that although individual systems are powerful, and declared state of the art (even including on their own, some fusion schemes), there is always the possibility of achieving a greater performance via an ad-hoc fusion system.

To achieve our goal, for each year of the MediaEval benchmark, we have considered the best systems. We tested different groupings of the participants' runs, e.g., all runs altogether and diverse selections of the runs, e.g., by taking only those above a certain empirically determined threshold in terms of official metric and/or MAP. For the free segment-level prediction, we keep either the intersection (*inter*) or the union (*union*) of the predicted segments. For shot-level and clip-level prediction, there is no need for such aggregation as the boundaries are the same for all the systems. Once the segments' aggregation is ready, the modified prediction scores are determined by considering only the scores of temporally coherent segments from the original systems, i.e., the segments should have an intersection of at least a given fixed duration (e.g., 0.02 s, empirically determined) to be tagged as coherent.

Finally, the late fusion is carried out via the fusion of the scores by taking the minimum (*min*), the maximum (*max*), or the average (*avg*). We also tested a combination of min-max (*minmax*) that consists of taking the minimum value for resulting non-violent segments and the maximum value for violent segments. However, this was applied only to the free segment-level prediction.

The results are summarized in Table 4. Except for the VSD2013H-subj/obj.shot data sets, the super system achieves better, sometimes significantly higher performance (+11 % on average) over, both, the best MediaEval runs and the state-of-the-art approaches from the literature. This leads to the idea that the corresponding individual systems were complementary in terms of detection and did not generate too many false alarms, which could be foreseen as those systems reached already high recall levels, at the cost of lower precision. For the 2013 data and shot-level prediction, on the contrary, the recall of individual systems was globally lower and even lower than the precision. In this specific case, this leads to the incapability of any of the super system approaches to outperform the individual systems. Their fusion could not increase their global recall sufficiently.

7 Discussion and Conclusions

Violent Scenes Detection (VSD) is an increasingly important topic, especially with the exponential proliferation of the Internet among young children. There is an urgent need for tools able to automatically predict affect content on the spot, and therefore of machine learning resources for developing them. We have introduced a *common evaluation framework* for VSD that comes with a very large annotated, *publicly available* data set, i.e., the VSD96 data set: 31 full Hollywood movies, 86 YouTube video clips, and 10,900 clips extracted from 199 Internet Hollywood-like movies, summing up to more than 96 hours of video. These resources

TABLE 4: Super system design (MediaEval — the best result from the benchmark, see Section 5.1; SotA — the best result from the literature, see Section 5.2; inter/union — segment aggregation; min/max/avg/minmax — late fusion scheme; any union scheme — union/avg, union/min, union/max).

VSD2011H-obj.shot data set			
parameters	any union scheme	MAP>0.17 inter/avg	MAP>0.17 union/min
	MediaEval cost	*MAP*	*MAP@100*
super system	1	0.369	0.501
MediaEval	0.761	0.339	0.406
SotA	NA	NA	NA
VSD2012H-obj.shot data set			
parameters	all best MAP runs/any union scheme	MAP>0.2 MAP100>0.6 inter/avg	MAP>0.2 MAP100>0.6 union/avg
	MediaEval cost	*MAP*	*MAP@100*
super system	0.908	0.633	0.77
MediaEval	1	0.318	0.651
SotA	NA	NA	0.42
VSD2013H-obj.shot data set			
parameters		MAP@100>0.49 union/avg	MAP@100>0.46 union/avg
		MAP	*MAP@100*
super system		0.444	0.544
MediaEval		0.511	0.553
SotA		0.229	0.72
VSD2013H-obj.seg data set			
parameters		MAP@100>0.35 union/minmax	MAP@100>0.35 union/avg
		MAP	*MAP@100*
super system		0.428	0.632
MediaEval		0.345	0.42
SotA		NA	NA
VSD2013H-subj.shot data set			
parameters		MAP@100>0.49 union/max	MAP@100>0.49 union/max
		MAP	*MAP@100*
super system		0.504	0.6
MediaEval		0.675	0.69
SotA		NA	0.761
VSD2014H-subj.seg data set			
parameters	MAP2014>0.5 union/minmax union/max	all union/minmax	MAP2014>0.37 union/avg
	MAP2014	*MAP*	*MAP@100*
super system	0.6398	0.719	0.788
MediaEval	0.63	0.706	NA
SotA	0.602	NA	NA
VSD2014YT-subj.seg data set			
parameters	MAP2014>0.5 union/avg inter/minmax	MAP2014>0.6 union/min	MAP2014>0.6 union/min
	MAP2014	*MAP*	*MAP@100*
super system	0.722	0.828	0.828
MediaEval	0.655	0.664	NA
SotA	0.678	NA	NA
VSD2015H-subj.clip data set			
parameters		MAP>0.14 union/avg inter/avg	
		MAP	
super system		0.384	
MediaEval		0.296	
SotA		0.303	

were developed and validated during the yearly MediaEval benchmarking initiative for multimedia evaluation.

We have provided an in-depth analysis of the crucial components of the VSD algorithms, by reviewing the capabilities and the evolution of the existing systems with the objective to offer a complete practitioner's guide for this task. We reviewed 236 systems that were submitted to MediaEval and selected 17 state-of-the-art systems from the literature that were tested on VSD96 data, which constitute a strong baseline. We analyzed the reliability of the annotations and system rankings, examined various aspects, e.g., overall trends and outliers, the influence of the employed content descriptors, the prediction methods employed, and the possibility of aggregating the systems' outputs into an ad-hoc super system to achieve even greater performance. Below we are summarizing the most important lessons learned and insights gained.

Where are we with the current capabilities of the algorithms? (i) Regardless of the subjectivity of the task, as results show, *machine learning is successfully employed*, the best approaches reaching a performance above 75% (MAP). The free segment-level prediction is intuitively harder than the *pre-segmented video shot prediction*. The prediction of a more general definition of violence, i.e., *subjective definition*, is more successful than the prediction of the less general objective definition. Methods are robust enough to *generalize well to different data types*, as proved by the evaluation on YouTube data when learning was performed on Hollywood content. The difficulty of the prediction increases significantly with the generalization of the task. An example is the prediction on the Hollywood-like movie clips, where systems were expected to be even more general and predict both violence and the emotional impact of a video. It is worth noting that, by far, the most predominant and successful approach is still the use of classic *Support Vector Machines*, despite all current popularity of deep learning. (ii) As for what concerns the modalities used, the *multimodal (audio-visual)* approaches are naturally the best performers given the characteristics of the data set. The variety of employed features is quite impressive, and this leads to the conclusion that accurate predictions can be made with any reliable content representation. Among the considered features, the *learning of mid-level concepts*, i.e., symbolic intermediate descriptors, such as the presence of "blood", "screams", or "gore", stands out as it outperforms the sole use of low-level features. The average MAP over the methods employing these concepts is 0.304, compared to 0.273 for the rest of the methods. The early fusion was by far more popular than the late fusion of systems, but surprisingly, the mix of the two approaches, *early-late fusion*, was the best performer. (iii) It is important to notice that regardless of the superiority of deep learning approaches in other tasks, for VSD, *deep learning methods are not native state-of-the-art* approaches. One explanation could be the inherently multimodal nature of the data as well as the subjectivity of the task, which requires more adapted data representation models. Deep learning is able to achieve high performance, mostly when addressing a specific modality or type of data. Among all the analyzed systems, only one pure deep learning approach is able to achieve state-of-the-art performance, while the deep features proved to be more effective in combination with other features and standard classifiers. An analysis of the performance of the 8 predominant classes of VSD techniques shows that the best performing approach are the hybrid methods, with an average MAP across all the analyzed systems of 0.423. Simple MLPs with one hidden layer are the second-best performers, with an average MAP of 0.408, while deep learning methods are

only third-best with an average MAP of 0.282. This would seem to further suggest that in building a top-performing system, researchers may need to develop several modality-specific approaches and fuse their results. (iv) The final experiment, i.e., the use of late fusion for creating a super system, proved that an *ad-hoc standard late fusion approach is able to boost the performance significantly* and even surpass the state-of-the-art individual approaches by a large margin, e.g., more than 11 percentage points on average (MAP). This opens the possibility of considering a "blind" hybrid multi-system over a heavily crafted individual approach. However, there are also some inherent limitations. First, there is the very high computational complexity of the super system, which makes it less suitable for real-time applications. Then, the experiments show that not all the systems are suitable for fusion. Fusing only high performing systems allowed to achieve better performance. The inclusion of lower-performing systems decreased performance.

Where are we heading to with the capabilities of the algorithms? (i) *Unsupervised* or weakly supervised systems are a future milestone and a promising alternative to example-based learning. Creating manually annotated data to account for the ever-increasing diversity of the video material is a tedious task. In this respect, the exploration of Generative Adversarial Networks for generating VSD data would be an interesting lead. Among the analyzed systems, only one attempted an unsupervised approach, but the performance is still unsatisfactory. However, it may be seen as a proof of concept, and it can encourage further developments. (ii) *Deep learning network architectures* tailored to the VSD task could be devised, including the *native integration of multimodal data processing*, that would allow multiple modalities to be analyzed in an end-to-end manner. These networks are already available for other tasks, such as video hyperlinking. (iii) The *text modality* could be exploited much more intensely, given the availability of this information in online materials (e.g., YouTube videos). Among the analyzed systems, very few leveraged the provided text information, but those that did, achieved promising results. (iv) Migrating to a more general approach by linking VSD with the prediction of *user perceptions of multimedia content*, e.g., emotions or memorability, would be the next step towards creating a computer system capable of holistically recognizing how humans perceive audio-visual input.

Acknowledgments

We would like to acknowledge first, MediaEval and in particular Martha Larson, for the constant support with organizing the VSD task. We acknowledge the work of Mohammad Soleymani, Cédric Penet, Yu-Gang Jiang, Vu Lam Quang, Emmanuel Dellandréa, Hanli Wang, Yoann Baveye, Liming Chen, co-organizers of the VSD tasks. We acknowledge also Technicolor France for founding and supporting the MediaEval task. Mihai Gabriel Constantin, Liviu-Daniel Ştefan, and Bogdan Ionescu's work was supported by the Ministry of Innovation and Research, UEFISCDI, project SPIA-VA, agreement 2SOL/2017, grant PN-III-P2-2.1-SOL-2016-02-0002. Mats Sjöberg's work was supported by the Academy of Finland via project 313988 "DeepGraph".

References

[1] A. F. Smeaton, P. Over, and W. Kraaij, "High-level feature detection from video in trecvid: A 5-year retrospective of achievements," in *Multimedia Content Analysis. Signals and Communication Technology*, Springer, 2009.

[2] P. Over, G. Awad, J. Fiscus, G. Sanders, B. Shaw, M. Michel, A. F. Smeaton, W. Kraaij, and G. Qunot, "Trecvid 2013–an overview of the goals, tasks, data, evaluation mechanisms and metrics," in *Proceedings of TRECVID 2013*, vol. 6040, 2013.

[3] G. Awad, A. Butt, K. Curtis, Y. Lee, J. Fiscus, A. Godil, D. Joy, A. Delgado, A. Smeaton, Y. Graham, *et al.*, "Trecvid 2018: Benchmarking video activity detection, video captioning and matching, video storytelling linking and video search," in *Proceedings of TRECVID 2018*, 2018.

[4] M. Sjöberg, Y. Baveye, H. Wang, V. L. Quang, B. Ionescu, E. Dellandréa, M. Schedl, C. Demarty, and L. Chen, "The mediaeval 2015 affective impact of movies task," in *Working Notes Proceedings of the MediaEval 2015 Workshop*, vol. 1436 of *CEUR Workshop Proceedings*, 14–15 Sep 2015.

[5] B. BJ and H. L, "Short-term and long-term effects of violent media on aggression in children and adults," *Archives of Pediatrics and Adolescent Medicine*, vol. 160, no. 4, pp. 348–352, 2006.

[6] C. Demarty, C. Penet, M. Soleymani, and G. Gravier, "Vsd, a public dataset for the detection of violent scenes in movies: design, annotation, analysis and evaluation," *Multimedia Tools Applications*, vol. 74, no. 17, pp. 7379–7404, 2015.

[7] W. H. Assembly, "Prevention of violence: Public health priority," 1996.

[8] L. Chen, H. Hsu, L. Wang, and C. Su, "Violence detection in movies," in *8th International Conference on Computer Graphics, Imaging and Visualization*, pp. 119–124, IEEE, 17–19 Aug 2011.

[9] E. B. Nievas, O. Déniz-Suárez, G. B. García, and R. Sukthankar, "Violence detection in video using computer vision techniques," in *Computer Analysis of Images and Patterns - 14th International Conference*, pp. 332–339, Springer, Aug 29-31 2011.

[10] T. Giannakopoulos, A. Makris, D. I. Kosmopoulos, S. J. Perantonis, and S. Theodoridis, "Audio-visual fusion for detecting violent scenes in videos," in *Artificial Intelligence: Theories, Models and Applications, 6th Hellenic Conference on AI, SETN 2010. Proceedings*, vol. 6040 of *Lecture Notes in Computer Science*, pp. 91–100, Springer, 4-7 May 2010.

[11] Y. Gong, W. Wang, S. Jiang, Q. Huang, and W. Gao, "Detecting violent scenes in movies by auditory and visual cues," in *9th Pacific Rim Conference on Multimedia, Tainan, Taiwan*, vol. 5353 of *Lecture Notes in Computer Science*, pp. 317–326, Springer, 2008.

[12] D. Weinland, R. Ronfard, and E. Boyer, "Free viewpoint action recognition using motion history volumes," *Computer Vision and Image Understanding*, vol. 104, no. 2-3, pp. 249–257, 2006.

[13] L. An-An, S. Yu-Ting, N. Wei-Zhi, and Y. Zhao-Xuan, "Jointly learning multiple sequential dynamics for human action recognition," *PloS one*, vol. 10, no. 7, 2015.

[14] E. L. Andrade, S. Blunsden, and R. B. Fisher, "Modelling crowd scenes for event detection," in *18th International Conference on Pattern Recognition*, pp. 175–178, IEEE, 20–24 Aug 2006.

[15] K. Soomro, A. R. Zamir, and M. Shah, "UCF101: A dataset of 101 human actions classes from videos in the wild," *CoRR*, vol. abs/1212.0402, 2012.

[16] E. Y. Fu, H. V. Leong, G. Ngai, and S. C. F. Chan, "Automatic fight detection in surveillance videos," *Int. J. Pervasive Computing and Communications*, vol. 13, no. 2, pp. 130–156, 2017.

[17] M. Marsden, K. McGuinness, S. Little, and N. E. O'Connor, "Resnetcrowd: A residual deep learning architecture for crowd counting, violent behaviour detection and crowd density level classification," in *14th International Conference on Advanced Video and Signal Based Surveillance*, pp. 1–7, IEEE, 29 Aug – 1 Sep 2017.

[18] J. Shao, K. Kang, C. C. Loy, and X. Wang, "Deeply learned attributes for crowded scene understanding," in *Conference on Computer Vision and Pattern Recognition*, pp. 4657–4666, IEEE, 7-12 Jun 2015.

[19] C. Demarty, C. Penet, G. Gravier, and M. Soleymani, "A benchmarking campaign for the multimodal detection of violent scenes in movies," in *ECCV 2012. Workshops and Demonstrations. Lecture Notes in Computer Science*, vol. 7585, pp. 416–425, Springer, 7–13 Oct 2012.

[20] C. Demarty, B. Ionescu, Y. Jiang, V. L. Quang, M. Schedl, and C. Penet, "Benchmarking violent scenes detection in movies," in

12th International Workshop on Content-Based Multimedia Indexing, pp. 1–6, IEEE, 18–20 Jun 2014.

[21] C. Demarty, C. Penet, B. Ionescu, G. Gravier, and M. Soleymani, "Multimodal violence detection in hollywood movies: State-of-the-art and benchmarking," in *Fusion in Computer Vision - Understanding Complex Visual Content*, Advances in Computer Vision and Pattern Recognition, pp. 185–208, Springer, 2014.

[22] M. Schedl, M. Sjöberg, I. Mironică, B. Ionescu, V. L. Quang, Y. Jiang, and C. Demarty, "VSD2014: A dataset for violent scenes detection in hollywood movies and web videos," in *13th International Workshop on Content-Based Multimedia Indexing*, pp. 1–6, IEEE, 10–12 Jun 2015.

[23] Y. Baveye, E. Dellandréa, C. Chamaret, and L. Chen, "LIRIS-ACCEDE: A video database for affective content analysis," *IEEE Transactions on Affective Computing*, vol. 6, no. 1, pp. 43–55, 2015.

[24] J. L. Fleiss, "Measuring nominal scale agreement among many raters.," *Psychological bulletin*, vol. 76, no. 5, p. 378, 1971.

[25] J. J. Randolph, "Free-marginal multirater kappa (multirater k [free]): An alternative to fleiss' fixed-marginal multirater kappa.," *Online submission*, 2005.

[26] M. L. McHugh, "Interrater reliability: the kappa statistic," *Biochemia medica: Biochemia medica*, vol. 22, no. 3, pp. 276–282, 2012.

[27] R. L. Brennan and D. J. Prediger, "Coefficient kappa: Some uses, misuses, and alternatives," *Educational and psychological measurement*, vol. 41, no. 3, pp. 687–699, 1981.

[28] J. R. Landis and G. G. Koch, "The measurement of observer agreement for categorical data," *Biometrics*, vol. 33, pp. 159–174, Mar 1977.

[29] P. Knees and M. Schedl, *Music Similarity and Retrieval - An Introduction to Audio- and Web-based Strategies*, vol. 36 of *The Information Retrieval Series*. Springer, 1 ed., 2016.

[30] J. van de Weijer, C. Schmid, J. J. Verbeek, and D. Larlus, "Learning color names for real-world applications," *IEEE Trans. Image Processing*, vol. 18, no. 7, pp. 1512–1523, 2009.

[31] M. A. Stricker and M. Orengo, "Similarity of color images," in *Storage and Retrieval for Image and Video Databases III*, vol. 2420, pp. 381–392, SPIE, 5–10 Feb 1995.

[32] T. Ojala, M. Pietikäinen, and D. Harwood, "Performance evaluation of texture measures with classification based on kullback discrimination of distributions," in *12th International Conference on Pattern Recognition*, vol. 1, pp. 582–585, IEEE, 9-13 Oct 1994.

[33] N. Dalal and B. Triggs, "Histograms of oriented gradients for human detection," in *2005 Conference on Computer Vision and Pattern Recognition*, pp. 886–893, IEEE, 20–26 Jun 2005.

[34] C. Buckley and E. M. Voorhees, "Evaluating evaluation measure stability," in *23rd Annual International Conference on Research and Development in Information Retrieval*, pp. 33–40, 24–28 Jul 2000.

[35] C. D. Manning, P. Raghavan, and H. Schütze, *Introduction to information retrieval*. Cambridge University Press, 2008.

[36] H. W. Kuhn, "The hungarian method for the assignment problem," *Naval research logistics quarterly*, vol. 2, no. 1-2, pp. 83–97, 1955.

[37] A. F. Martin, G. R. Doddington, T. Kamm, M. Ordowski, and M. A. Przybocki, "The DET curve in assessment of detection task performance," in *5th European Conference on Speech Communication and Technology*, 22–25 Sep 1997.

[38] C. Penet, C. Demarty, G. Gravier, and P. Gros, "Technicolor and INRIA/IRISA at mediaeval 2011: learning temporal modality integration with bayesian networks," in *Working Notes Proceedings of the MediaEval 2011 Workshop*, vol. 807 of *CEUR Workshop Proceedings*, 1–2 Sep 2011.

[39] H. Glotin, J. Razik, S. Paris, and J. Prevot, "Real-time entropic unsupervised violent scenes detection in hollywood movies – DYNI @ mediaeval affect task 2011," in *Working Notes Proceedings of the MediaEval 2011 Workshop*, vol. 807 of *CEUR Workshop Proceedings*, 1–2 Sep 2011.

[40] S. Paris, H. Glotin, and Z.-Q. Zhao, "Real-time face detection using integral histogram of multi-scale local binary patterns," in *International Conference on Intelligent Computing*, pp. 276–281, Springer, 2011.

[41] B. Safadi and G. Quénot, "LIG at mediaeval 2011 affect task: use of a generic method," in *Working Notes Proceedings of the MediaEval 2011 Workshop*, vol. 807 of *CEUR Workshop Proceedings*, 1–2 Sep 2011.

[42] G. Gninkoun and M. Soleymani, "Automatic violence scenes detection: A multi-modal approach," in *Working Notes Proceedings of the MediaEval 2011 Workshop*, vol. 807 of *CEUR Workshop Proceedings*, 1–2 Sept 2011.

[43] J. Schlüter, B. Ionescu, I. Mironică, and M. Schedl, "ARF @ mediaeval 2012: An uninformed approach to violence detection in hollywood movies," in *Working Notes Proceedings of the MediaEval 2012 Workshop*, vol. 927 of *CEUR Workshop Proceedings*, 4–5 Oct 2012.

[44] C. Penet, C. Demarty, M. Soleymani, G. Gravier, and P. Gros, "Technicolor/inria/imperial college london at the mediaeval 2012 violent scene detection task," in *Working Notes Proceedings of the MediaEval 2012 Workshop*, vol. 927 of *CEUR Workshop Proceedings*, 4–5 Oct 2012.

[45] Y. Jiang, Q. Dai, C. C. Tan, X. Xue, and C. Ngo, "The shanghai-hongkong team at mediaeval2012: Violent scene detection using trajectory-based features," in *Working Notes Proceedings of the MediaEval 2012 Workshop*, vol. 927 of *CEUR Workshop Proceedings*, 4–5 Oct 2012.

[46] V. Lam, D. Le, S. P. Le, S. Satoh, and D. A. Duong, "Nii, japan at mediaeval 2012 violent scenes detection affect task," in *Working Notes Proceedings of the MediaEval 2012 Workshop*, vol. 927 of *CEUR Workshop Proceedings*, 4–5 Oct 2012.

[47] C. C. Tan and C. Ngo, "The vireo team at mediaeval 2013: Violent scenes detection by mid-level concepts learnt from youtube," in *Proceedings of the MediaEval 2013 Workshop*, vol. 1043 of *CEUR Workshop Proceedings*, 18–19 Oct 2013.

[48] R. Speer, J. Chin, and C. Havasi, "Conceptnet 5.5: An open multilingual graph of general knowledge," in *Thirty-First AAAI Conference on Artificial Intelligence*, 2017.

[49] J. D. Lafferty, A. McCallum, and F. C. N. Pereira, "Conditional random fields: Probabilistic models for segmenting and labeling sequence data," in *Proceedings of the 18th International Conference on Machine Learning*, pp. 282–289, 2001.

[50] Q. Dai, J. Tu, Z. Shi, Y. Jiang, and X. Xue, "Fudan at mediaeval 2013: Violent scenes detection using motion features and part-level attributes," in *Proceedings of the MediaEval 2013 Workshop*, vol. 1043 of *CEUR Workshop Proceedings*, 18–19 Oct 2013.

[51] Y.-G. Jiang, Q. Dai, X. Xue, W. Liu, and C.-W. Ngo, "Trajectory-based modeling of human actions with motion reference points," in *European Conference on Computer Vision*, pp. 425–438, Springer, 2012.

[52] Y. Zheng, Y.-G. Jiang, and X. Xue, "Learning hybrid part filters for scene recognition," in *European Conference on Computer Vision*, pp. 172–185, Springer, 2012.

[53] S. Goto and T. Aoki, "TUDCL at mediaeval 2013 violent scenes detection: Training with multi-modal features by MKL," in *Proceedings of the MediaEval 2013 Workshop*, vol. 1043 of *CEUR Workshop Proceedings*, 18–19 Oct 2013.

[54] H. Wang, A. Kläser, C. Schmid, and C. Liu, "Dense trajectories and motion boundary descriptors for action recognition," *International Journal of Computer Vision*, vol. 103, no. 1, pp. 60–79, 2013.

[55] N. Derbas, B. Safadi, and G. Quénot, "LIG at mediaeval 2013 affect task: Use of a generic method and joint audio-visual words," in *Proceedings of the MediaEval 2013 Workshop*, vol. 1043 of *CEUR Workshop Proceedings*, 18–19 Oct 2013.

[56] C. Penet, C. Demarty, G. Gravier, and P. Gros, "Technicolor/inria team at the mediaeval 2013 violent scenes detection task," in *Proceedings of the MediaEval 2013 Workshop*, vol. 1043 of *CEUR Workshop Proceedings*, 18–19 Oct 2013.

[57] Q. Dai, Z. Wu, Y. Jiang, X. Xue, and J. Tang, "Fudan-njust at mediaeval 2014: Violent scenes detection using deep neural networks," in *Working Notes Proceedings of the MediaEval 2014 Workshop*, vol. 1263 of *CEUR Workshop Proceedings*, 16–17 Oct 2014.

[58] Z. Wu, Y. Jiang, J. Wang, J. Pu, and X. Xue, "Exploring inter-feature and inter-class relationships with deep neural networks for video classification," in *Proceedings of the International Conference on Multimedia*, pp. 167–176, ACM, 03–07 Nov 2014.

[59] J. Tu, Z. Wu, Q. Dai, Y. Jiang, and X. Xue, "Challenge huawei challenge: Fusing multimodal features with deep neural networks for mobile video annotation," in *International Conference on Multimedia and Expo Workshops*, pp. 1–6, IEEE, 14–18 Jul 2014.

[60] M. Sjöberg, I. Mironică, M. Schedl, and B. Ionescu, "FAR at mediaeval 2014 violent scenes detection: A concept-based fusion approach," in *Working Notes Proceedings of the MediaEval 2014 Workshop*, vol. 1263 of *CEUR Workshop Proceedings*, 16–17 Oct 2014.

[61] B. Zhang, Y. Yi, H. Wang, and J. Yu, "Mic-tju at mediaeval violent scenes detection (vsd) 2014," in *Working Notes Proceedings of the MediaEval 2014 Workshop*, vol. 1263 of *CEUR Workshop Proceedings*, 16–17 Oct 2014.

[62] H. Wang and C. Schmid, "Action recognition with improved trajectories," in *Proceedings of the IEEE international conference on computer vision*, pp. 3551–3558, 2013.

[63] Q. Dai, R. Zhao, Z. Wu, X. Wang, Z. Gu, W. Wu, and Y. Jiang, "Fudan-huawei at mediaeval 2015: Detecting violent scenes and affective impact in movies with deep learning," in *Working Notes Proceedings of the MediaEval 2015 Workshop*, vol. 1436 of *CEUR Workshop Proceedings*, 14–15 Sep 2015.

[64] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," in *26th Conference on Neural Information Processing Systems.*, pp. 1106–1114, IEEE, 3–6 Dec 2012.

[65] H. Ye, Z. Wu, R.-W. Zhao, X. Wang, Y.-G. Jiang, and X. Xue, "Evaluating two-stream cnn for video classification," in *International Conference on Multimedia Retrieval*, pp. 435–442, ACM, 2015.

[66] V. Lam, S. P. Le, D. Le, S. Satoh, and D. A. Duong, "NII-UIT at mediaeval 2015 affective impact of movies task," in *Working Notes Proceedings of the MediaEval 2015 Workshop*, vol. 1436 of *CEUR Workshop Proceedings*, 14–15 Sep 2015.

[67] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," *CoRR*, vol. abs/1409.1556, 2014.

[68] O. Seddati, E. Kulah, G. Pironkov, S. Dupont, S. Mahmoudi, and T. Dutoit, "Umons at mediaeval 2015 affective impact of movies task including violent scenes detection," in *Working Notes Proceedings of the MediaEval 2015 Workshop*, vol. 1436 of *CEUR Workshop Proceedings*, 14–15 Sep 2015.

[69] J. S. Pérez, E. Meinhardt-Llopis, and G. Facciolo, "Tv-l1 optical flow estimation," *Image Processing On Line*, vol. 2013, pp. 137–150, 2013.

[70] Y. Yi, H. Wang, and B. Zhang, "Mic-tju in mediaeval 2015 affective impact of movies task," in *Working Notes Proceedings of the MediaEval 2015 Workshop*, vol. 1436 of *CEUR Workshop Proceedings*, 14–15 Sep 2015.

[71] K. Chatfield, K. Simonyan, A. Vedaldi, and A. Zisserman, "Return of the devil in the details: Delving deep into convolutional nets," *arXiv preprint arXiv:1405.3531*, 2014.

[72] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Conference on Computer Vision and Pattern Recognition*, pp. 770–778, 2016.

[73] A. Nguyen, J. Clune, Y. Bengio, A. Dosovitskiy, and J. Yosinski, "Plug & play generative networks: Conditional iterative generation of images in latent space," in *Conference on Computer Vision and Pattern Recognition*, pp. 4467–4477, 2017.

[74] X. Yang, X. Yang, M.-Y. Liu, F. Xiao, L. S. Davis, and J. Kautz, "Step: Spatio-temporal progressive learning for video action detection," in *Conference on Computer Vision and Pattern Recognition*, pp. 264–272, 2019.

[75] Z. Wu, X. Wang, Y. Jiang, H. Ye, and X. Xue, "Modeling spatial-temporal clues in a hybrid deep learning framework for video classification," in *23rd Annual Conference on Multimedia Conference*, pp. 461–470, ACM, 26–30 Oct 2015.

[76] M. Sjöberg, J. Schlüter, B. Ionescu, and M. Schedl, "FAR at mediaeval 2013 violent scenes detection: Concept-based violent scenes detection in movies," in *Proceedings of the MediaEval 2013 Workshop*, vol. 1043 of *CEUR Workshop Proceedings*, 18–19 Oct 2013.

[77] J. Urbano, M. Marrero, and D. Martín, "On the measurement of test collection reliability," in *36th International conference on research and development in Information Retrieval*, pp. 393–402, ACM, 28 Jul–1 Aug 2013.

[78] M. Sanderson and J. Zobel, "Information retrieval system evaluation: effort, sensitivity, and reliability," in *28th Annual International Conference on Research and Development in Information Retrieval*, pp. 162–169, ACM, 15–19 Aug 2005.

[79] H. Abdi, "The kendall rank correlation coefficient," *Encyclopedia of Measurement and Statistics. Sage*, pp. 508–510, 2007.

[80] E. M. Voorhees, "Variations in relevance judgments and the measurement of retrieval effectiveness," in *Proceedings of the 21st Annual International Conference on Research and Development in Information Retrieval*, pp. 315–323, ACM, 24–28 Aug 1998.

[81] S. Vigna, "A weighted correlation index for rankings with ties," in *24th International Conference on World Wide Web*, pp. 1166–1176, ACM, 18–22 May 2015.

[82] F. Eyben, F. Weninger, N. Lehment, B. Schuller, and G. Rigoll, "Affective video retrieval: Violence detection in hollywood movies by large-scale segmental feature extraction," *PloS one*, vol. 8, no. 12, 2013.

[83] E. Acar, F. Hopfgartner, and S. Albayrak, "Violence detection in hollywood movies by the fusion of visual and mid-level audio cues," in *Multimedia Conference*, pp. 717–720, ACM, 21–25 Oct 2013.

[84] S. Goto and T. Aoki, "Violent scenes detection using mid-level violence clustering," *Computer Science. CSCP*, pp. 283–296, 2014.

[85] S. Goto and T. Aoki, "Violent scenes detection based on automatically-generated mid-level violent concepts," in *19th Computer Vision Winter Workshop*, 2014.

[86] D. Moreira, S. E. F. de Avila, M. Perez, D. Moraes, V. Testoni, E. Valle, S. Goldenstein, and A. Rocha, "Temporal robust features for violence detection," in *Winter Conference on Applications of Computer Vision*, pp. 391–399, IEEE, 24–31 Mar 2017.

[87] C. C. Tan and C. Ngo, "On the use of commonsense ontology for multimedia event recounting," *IJMIR*, vol. 5, no. 2, pp. 73–88, 2016.

[88] V. Lam, D. Le, S. P. Le, S. Satoh, D. A. Duong, and T. D. Ngo, "Evaluation of low-level features for detecting violent scenes in videos," in *International Conference on Soft Computing and Pattern Recognition*, pp. 213–218, IEEE, 15-18 Dec 2013.

[89] V. Lam, S. Phan, T. D. Ngo, D. Le, D. A. Duong, and S. Satoh, "Violent scene detection using mid-level feature," in *International Symposium on Information and Communication Technology*, pp. 198–205, ACM, 5-6 Dec 2013.

[90] N. Derbas and G. Quénot, "Joint audio-visual words for violent scenes detection in movies," in *International Conference on Multimedia Retrieval*, p. 483, ACM, 1–4 Apr 2014.

[91] I. Mironică, I. Duţă, B. Ionescu, and N. Sebe, "Beyond bag-of-words: Fast video classification with fisher kernel vector of locally aggregated descriptors," in *International Conference on Multimedia and Expo*, pp. 1–6, IEEE, 29 Jun–3 Jul 2015.

[92] M. R. Khokher, A. Bouzerdoum, and S. L. Phung, "Violent scene detection using a super descriptor tensor decomposition," in *International Conference on Digital Image Computing: Techniques and Applications*, pp. 1–8, IEEE, 23–25 Nov 2015.

[93] A. Ali and N. Senan, "Violence video classification performance using deep neural networks," in *3rd International Conference on Soft Computing and Data Mining*, vol. 700 of *Advances in Intelligent Systems and Computing*, pp. 225–233, Springer, 6–7 Feb 2018.

[94] V. Lam, S.-P. Le, T. Do, T. D. Ngo, D.-D. Le, and D. A. Duong, "Computational optimization for violent scenes detection," in *International Conference on Computer, Control, Informatics and its Applications*, pp. 141–146, IEEE, 3–5 Oct 2016.

[95] V. Lam, S. Phan, D.-D. Le, D. A. Duong, and S. Satoh, "Evaluation of multiple features for violent scenes detection," *Multimedia Tools and Applications*, vol. 76, pp. 7041–7065, Mar 2017.

[96] S. Sarman and M. Sert, "Audio based violent scene classification using ensemble learning," in *6th International Symposium on Digital Forensic and Security*, pp. 1–5, IEEE, 2018.

[97] E. Acar, M. Irrgang, D. Maniry, and F. Hopfgartner, *Detecting Violent Content in Hollywood Movies and User-Generated Videos*, pp. 291–314. Springer International Publishing, 2015.

[98] E. Acar, F. Hopfgartner, and S. Albayrak, "Breaking down violence detection: Combining divide-et-impera and coarse-to-fine strategies," *Neurocomputing*, vol. 208, pp. 225–237, 2016.

[99] X. Li, Y. Huo, Q. Jin, and J. Xu, "Detecting violence in video using subclasses," in *2016 Conference on Multimedia*, pp. 586–590, ACM, 15–19 Oct 2016.

[100] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. E. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich, "Going deeper with convolutions," in *Computer Vision and Pattern Recognition*, pp. 1–9, 2015.

[101] P. Mettes, D. C. Koelma, and C. G. M. Snoek, "The imagenet shuffle: Reorganized pre-training for video event detection," in *International Conference on Multimedia Retrieval*, pp. 175–182, 2016.

Mihai Gabriel Constantin is currently a PhD candidate with the Faculty of Electronics, Telecommunications and Information Technology and a researcher with the Multimedia Lab, CAMPUS Research Center, University Politehnica of Bucharest, Romania. His research focused on the study of methods for analyzing the visual impact of multimedia data. He has authored over 11 scientific publications and was involved in several Romanian/EU funded research projects. He was a member of the organizing team for several conferences (e.g., IEEE/ACM CBMI 2016, ACM ICMR 2017) and benchmarking tasks (e.g., 2018 MediaEval Recommending Movies Using Content task).

Liviu-Daniel Ştefan is a PhD candidate with the Faculty of Electronics, Telecommunications and Information Technology, University Politehnica of Bucharest (UPB), and a researcher with the Multimedia Lab, CAMPUS Research Center, UPB, Romania. His research interests cover deep learning for multimedia classification, and multimedia/video/image processing and analysis for various applications, e.g. video surveillance, medical diagnosis. He has been a member of the organizing team for several benchmarking campaigns (e.g., ChaLearn ICPR Multimedia Information Processing for Personality & Social Networks Analysis Challenge), conferences (e.g., ACM ICMR 2017, ICPR workshops 2018) and part of the research team of several Romanian/EU funded research projects.

Bogdan Ionescu is general manager of the Research Center CAMPUS at University Politehnica of Bucharest (UPB). He holds a double PhD degree in image/video processing from UPB and University of Savoie, France. He is currently a tenured Professor with ETTI-UPB. His main research interests cover: multimedia/video/image processing and analysis, multimedia content-based retrieval and machine learning for multimedia. He has authored over 160 scientific publications. He serves/served as guest co-editor for Image and Vision Computing, Multimedia Tools and Applications, and International Journal of Computer Vision; conference committee chair for various conferences; lead organizer/co-organizer for several benchmark campaigns: MediaEval Retrieving Diverse Social Images, Violent Scenes Detection, Affective Impact of Movies Task, Predicting Media Interestingness, and Predicting Media Memorability, ImageCLEF 2017-2019, ChaLearn ICPR Multimedia Information Processing for Personality & Social Networks Analysis Challenge.

Claire-Hélène Demarty is senior scientist at InterDigital R&I, Rennes, France. She graduated from Telecom ParisTech in 1994 and received a Ph.D. degree in Computer Science, Mathematical Morphology, from Mines ParisTech in 2000. Prior to InterDigital, she was senior researcher and then senior scientist at Technicolor, Research & Innovation Center, France (2004 to 2019). Prior to Technicolor, she worked at LTUTechnologies (2000 to 2002) and at INRIA Rennes IRISA (2003 to 2004), a French public research center, as a researcher in image analysis and video indexing technologies. Her research focuses on multimedia indexing technologies and perceptual understanding of content, through the use of machine learning. She is author or co-author of more than 30 publications and is holding several patents. She was lead organizer of the Affect Task - Violent Scenes Detection from 2011 to 2013, of the Predicting Media Interestingness Task in 2016-2017 and of the Predicting Media Memorability in 2018, in the MediaEval benchmark.

Mats Sjöberg works as a Machine Learning Specialist at CSC – IT Center for Science Ltd in Finland. He received his doctorate degree in computer science from Aalto University, Finland in 2014. During 2014-2017 he worked in the Intelligent Interactive Information Access research group at the Department of Computer Science at the University of Helsinki. In 2018 he worked in the Content-Based Image and Information Retrieval Group at Aalto University. In 2019 Dr. Sjöberg joined CSC, Finland's national supercomputing center for research. His research has focused on applying machine learning to real-world problems, in particular multimedia retrieval, affective computing, visual concept detection, and intelligent personal information access. He was the lead organizer of the MediaEval benchmark's "Violent Scenes Detection" Task in 2014, and the "Affective Impact of Movies" Task in 2015, and has been a co-organiser in several multimedia affect-related tasks since. Dr. Sjöberg is the author of more than 50 scientific publications.

Markus Schedl is professor at the Johannes Kepler University (JKU) Linz, Institute of Computational Perception, and Linz Institute of Technology (LIT), AI Lab. He graduated in Computer Science from the Vienna University of Technology and earned his Ph.D. in Computer Science from the JKU. His main research interests include web and social media mining, data analytics, information retrieval, recommender systems, multimedia, and music information research. Markus (co-)authored more than 200 refereed conference papers and journal articles (among others, published in ACM Multimedia, RecSys, ICMR, SIGIR, ECIR, ISMIR, WEB/WWW, IEEE Visualization; Journal of Machine Learning Research, ACM Transactions on Information Systems, IEEE Transactions on Affective Computing, IEEE Multimedia, User Modeling and User-Adapted Interaction, PLOS ONE). Furthermore, he is associate editor of the Springer International Journal of Multimedia Information Retrieval and the Transactions of the International Society for Music Information Retrieval. He serves on the program committee of top-tier conferences, including WWW, RecSys, ICMR, and ACM Multimedia, and as reviewer for top-tier journals, including ACM Computing Surveys, Elsevier Pattern Recognition Letters, IEEE Transactions on Multimedia, Journal of the Association for Information Science and Technology, and ACM Transactions on Intelligent Systems and Technology.

Guillaume Gravier is senior research scientist at CNRS, France, where he is today deputy director of the Research Institute on Informatics and Random Systems (IRISA), one of the largest lab in France in the field of computer science. Over the past years, he has been leading the research group on content-based media analysis, indexing and linking, with the ultimate goal of enabling better multimedia applications and new innovative services. His current research interests are in media analytics, multimedia collection modeling, deep learning and multimodality, graph-based methods for multimedia content representation. He is the authors of 150+ publications in international multimedia venues and served as associate editor of IEEE Trans. on Multimedia. He has a long-standing involvement in community management in particular via conference and benchmark organization. He has been one of the founding members of the ISCA SIG on Speech and Language in Multimedia and of the MediaEval Community Council.