IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received 27 March 2024, accepted 18 April 2024, date of publication 1 May 2024, date of current version 9 May 2024.

Digital Object Identifier 10.1109/ACCESS.2024.3395922

RESEARCH ARTICLE

SemanticAd: A Multimodal Contextual Advertisement Framework for Online Video Streaming Platforms

BOGDAN MOCANU, (Member, IEEE), AND RUXANDRA TAPU, (Member, IEEE)
Telecommunications Department, Faculty of ETTI, National University of Science and Technology Politehnica Bucharest, 060042 Bucharest, Romania
Institute Polytechnique de Paris, Laboratoire SAMOVAR, Télécom SudParis, 91000 Paris, France

Corresponding author: Ruxandra Tapu (ruxandra.tapu@telecom-sudparis.eu)

This work was supported in part by Romanian Ministry of Research, Innovation and Digitization, Consiliul Naţional al Cercetării Ştiinţifice [National University Research Council (CNCS)]/Colegiul Consultativ pentru Cercetare-Dezvoltare şi Inovare [Advisory Board for Research, Development and Innovation (CCCDI)]—Unitatea Executiva pentru Finantarea Invatamantului Superior, a Cercetarii, Dezvoltarii si Inovarii [Executive Agency for Higher Education Research Development and Innovation Funding (UEFISCDI)], within Planul Național de Cercetare Dezvoltare și Inovare [National Research, Development and Innovation Plan (PNCDI)] III under Project PN-III-P1-1.1TE-2021-0393; and in part by the Industrial Grant Fotonation—Image Quality Experimental Methods for Data Collection for In-Cabin Sensing Technologies (4/05.12.2023).

ABSTRACT In the past few years, the online video streaming market has witnessed rapid growth and has become the most important form of entertainment. Motivated by the huge business opportunities, the advertisement insertion mechanisms have become a hot topic of research and represent the most important component of an online delivery ecosystem. In this paper, we introduce *SemanticAd*, a multimodal ad insertion framework designed from the viewers' perspective in terms of the quality of experience and degree of intrusiveness. The core of the proposed approach involves a novel temporal segmentation algorithm that extracts story units with a frame level precision. To the best of our knowledge, the proposed solution is the most robust and accurate solution dedicated to TV news videos. In addition, by taking into consideration ad temporal distribution and semantic information, the framework proposes commercials that are contextually relevant with respect to video content. The quantitative and qualitative experimental results conducted on a challenging set of 50 multimedia documents validate the *SemanticAd* methodology, returning a F1-score superior to 92%. Moreover, when compared to other state-of-the-art methods, our system demonstrates its superiority with gains in performance ranging in the [4.19%, 10.22%] interval.

INDEX TERMS Content targeted ad insertion, multimodal video analysis, story unit extraction, video temporal segmentation.

I. INTRODUCTION

In recent years, with the development of the online streaming platforms, traditional media (i.e., television, radio, or newspaper) witness their reality change and domination cease to be absolute. Video streaming is now responsible for most of the Internet traffic and is expected to increase by 5% each year [1]. On the other hand, we have witnessed the quick and consistent growth of the online advertisement market and is anticipated that the industry will reach 560-580 billion USD in 2024.

Motivated by the huge business opportunities and the semantic information included into a multimedia file, the online video advertisement field has become a hot topic of research [2], [3]. The goal is to insert the commercial clips efficiently and effectively within the multimedia documents to maximize the preferences of target consumers. However, if the commercials are displayed in inappropriate locations, are too often played or are uncorrelated with the video semantic content, the ads can easily disturb the viewer and affect their quality of experience (QoE) [4].

In practice, a wide range of techniques have been developed such as: pre-roll and post-roll commercials, in-stream ads (inserted during the video) or overlaid advertisement.

The associate editor coordinating the review of this manuscript and approving it for publication was Lei Shu.

However, most online video platforms insert the advertisement before or at the end of an image sequence. Using such strategy, the users are compelled to view it, but in most of the cases they disregard the content because of the reduced commercial relevance. In addition, from the subjective evaluation studies [5], it has been observed that ''in-stream'' video ads are more tolerated by viewers.

In the area of professional production of video content, the ads insertion points are established at fixed locations, manually specified by the producers, and transmitted to the video players (''digital cue tones''). These positions are not always optimal and can occur in inappropriate moments being considered intrusive, while affecting the viewers' semantic comprehension over the video content. Other approach adopted by the current ads insertion systems is to use text processing techniques to match the ad textual descriptors with video metadata (*i.e.*, title, keywords, or short summary). As a result, the inserted commercial is only relevant with respect to the whole multimedia document and not with the current/viewed content.

Motivated by the above observations, we argue that it is essential to develop a methodology designed to improve the viewers quality of experience, while making it interested in the proposed commercial. An efficient ad insertion system will: (1). correlate the ads and video content by targeting the appropriate marketing audience; (2). adaptively determine the total number of ads by considering the video temporal length and the total active watch duration; (3). dynamically estimate the optimal ads' locations by performing a temporal segmentation of the video stream in semantic units (*i.e.*, stories/scenes). The commercial position needs to be determined to maximize the video semantic discontinuity (*i.e.,* the beginning/end timestamps of an action sequence). Based on such strategy, the viewer can comprehend why a particular advertisement is being proposed at a specific temporal location, increasing their level of acceptability and receptiveness.

In this paper, we introduce *SemanticAd*, a multimodal ad insertion system dedicated to online video streaming platforms. Our solution is designed from the viewer's perspective, in terms of commercial contextual relevance and degree of intrusiveness. The *SemanticAd* framework dynamically establishes the most appropriate ads' locations by analysing two sources of information: the audio and visual channels. The optimal ads' locations are determined using a high-level video structuring method that temporally segments the TV news videos into story units. The system adaptively selects contextually relevant advertisement clips by matching the commercial with the video semantic content. In addition, by taking advantage of the story units' boundaries, the proposed insertion points are distributed as uniformly as possible along the video timeline. The major contributions of the paper are:

1. A novel completely automatic ad insertion framework dedicated to online video streaming platforms. The system is designed from the user perspective to ensure minimal impact on user quality of experience while dynamically estimating optimal commercial insertion points based on analysis of audio and visual stream content discontinuity. The proposed framework addresses the critical need for seamless ad integration in online video streaming, enhancing both user satisfaction and advertising effectiveness.

2. At the technical level the paper introduces a novel high-level temporal structuring method that adaptively segments the video stream into stories by adopting a multimodal fusion of information from the visual and audio channels. The proposed solution is dedicated to TV news videos. Unlike existing techniques, our approach offers frame-level precision in identifying scene boundaries within TV news videos. Notably, our system exclusively proposes ad insertion points at scene/story boundaries, ensuring a non-intrusive viewing experience. Moreover, the proposed method distributes ad locations uniformly throughout the video stream, maximizing advertising exposure while maintaining viewer engagement. To the best of our knowledge, our solution is the most robust and accurate system designed to segment such type of video genre (*i.e.*, TV news).

3. A novel technique designed to offer a comprehensive understanding of both video and ad content, with the aim of augmenting the semantic relevance of advertisements in relation to the video content. Through the deep analysis of the semantic similarity between the commercial and the multimedia document, the proposed advertisements are contextually relevant with respect to the current story unit. This approach significantly enhances the effectiveness of advertising by delivering ads that seamlessly integrate with the thematic content of the video, enhancing viewer engagement and ad recall.

In this paper we have focused our attention on a specific type of video content (*i.e.,* TV news videos) broadcasted over online streaming platforms. In this section, we present several crucial points to substantiate our choice of such video genre, elucidating its relevance to our research objectives:

1. TV news programs exhibit a highly structured format characterized by segments dedicated to news, sports, weather, and entertainment. This structured nature aligns with the content classification requirement imposed to the online video streaming platforms. The temporal segmentation of the video content into distinct categories facilitates comparative analysis and provides insights into audience engagement and preferences across different genres.

2. The dynamics of advertising placements used in TV news programming can be applied to any other type of content broadcasted by the online video streaming platforms. The key objective is to balance the user experience with the advertising revenue generation, highlighting the importance of optimizing ad placements to enhance viewer engagement and maximize the effectiveness of advertisements.

3. TV news attracts a wide range of audience, spanning from different age groups, socio-economic backgrounds, and geographical regions. Studying the audience demographics and viewing habits in TV news programming offers

valuable insights into consumer behavior and preferences. These insights are directly relevant for understanding user engagement on online video streaming platforms.

4. Previous research studies [46] and industry reports [47] have highlighted the importance of TV news content in understanding broader trends in media consumption and advertising effectiveness. In addition, both studies highlighted that TV news serves as a representative dataset for examining advertising dynamics across diverse media channels.

The rest of the paper is structured as follows. Section II reviews the technical literature dedicated to advertisement methodologies for online video platforms. Section III describes the proposed architecture with the main steps involved. Section IV introduces the experimental evaluation and discusses the results obtained on a large set of video documents. Finally, Section V provides the conclusions of the paper and opens some perspectives of further work.

II. RELATED WORK

In this section we review the related work dedicated to automatic ads insertion in online video streams. Most existent state-of-the-art techniques can be classified in two categories: user behaviour-based methods and content-based analysis techniques.

A. ADS INSERTION METHODS BASED ON USERS' BEHAVIOR

The main idea of such methods is to analyse and understand the user behaviour by performing marketing and psychological studies. Based on the users' searching, browsing, or clicking Internet history such methods determine a set of rules for inserting and displaying the ads that match the viewers' personal interest.

Rijsbergen et al. [6] proposed to use the viewers Internet search history to identify behavioral characteristics and to infer semantic similar advertisement. Wang et al. introduces in [7] a study that measures the consumers' perception over commercials for various purposes (*e.g.,* brand building) and on different media platforms (traditional and Internet-based). By understanding the user behaviour, designers and marketers can improve the advertisement strategy, increase the click rates, and enhance the effectiveness of interactive media. In [8] an empirical study over the ads click-through log (collected from a commercial search engine) is presented. After analysing the results, the following set of conclusions can be highlighted: (1). users clicking on the same ad present a similar behaviour over the Internet; (2). the click-through rate can by increased by 670% by performing a customer targeting advertising; (3). the user behaviour is better modelled using short time analysis methods. Malheiros et al. [9] conducted a study using 30 participants on a holiday booking website, each page showing ads with different degrees of personalization. The participants reported that semantic correlated ads are most likely to the noticed. However, the subjects indicated an increase level of discomfort with high personalized advertisement.

In [10] Kaytoue et al. propose analysing the number of viewers on the Twitch platforms, to predict the popularity of a live video game session, while in [11] the authors analyse various parameter of the Twitch streaming network to learn how broadcasters and viewers interact within the environment. Both papers [10], [11] focus on the viewers changing channel which is only a part of the information that can be extracted from live streaming scenarios. In [12], broadcasters and customers are interviewed to identify the personalities influencing viewers to stay more time on a specific channel.

The advertisement insertion based on viewers' behaviour is complex to design and is very difficult to develop a mathematical model appropriate for all users. In addition, such studies cannot scale for a large set of subjects and are expensive to perform. Moreover, the subject behaviour can modify in time and is influenced by various environmental factors like stress, state of mind or wellbeing.

B. ADS INSERTION METHODS BASED ON VIDEO CONTENT ANALYSIS

The ad insertion based on the text correlation is one of first approach included in the content-oriented advertisement methods. The goal of such technique is to match the metadata associated to a video document to a set of keywords used to describe the commercials. In this context, the matching algorithm has a direct impact over the ads' relevance. The most popular frameworks adopting such strategy are AdSense [13] and AdWords [14]. In [15] ten kinds of matching strategies are proposed and used to rank the ads dataset with respect to the embedded video web page content. The methods have been extended in [16] were Lacerda et al. introduced the idea of dynamic programming and defines a cost function to compute the degree of similarity between two different text descriptors.

Even though the text correlation methods are highly popular in commercial systems such frameworks suffer from a set of limitations: (1). the textual data is by far insufficient and fails to provide a precise and complete description over the information included in an image sequence; (2). the metadata associated to a video/ad document suffers from the subjectivity of the annotation process and may lead to incomplete/wrong matching; (3). the entire process is highly dependent on the matching strategy and the cost function minimization.

To address the above limitations most researchers, follow a unified rule of placing the commercial at the level of the story units. In addition, the system can dynamically establish contextually relevant ads (*i.e.,* with respect to the content of the video). Moreover, the video segmentation at different levels of granularity offers the advantage of proposing insertion points that minimize the visual and audio content discontinuity.

One of the first methods addressing the problem of video temporal segmentation in the context of online advertisement is introduced in [17]. vADeo determines ads' locations

by using the scenes boundaries. Then, the commercials are selected through a high-level analysis of the surrounding video segments. vADeo encourage the user interaction and ad click without disrupting the user viewing experience.

The VideoSense system introduced [3] is designed as a contextual video advertisement framework that automatically determines optimal locations and relevant ads. Unlike traditional systems that insert the commercials at the beginning or the end of an image sequence, the VideoSense analyses the Web page containing the online video and extracts the surrounding text. The locations are established based on the content discontinuity and level of attractiveness. Finally, the commercials are determined at the output of a multimodal relevance function.

VideoAder [18] leverage the structure of multimedia documents for embedding visual relevant ads (inserted into a set of precise locations). The framework uses content-based image retrieval techniques (*i.e.,* Laplacian of Gaussian [19] or Scale Invariant Feature Transform [20]) to identify relevant matches between the current video content and the ads dataset. Then, the ads association is formulated as an optimization function designed to maximize the total revenue of the system. Similarly, in [21] Wang et al. propose an interactive recommendation framework based on ad concept hierarchy and contextual search. The work was extended in [22] where ActiveAd is introduced. The system is designed to improve the effectiveness of advertising by combining online shopping information with the video content and directly forward the viewers to proper online shopping websites. The framework performs a video analysis to identify both syntactic and semantic elements. Then, a visual linking by search component is proposed, while the relevant query keywords are selected through a tag matching procedure.

In [5] the authors propose an object-level video advertising approach based on deep neural networks architectures. The framework aims to embed content relevant ads, representing human clothing, while minimizing the degree of intrusiveness. The matching between the ads and the video content is performed using a heuristic algorithm. In [23] an effective content-targeted method for online video advertising is proposed. The matching between the video and the ads is performed based on their associated semantic description. In addition, the scene characteristics are considered to select the optimal insertion locations. In [24] the authors propose inserting the ads at the level of the story boundaries determined using a scene segmentation algorithm relaying on the visual features extracted using multiple CNN architectures and shot grouping using an agglomerative clustering technique.

After analysing the technical literature dedicated to the subject of advertisement insertion in video streams, we can derive the following set of conclusions: (1). Most methods perform high level video structuring into scenes/story units by using only perceptual cues such as colour features, contrast measures or motion vectors. Such algorithms quickly show their limitation when confronted to the high diversity of the video content. In addition, most methods fail to identify the scenes boundaries for adjacent scenes containing similar visual patterns but are developed in different locations; (2). Most techniques can identify only the core of a story unit but fail to extract the exact scene boundaries. Consequently, the advertisement risks to be inserted in the middle of the action and can easily disturb the viewer and affect the quality of experience; (3). In some cases, the semantic correlation between the video and ad content is neglected.

To address the above drawbacks, in this paper we have introduced the *SemanticAd* architecture that performs a multimodal analysis and connects audio, visual and semantic cues to adaptively extract the story units' boundaries with a frame level precision. Our method dynamic estimates the ads' locations based on the visual and audio content discontinuity. In addition, we propose inserting the commercial by taking into account the semantic content correlation between the ads and the current video story.

III. PROPOSED METHODOLOGY

Fig. 1 presents a schematic diagram of the proposed system architecture with the main steps involved. First, the video stream is divided into a set of shots using a shot boundary detection algorithm. Then, we tackle the problem of video temporal segmentation into scenes/story units by using a multimodal approach that exploits visual and audio information. Finally, content relevant commercials are proposed using the ads and video semantic description.

A. SHOT BOUNDARY DETECTION

The shot boundary detection algorithm is based on the method introduced in [25]. In this context, each video frame is considered as a node in a graph structure connected with other vertices by edges. The weight associated to an edge indicates the degree of resemblance between two frames. In our case, we have defined the edge weight as the cosine similarity measure between the extracted visual descriptors.

We have considered as visual features the concatenation of the HSV colour histograms with the high-level characteristics extracted from the last convolutional layer of the ResNet50 [26] architecture. Then, a shot boundary can be identified by optimizing a min-max cost function computed within a temporal analysis window. The system proves to be very effective (returning average precision and recall scores superior to 95%) in detecting abrupt transitions (cuts) and gradual transitions with a temporal duration inferior to 10 frames. However, for long lasting gradual transitions the system performance decreases because of the different types of objects and video camera motion.

To address such limitation, in this article we have introduced a simple and efficient strategy that allows increasing the proposed framework robustness. For each possible transition a further analysis is performed, using a 2 second video sub-segment (centred in the current frame). On the first frame of the considered segment, we start extracting interest points using a regular grid structure. The step of the grid element

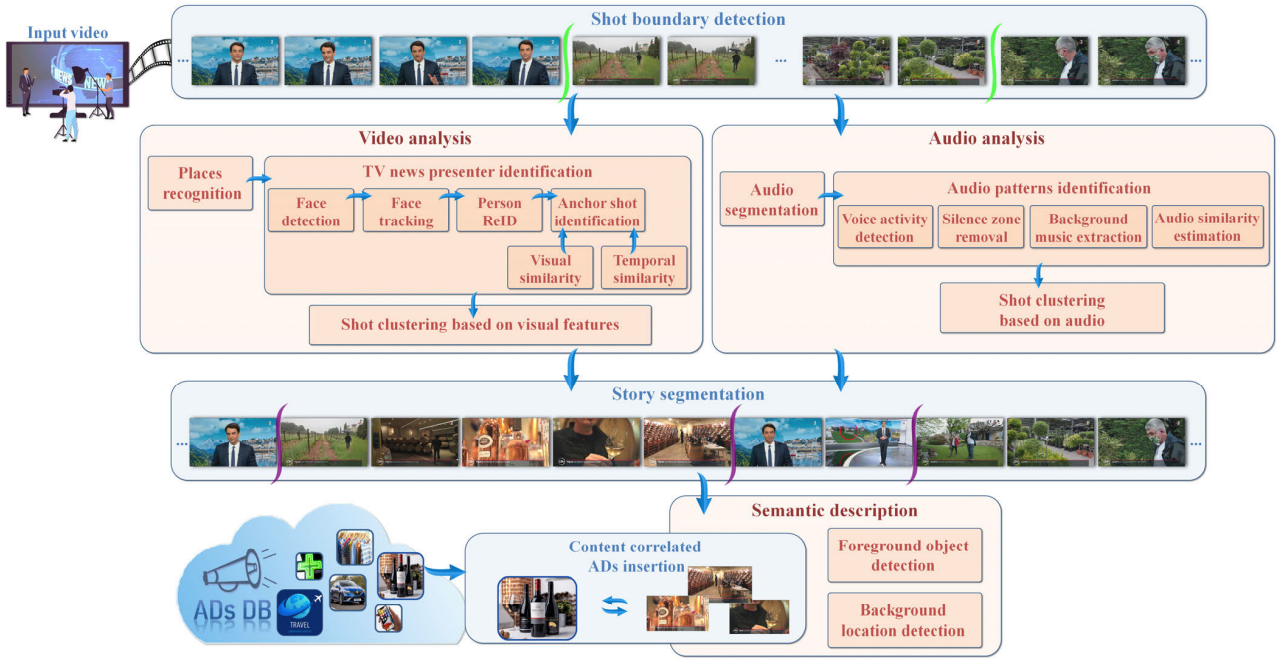

FIGURE 1. The overall architecture of the proposed SemanticAd system.

is computed as: $S = W \cdot H / no$, where W and H represent the frame size (width and height), while no denotes the maximum number of interest points. Then, the extracted points are applied as input to a multiscale Lucas-Kanade tracker [27] that determines their position within the adjacent video frames of the considered sub-segment.

The actual location of a shot boundary can be identified by analysing the remaining number of points and their distribution within the last frame of the video sub-segment. When a cut transition occurs (Fig. 2a), no interest point will remain on the last frame because the tracking cannot be correctly performed. For long gradual transitions (Fig. 2b), even if the first and last frames are completely different, because of the slow content variation, the Lucas-Kanade algorithm can still track more than half of the initial number of interest points extracted. While, for video sub-segments depicting large camera/object motion (Fig. 2c) it can be observed that the number of points will drastically decrease and will concentrate only on a specific frame region.

Using the detected shots, the next stage of our framework is to extract story units based on the visual and audio information and on a novel temporally constrained clustering technique.

B. HIGH LEVEL VIDEO SEGMENTATION INTO STORY UNITS

In this section we address the problem of video temporal segmentation into scenes for TV news content. Fig. 3 illustrates the synoptic diagram of the proposed video segmentation method.

The framework accepts as input video shots, extracted using the shot boundary detection module (*cf.* Section III-A) and determines the story units based on an agglomerative clustering technique adopting visual and audio cues.

1) PLACES RECOGNITION

Because each shot is characterized by a smooth variation of the visual content, we have decided to construct its associated static summary using a variable number of representative frames. The keyframes are selected as the images that maximize the saliency degree [28] with respect to the visual descriptors (*i.e.,* color histograms). The place recognition module is designed to determine the location where a specific shot is taking place. In the context of our application, due to the visual characteristic of the analyzed video streams (*i.e.,* TV news) we have adopted the ResNet152 [26] that has been trained on Places365 database [29]. From the considered dataset we have retained only two classes: television studio and reportage. We have trained the network to extract the background information and infer knowledge about a shot with respect to the filming setting within the studio or in different indoor/outdoor locations.

The corresponding recognition probabilities are also returned. Let us denote by Set_{studio}, the set of shots labeled as studio which can potentially contain the news presenter.

2) TV NEWS PRESENTER IDENTIFICATION

For each shot labelled as television studio we have performed the face detection, tracking and person re-identification process.

The face detection is based on Faster R-CNN [30] extended with region proposal [31]. We have used 3 scales (starting from 512 × 512 with a sub-sampling frequency of 2) and

FIGURE 2. Video segmentation into shots based on interest points' analysis: (a) CUT transition; (b) Dissolve transition; (c) A video shot exhibiting important camera motion.

three aspect ratios (2:1, 1:2 and 1:1) that translate into nine anchors for each possible location.

The faces are tracked using the ATLAS algorithm [32]. Each detected face is characterized by a set of high-level characteristics selected from the last convolutional layer of the VGG16 [33] architecture trained for person recognition tasks.

The network output is a 4096-visual face descriptor, which is further normalized to a unit vector. We denote by $F = \{f_1, f_2, \ldots, f_N\}$ a face track extracted from a video shot with N frames, where f_i denotes the face instance feature descriptor tracked in the i^{th} frame. For each frame we extract all faces with their associated descriptors (f_i) and the normalized feature representation ($\bar{f}_i$). Our objective is to develop a global, fixed length, visual facial descriptor ($S_{FACE}(F)$), constructed at the level of a track that aggregates all features:

$$S_{FACE}(F) = \sum_i weights_i \cdot \bar{f}_i; \tag{1}$$

where $\{weights_i\}_{i=1}^{N}$ is a set of weights that controls the relevance of each feature to the overall face representation ($weights_i$ are associated to the $\bar{f}_i$ descriptor). In this way, the $S_{FACE}(F)$ descriptor will have the same size regardless on the track length.

Next, we focused our attention in estimating the set of $weights_i$ that allows to distinguish between the various face poses involved.

In this context, we have trained a different convolutional neural network (*i.e.,* VGG16 [33]) with two categories denoted: *representative* and *trivial*. The *representative* class contains frontal, high-quality, unblurred and un-occluded face instances, while in the *trivial* category we have included low-quality, profile and blurred face images. In addition, the following transformations have been applied over the images from the *trivial* class: liner motion, optical blur, various types of noise and scale transformation. For training we have adopted the Multi-Task Facial Landmark (MTFL) dataset [34] and used more than 20k instances per class. The network will return as output the probability (p_i) of a face descriptor $\bar{f}_i$ to be assigned to the representative class. Finally, all the output probabilities are passes through a softmax operator so that: $\sum_i weights_i = 1$. The proposed weight adaptation module ensures the robustness of our system to the number of faces instances and to the order the data is applied as input.

Based on the observation that TV news videos are characterized by a highly structured content, we have focused our attention in identifying anchor shots (*i.e.,* video shots that are recorded in the studio and contain the presenter).

In this context, each face descriptor $S_{FACE}(F)$ is treated independently and applied as query (q) over Set_{studio}. For each query, the module returns an ordered list of shots, ranked based on the visual similarity score computed between the associated global facial descriptors. The distance between the visual descriptors is estimated using the cosine similarity metric.

For each query instance (q) we retain the top k ranked results $No_k(q)$ in order to construct a cluster of elements for which we determine the interclass similarity:

$$VisSim_{q-cluster} = \sum_{j=1}^{k} \frac{S_{FACE}(q) \cdot S_{FACE}(F_j)}{\|S_{FACE}(q)\| \cdot \|S_{FACE}(F_j)\|}, \tag{2}$$

with $F_j \in No_k(q)$.

The parameter $VisSim_{q-cluster}$ measures the degree of similarity between the various elements included in the current cluster. In the context of presenter identification in TV news videos, the clusters returning the highest score for the $VisSim_{q-cluster}$ parameter may reveal the identity of the presenter, while their elements may be marked as anchor shots. However, if multiple people are presented within the video sequences (*e.g.,* presenter, guests, reporters, etc.), such method fails to distinguish between the actual news presenter and other persons.

FIGURE 3. TV news video segmentation into story units.

In order to address such limitation, we propose to analyse the temporal distribution of the various global face descriptors within the video timeline. So, for each cluster of persons we compute not only the visual similarity based on the associated facial descriptors but also the temporal variance of shots positions as:

$$TempSim_{q-cluster} = \frac{1}{k}\sum_{j=1}^{k}\left(t_j - \mu_{q-cluster}\right)^2, \qquad (3)$$

where k denoted the number of shots, t_j is the temporal location of a studio shot within the considered cluster, while $\mu_{q-cluster}$ is the average temporal location of all shots included in the analysed cluster $(q - cluster)$.

For video shots displaying guests or other persons, the $TempSim_{q-cluster}$ will return low scores due to the compact temporal distribution.

The temporal distribution of such persons is restricted to specific time intervals, rather the entire video timeline. In contrast, the presenter will have a uniform distribution over the video content that will translate in higher values for the $TempSim_{q-cluster}$ parameter.

Each cluster will be characterized by the associated score $Score_{q-cluster}$ given by the following equation:

$$Score_{q-cluster} = VisSim_{q-cluster} \cdot TempSim_{q-cluster} \qquad (4)$$

Here, $VisSim_{q-cluster}$ is normalized to the unit vector using the maximum visual similarity score retrieved, while

$TempSim_{q-cluster}$ is normalized to the highest variance of all classes.

The cluster with the maximum value for the $Score_{q-cluster}$ parameter is labeled as the presenter cluster and all its constituent shots are marked as anchor shots.

We need to point out that the cluster containing anchors shots will not include an exhaustive group of shots showing the presenter, but a visual appearance model associated to the TV news presenter characterized by the maximum visual and temporal characteristics.

Finally, all presenter locations along the video timeline are retrieved by comparing all shots against all anchor shots (Fig. 4). If the similarity score is above a pre-established threshold (Th_1) the shot is marked as anchor shot and the presenter cluster is extended with an additional element. In our experiments we set $Th_1 = 0.85$.

3) SHOT CLUSTERING BASED ON VISUAL FEATURES

The initial segmentation of the video stream into stories can be performed by grouping video shots containing the presenter using the similarity between the associated facial descriptors. We have introduced the following agglomerative clustering technique. We group into the same cluster anchor shots only within a temporal sliding window ($window_{size}$) computed as:

$$window_{size} = \begin{cases} \sum_{m=1}^{N_{intermShots}} T_{shot_m}; & N_{intermShots} \leq 5 \\ 120sec; & N_{intermShots} > 5, \end{cases} \quad (5)$$

where T_{shot_m} represents the temporal duration of a shot expressed in seconds, while $N_{intermShots}$ is the number of intermediary shots situated between two anchor shots depicting the presenter. In our experiments we have fixed the $N_{intermShots}$ parameter to 5. Next, we have imposed a natural condition to our shots merging process: the video shots need to be temporally continuous. All intermediary shots situated between two anchor shots (assigned to the same cluster) are automatically merged into the same story unit.

However, such a strategy does not guarantee that all video shots included into different clusters will form a complete story unit. To deal with the video shots not assigned to any scenes we have introduced a shot grouping strategy based on the visual descriptors.

Each shot is characterized by the associated set of representative keyframes. For each keyframe (K) we construct the global representation based on HVS colour histogram and the high-level features extracted from the last layer before classification of a ResNet50 [26] architecture ($S_{VIS}(K)$). Using the $S_{VIS}(K)$ descriptors we can determine the visual similarity between two shots as:

$$Vis_{similarity}(s_m, s_n) =) = \max_{\substack{i=1,\ldots,K\ F(s_m) \\ j=1,\ldots,K\ F(s_n)}} \frac{S_{VIS}(K_i) \cdot S_{VIS}(K_j)}{\|S_{VIS}(K_i)\| \cdot \|S_{VIS}(K_j)\|}, \quad (6)$$

where $KF(s)$ represents the total number of keyframes included in the shot s static summary. Finally, the shots are assigned into clusters using the agglomerative clustering techniques described above (Fig. 5).

4) SHOT CLUSTERING BASED ON AUDIO FEATURES

The analysis of the audio information starts by splitting the audio signal into smaller parts (denoted *audio chunks*) using the timestamps indicated by the shot boundary detection algorithm (*cf.* Section III-A). Then, our objective is to cluster shots to the corresponding story units-based similarity of the various audio pattern involved.

Next, we have applied traditional speech processing techniques, including voice activity detection, silence zone identification and removal or background music extraction [35] on each audio chunk. We treated the shots clustering based on audio patterns identification as a multi-category classification problem. On the filtered audio chunks, we computed the image spectrograms [36], represented as tensors of size $257 \times T \times 1$, where 257 are the spectral components of a Shot Time Fourier Transform (STFT), T is the temporal length of the analysed audio segment (expressed in seconds), while 1 is the number of channels used for spectrogram representation. We performed the mean and variance normalization on each bin of the frequency spectrum.

To reidentify the various audio patterns involved, we have modified a CNN architecture (*i.e.,* ResNet34 [26]) and we have adapted it to spectrograms inputs. In addition, we have performed a batch normalization stage before passing the data through the non-linear activation function (*i.e.,* ReLUs). The system accepts as input audio signals of arbitrary lengths and returns as output a fixed size utterance descriptor. Since we treat the problem of audio pattern re-identification as an image classification problem, we have used the extracted audio descriptor ($S_{AUDIO}(a)$) to establish the audio similarity between the corresponding shots:

$$Audio_{similarity}(s_m, s_n) = max_{\substack{i=1,\ldots,A(s_m) \\ j=1,\ldots,A(s_n)}} \frac{S_{AUDIO}(a_i) \cdot S_{AUDIO}(a_j)}{\|S_{AUDIO}(a_i)\| \cdot \|S_{AUDIO}(a_j)\|}, \quad (7)$$

where $A(s)$ indicates the number of audio sub-segments extracted from a video shot.

Using the audio descriptors, we start clustering shots into story units based on the grouping strategy introduced in Section III-B.3. Finally, we analysed the two predictions regarding the story boundaries: one returned by the visual module and the second predicted by the audio framework. The resulted story units are identified based on the union of the various clusters involved. So, clusters containing at least one common shot are assigned to the same video story unit.

C. STORY UNITS' SEMANTIC DESCRIPTION

After performing the high-level temporal segmentation of the video stream into story units, the next stage of our framework is to construct the associated semantic representation. To this

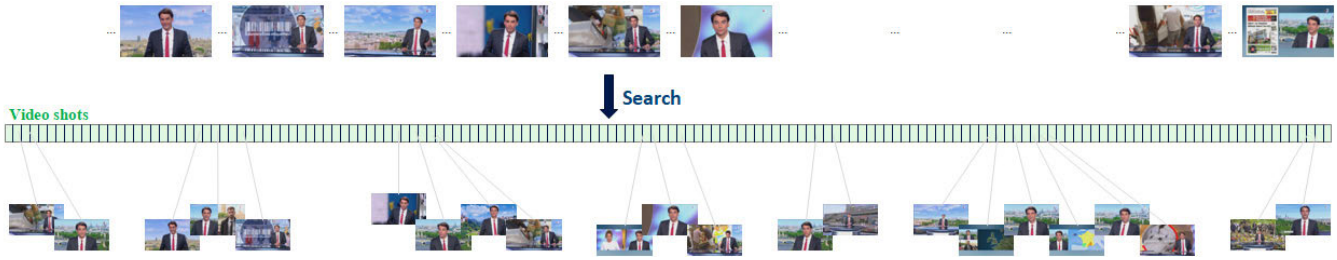

FIGURE 4. Extending the presenter clusters based on visual similarity of facial features.

FIGURE 5. Shots grouping into story units based on visual descriptors.

purpose we have trained two CNNs architectures to identify the foreground and background objects existent in the various scenes.

Because we are focused on TV news videos, we propose associating a semantic description only to the reportage story units, while avoiding studio scenes. We decided to remove studio scenes because in such cases the visual informational content is uniform in time and is unrepresentative in the context of ads insertion.

In order to identify the foreground objects, we have trained YOLOv7 [37] on COCO dataset [38]. CNN can detect and recognize the most relevant 80 classes of objects a person can interact with in an indoor/outdoor scenario. Some of the considered classes involve various types of transportation systems, animals, eating tools, fruits and vegetables, food types etc. We propose applying as input to YOLOv7 the shot keyframes (used to form the static summary) and retain the top-3 predictions.

For the background object detection, we have trained the ResNet152 architecture [26] on Places365-Standard [29] dataset to identify the most common indoor/outdoor locations. Some of the retained classes include beach, church, cafeteria, bookstore, hotel, pharmacy, harbor, river, restaurant, etc. The locations are extracted by applying as input to the network the shots keyframes for which we have retained the top-5 predictions.

Finally, a video shot is described by the associated list of objects/places, the occurrence rates, and the recognition scores, while a story unit is semantically represented as a union of its associated objects/locations from its constituent shots.

D. ADVERTISEMENT INSERTION SYSTEM

In this section, we introduce the objective parameters that are associated to the advertisement insertion module. The key principle of the advertisement location estimation is based on the temporal segmentation algorithm introduced in Section III-B. Using the visual and audio discontinuity of story units' boundaries, we argue that such locations represent good candidates as ads insertion points.

After extensive consultations/discussions with users, TV broadcasting stations, researchers, and software developers, we have collected a set of requirements and features an online advertisement system should contain in order to be accepted by the community. The identified structural elements refer mostly to the location and the number of ads to be included in a video stream. Regarding the position of the commercial, most users prefer to have ads situated at the end of a reportage section rather than at its beginning. Users indicated a higher degree of discomfort when watching ads located between a studio video segment and a reportage scene and pointed out that in some cases the two elements are semantically related. With respect to the number of ads, all participants agreed this parameter should consider the video temporal length and the total active watch duration. The *SemanticAd* framework notably aims at addressing the above observations. For each candidate location we have computed a score using the following set of parameters designed to determine the users' comprehension over the video content and their receptiveness with respect to the proposed commercials.

1) ADS TEMPORAL DISTRIBUTION

The commercials dispersion determines the ads temporal distribution over the video timeline. It is expected that ads insertion points to present a uniform distribution within the source video, while avoiding densely insert commercials in a limited number of points. For the proposed methodology, by using the temporal segmentation algorithm into story units, due to the intrinsic structure of the image sequences, the uniform ads dispersion condition is naturally fulfilled.

In order to prioritize between reportage story units, we introduce the TD_i parameter computed as: the normalized temporal durations of a reportage scene with respect to the maximum temporal duration of a story unit. The system privileges the ads insertion points that maximize the value of TD_i.

2) ADS CONTEXTUAL RELEVANCE

The contextual relevance determines the semantic correlation between the story units and the considered advertisement. To this purpose we develop the ads associated list of foreground/background objects by applying the method presented in Section III-C. Between a scene $Scene_i$ and an advertisement ad_j, the commercial relevance (CR_i) can be

computed as:

$$CR_i = \max_j \sigma\left(Scene_i, ad_j\right), \tag{8}$$

where $\sigma(\cdot)$ is a function that counts the number of common recognized objects between the j^{th} commercial and $Scene_i$. The function output is normalized by using the maximum number of common objects a scene and an advertisement contain. The system privileges ads that maximize the value of the contextual relevance (CR). For a better semantic correspondence, the CR_i parameter can be further extended to consider: the level of confidence for the detected objects, the occurrence frequency of an object and the distance between the object and the centre of the screen.

3) ADS DEGREE OF INTRUSIVENESS

To reduce to minimum, the ads degree of intrusiveness, we propose prioritizing the ads' location based on the audio stream content discontinuities. With this respect, for each candidate location we compute the silence length (SL_i) between a reportage and a studio scene. The value of SL_i is normalized to the maximum silence interval established between the analyzed story units.

The ads temporal distribution, degree of intrusiveness and contextual relevance are fused within a score function as follows:

$$AdsLocation_i = TD_i + CR_i + SL_i \tag{9}$$

Finally, the locations are sorted in descended order of relevance.

IV. EXPERIMENTAL RESULTS

This chapter presents the experimental methodology employed to determine the performance of the *SemanticAd* framework. To validate the capacity of our system, we have conducted extensive objective experiments of the proposed high-level video segmentation algorithm into story units together with the ads insertion methodology. Finally, a subjective evaluation is provided to determine the users' quality of experience after using our system.

A. THE EVALUATION DATASET

Because there is no openly accessible dataset devoted to the topic of ads insertion in TV news videos, we have stated our analysis by developing a novel database used to determine the performance of the proposed high-level video structuring algorithm. The dataset consists in 50 multimedia documents with more than 40 hours of content, with the average duration ranging between 20 minutes and 3 hours. We have included 30 documents, with the resolution of 1024 × 576 pixels, from the archive of France National Television (FTV) that includes the following shows: ''JT13 – Le journal de 13 heures'', ''JT20 - Le journal de 20 heures'', ''Télématin'' and ''Envoyé spécial'' and 20 videos selected from the YouTube platform, with the resolution of 704 × 396 pixels, that include TV news broadcasted on the ABC and CNN stations.

The video streams are highly challenging containing various types of objects and camera movement, lighting condition or scale change variation. In addition, the ''Télématin'' show is very particular because it has a temporal duration of 3 hours and contains some parts of the program that are broadcasted multiple times (*e.g.,* the news information, the weather section, the salient reportages, etc.).

The partition of the dataset into story units has been performed using five human observers. After the annotation process, the dataset can be globally characterized by 17316 video shots and 982 story units.

We acknowledge the importance of testing the robustness and generalizability of our method to ensure its effectiveness in optimizing ad placements across various types of TV news video broadcast over the online streaming platforms. In order to demonstrate the powerfulness of our temporal segmentation algorithm we have evaluated the proposed method on two publicly available datasets: BBC Rushes [48] and TRECVid [49]. BBC Rushes provides a database of raw video rushes, which are unedited footage collected during the production of TV news stories. This dataset contains a diverse collection of content, including raw footage from interviews, on-location reporting, and b-roll footage. The National Institute of Standards and Technology (NIST) organizes the TRECVid benchmark evaluation for video retrieval and analysis tasks. TRECVid datasets include a subset of TV news videos from various sources, such as CNN, ABC, and NBC, with annotations for temporal segmentation tasks.

B. THE CNN TRAINING

For the face re-identification module, we have considered the VGG face dataset [39] that contains 2.6 million images with 2622 categories of known individuals. The detected and tracked faces (*cf.* Section III-B.2) are further aligned. The spatial alignment process is based on facial landmarks, extracted as described in [40]. All face images are resized to a common resolution of 224 × 224 pixels. The training of the CNN architecture is performed using a batch size of 128 elements, with a learning rate of 0.0001 and for 50k epochs. We need to highlight that in the online stage no weights adjustment of the CNN network is allowed. For the VGG16 architecture involved in the weight adaptation module we have adopted similar values for the training parameters as for the face re-identification module.

The visual descriptors used to characterize the shot static summary, further used for shot clustering, are obtained as the concatenation of the HSV colour histograms with the high-level features extracted from the last convolutional layer of the ResNet50. The ResNet50 architecture has been trained on ImageNet database [41].

In the context of shot grouping based on audio features we have trained a modified version of the ResNet34 architecture on VoxCeleb2 dataset [42], that contains 1 million utterances extracted for 6,112 celebrities. We have imposed a minimum length for each speech segment of 2.5 seconds. The audio

elements from VoxCeleb2 are highly challenging: multiple acoustic environments are used for the recordings, with real word noise, background clutter or overlapping speech.

For the semantic story/ad representation framework we have used two CNN architectures: YOLOv7 [37] trained on COCO dataset [38] and ResNet152 [26] trained on Places365-Standard [29].

C. QUANTITATIVE SYSTEM EVALUATION

The proposed high level temporal segmentation framework has been tested on the set of 50 video documents (*cf.* Section IV-A). To evaluate the performance of video temporal segmentation algorithm, we have considered the following objective evaluation parameters: accuracy (A), recognition score (R) and F1 rate defined as follows:

$$A = \frac{TP}{TP + FP}, R = \frac{TP}{TP + FN}, F1 = \frac{2 \cdot A \cdot R}{A + R}, \quad (10)$$

where TP denotes the total number of true positive samples (*i.e.,* story boundaries that are correctly identified), FP indicates the false positive elements (*i.e.,* story boundaries that are wrongly labelled as belonging to a novel scene) and FN counts the number of false negative instances (*i.e.,* not detected story boundaries).

We have conducted ablation studies to verify the individual impact of the various components involved on the overall performance. More precisely, the main components under evaluation are the following:

1. A temporal segmentation strategy into story units based solely on the global facial features associated to the various persons existent in the multimedia document (*cf.* Section III-B.2).

2. A temporal segmentation algorithm that uses the face features as in the first testing scenario extended with the visual descriptors extracted from each keyframe of the video shot static summary (*cf.* Section III-B.3).

3. Story units' extraction using the visual features as in the second testing scenario that includes the shot classification into studio/reportage (*cf.* Section III-B.1). The clustering process is restricted to grouping into the same category only video shots belonging to one of the considered classes.

4. The complete story high-level temporal structuring framework introduced in this paper involves the multimodal clustering of the video shots using the visual and audio descriptors (*cf.* Section III-B.4).

The experimental results obtained by the *SemanticAd* system in the various testing scenarios considered are given in Table 1. After analyzing the experimental results presented in Table 1 the following set of conclusions can be derived:

1. The temporal segmentation based on facial features is very appropriate in clustering video shots focused on the presenter. It can be observed that an analysis of the global facial descriptors is suitable in identifying the core of a story unit. However, such an approach fails to extract the scene boundaries with a frame level precision. This behavior can be explained by the semantic nature of the image sequence that may contain video shots not showing the presenter's face. In addition, in some shots the facial descriptors may be affected by various external factors such as: lighting conditions, image resolution, and compression noise.

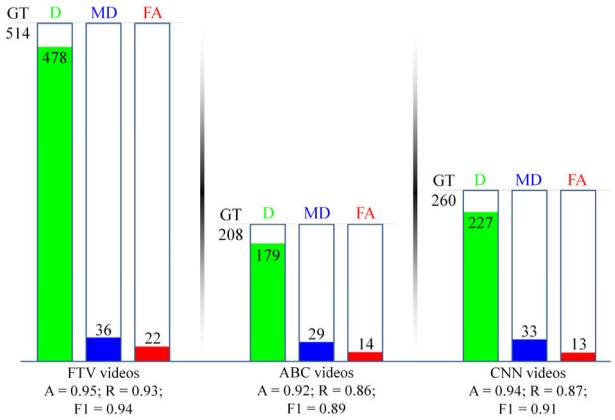

FIGURE 6. Performance evaluation of the SemanticAd temporal segmentation algorithm for various types of video content.

2. The visual descriptors extracted as the concatenation of the HSV color histograms with the high-level features from the last convolutional layer of the ResNet50 are useful in grouping video shots that share similar visual characteristics. We observe that the visual feature analysis can offer an increase in the accuracy score of more than 3%.

3. The semantic description of the video shots and their classification into studio/reportage categories can further increase the robustness of the clustering process. Such classification helps in disambiguating between adjacent story units, with low similarity scores of the visual descriptors.

4. The multimodal story unit segmentation method, introduced in this paper involves the joint analysis of the visual and audio information returns the best results (over 92% in term of F1 score). This behavior can be explained by the complexity of our method that is able to cluster to the corresponding scenes video shots that contain the same speaking personage or share a similar audio pattern.

The 982 story boundaries were further analyzed to determine the proposed framework robustness with respect to different types of content addressed. Thus, the testing dataset has been split into the following categories: France TV, ABC and CNN videos. In Fig. 6, we illustrate the experimental results obtained on each category used for testing.

Finally, we have compared the proposed framework with salient state-of-the-art temporal segmentation algorithms, addressing the problem of TV news story extraction: Ji et al. [43], Chen et al. [44] and Broilo et al. [45]. The experimental results obtained are summarized in Table 2.

From the experimental results reported in Table 2 we can observe that the proposed framework outperforms the other models, with more than 4%, regardless of the broadcasting station.

TABLE 1. Ablation study of the different attention modules involved in the *SemanticAD* framework.

Method	GT story units	Accuracy (%)	Recall (%)	F1 score (%)
(1). Video segmentation based on facial descriptors	982	87.13%	78.61%	82.65%
(2). Story units' extraction using facial and low-level visual descriptors		90.01%	81.77%	85.69%
(3). Video segmentation using visual features and shot type classification		92.12%	85.73%	88.81%
(4). Multimodal story segmentation using visual and aural features		94.74%	90.01%	92.32%

TABLE 2. Comparison between the proposed framework (*SemanticAd*) and several state-of-the-art methods on the FTV dataset.

Method	Accuracy (%)	Recall (%)	F1 score (%)
Ji et al. [43]	87.13%	78.61%	82.65%
Chen et al. [44]	90.01%	81.77%	85.69%
Broilo et al. [45]	92.12%	85.73%	88.81%
SemanticAd	94.74%	90.01%	92.32%

TABLE 3. Experimental results of *SemanticAd* framework and the considered state-of-the-art methods on the BBC rushes and TRECVid datasets.

Method	F1 score (%) BBC Rushes	F1 score (%) TRECVid
Ji et al. [43]	67.23%	69.43%
Chen et al. [44]	71.89%	74.12%
Broilo et al. [45]	75.45%	79.46%
SemanticAd	85.22%	88.69%

For comprehensive validation of the proposed framework, we conducted system evaluation using two TV news datasets: BBC Rushes and TRECVid. These datasets were chosen to facilitate a robust assessment of our approach across diverse contexts and scenarios. In addition, we have conducted a comparative analysis of our results with the techniques introduced by Ji et al. [43], Chen et al. [44], and Broilo et al. [45].

The experimental results presented in Table 3 demonstrate the performance of the proposed framework when compared with the state-of-the-art methods on both the BBC Rushes and TRECVID datasets. The SemanticAd framework showcases its performance, outperforming existing methods with notably higher F1 scores. Specifically, SemanticAd achieves an F1 score of 85.22% on the BBC Rushes dataset and 88.69% on the TRECVID dataset, showcasing its effectiveness for automatic TV news segmentation.

After analyzing the experimental results in Table 3, the following observation can be highlighted: while the video temporal segmentation plays a crucial role in the context of ads insertion, it's important to acknowledge that the performance of existing methods may vary when it comes to determine precise locations for ad insertion. This is primarily since the existing segmentation techniques may not achieve the required level of granularity, particularly at a frame-level precision.

To address this challenge, we devised a comprehensive approach that goes beyond traditional segmentation methods. For each candidate location identified, we've implemented a scoring mechanism designed to assess two key factors: users' comprehension of the video content and their receptiveness towards the proposed commercials.

In determining users' comprehension of the video content, we have considered factors such as scene transitions, silence zones, and visual coherence. These aspects contribute to the overall understanding and engagement of viewers with the video material. On the other hand, assessing users' receptiveness towards commercials involves evaluating parameters such as ads temporal distribution, contextual relevance, and degree of intrusiveness. By examining these parameters, we aim to ensure that the proposed commercials seamlessly integrate with the thematic content of the video, enhancing viewer engagement and ad recall. The proposed approach not only addresses the limitations of existing segmentation techniques but also aligns with our goal of delivering a superior user experience while optimizing advertising effectiveness.

For the advertisement insertion strategy, it is difficult to conduct an objective evaluation procedure because of the subjective nature of such a process. In the following section we present the qualitative evaluation study conducted over the proposed approach.

D. QUALITATIVE SYSTEM EVALUATION

Given the nature of the research, it is challenging to do a live test to evaluate our advertisement insertion strategy. Therefore, we have conducted a complex user-study experiment to determine the views' subjective evaluation after using the *SemanticAd* framework.

For the usability study we have considered all France TV videos except "Télématin" because of the long temporal length (more than 3 hours of broadcasting). Finally, the testing dataset contains 20 videos (*cf.* Section IV-A) with the

average duration of each element around 20 minutes. Concerning the end users, 30 anonymous volunteers (19 males and 11 females) with an average age of 23 years were asked to participate in our experiment. It is worth mentioning that all of them have previous experience in watching online videos.

We have compared the proposed advertisement insertion framework with traditional ad display strategies such as: pre-roll and mid-roll. During the evaluation, the subjects were not informed about the experimental settings and were invited to watch videos as usual. At the end, all participants were asked to score (on a scale from 1 to 5 – with 5 the highest value) each ads insertion strategy with respect to their overall satisfaction and responded to the following set of questions: (1). What is the longest time you watched the ad? (2). What is your degree of satisfaction with respect to the location where the ad has been inserted? (3). What is your degree of satisfaction with respect to the ad content? (4). From your opinion what is the acceptable number of ads to insert in TV news videos?

The analysis of results highlights the following conclusions:

1. The mid-roll ad insertion strategy received the lowest score since it interrupts the user quality of experience. The mid-roll has been found somehow disturbing, especially when the commercial is inserted at a fixed moment in time (*i.e.,* after 8 minutes of content) without considering either the visual nor the audio content.2. In the case of pre-roll, the users indicated a lower degree of intrusiveness when compared with mid-roll and found such strategy having a reduced impact on the viewers' comprehension over the video content. However, in the pre-roll setting the users tend to skip the commercials.

3. The proposed approach that dynamically establishes the ads insertion location based on the visual and audio content discontinuity has been found interesting. The user reported an increased degree of acceptability and receptiveness. In addition, the viewers have indicated that most ads are inserted during the less interesting moments of the video stream.

4. The participants preferred the proposed ads' locations with more than 75.32% over the traditional mid-rolls and 52.72% over the pre-roll insertion strategies. In addition, the users observed the semantic correlation between the ads and the video content and understood why a given commercial is proposed at a specific moment in time.

5. With respect to the number of commercials for a video stream of 20 minutes, most users indicated that the maximum number of ads should not exceed 3 elements.

As a general observation most participants considered the proposed system (*SemanticAd*) user-friendly and left the ads play until the end, without wanting to skip it. In our opinion, such a system can create more advertising opportunities for the dedicated market.

V. CONCLUSION

In this paper, we have proposed a novel completely automatic advertisement insertion methodology, denoted *SemanticAd*, designed from the viewers' perspective in terms of ad contextual relevance and degree of intrusiveness. The proposed approach is based on multimodal analysis of the audio and visual information fusion designed to provide a high-level understanding of the video stream.

From the methodological point of view, the core of the approach relies on a temporal structuring algorithm into story units, that leverage the mutually complementary nature of various features involved while maintaining the modality-specific information. The proposed system is designed to provide commercial insertion points that minimize the content discontinuity. Our approach ensures that the proposed locations are uniformly distributed along the video timeline. In addition, by analyzing the semantic similarity between the ad and the story unit, the system adaptively selects contextually relevant commercials with respect to both contents.

The experimental results conducted on a challenging set of 50 multimedia documents taken from France National Television, ABC and CNN television stations validates the *SemanticAd* methodology, which returns a F1-score superior to 92%. In addition, when compared to other methods [43], [44], [45], our methodology demonstrates its superiority with gains in performances ranging in the [4.19%, 10.22%] interval.

For further work and development, we envisage extending the proposed system to various types of video gender including movies, documentaries, TV games or talk shows. In addition, a speech to text module can further increase the semantic similarity between the ad and the video content.

REFERENCES

[1] R. Dubin, A. Dvir, O. Hadar, T. Frid, and A. Vesker, "Novel ad insertion technique for MPEG-DASH," in *Proc. 12th Annu. IEEE Consum. Commun. Netw. Conf. (CCNC)*, Jan. 2015, pp. 582–587, doi: 10.1109/CCNC.2015.7158038.
[2] W.-S. Liao, K.-T. Chen, and W. H. Hsu, "AdImage: Video advertising by image matching and ad scheduling optimization," in *Proc. 31st Annu. Int. ACM SIGIR Conf. Res. Develop. Inf. Retr.*, Jul. 2008, pp. 767–768.
[3] T. Mei, X.-S. Hua, and S. Li, "VideoSense: A contextual in-video advertising system," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 19, no. 12, pp. 1866–1879, Dec. 2009.
[4] T. Hoßfeld, M. Seufert, M. Hirth, T. Zinner, P. Tran-Gia, and R. Schatz, "Quantification of Youtube QoE via crowdsourcing," in *Proc. IEEE Int. Symp. Multimedia*, Dec. 2011, pp. 494–499, doi: 10.1109/ISM.2011.87.
[5] H. Zhang, X. Cao, J. K. L. Ho, and T. W. S. Chow, "Object-level video advertising: An optimization framework," *IEEE Trans. Ind. Informat.*, vol. 13, no. 2, pp. 520–531, Apr. 2017, doi: 10.1109/TII.2016.2605629.
[6] C. van Rijsbergen, S. Robertson, and M. Porter, "New models in probabilistic information retrieval," Brit. Library R&D Rep., Cambridge, U.K., Tech. Rep. 025.524 VAN-RIJ n, 1980.
[7] R. Wang, P. Zhang, and M. Eredita, "Understanding consumers attitude toward advertising," in *Proc. 8th Americas Conf. Inf. Syst.*, 2002, pp. 1143–1148.
[8] J. Yan, N. Liu, G. Wang, W. Zhang, Y. Jiang, and Z. Chen, "How much can behavioral targeting help online advertising?" in *Proc. 18th Int. Conf. World Wide Web*, Apr. 2009, pp. 261–270.
[9] M. Malheiros, C. Jennett, S. Patel, S. Brostoff, and M. A. Sasse, "Too close for comfort: A study of the effectiveness and acceptability of rich-media personalized advertising," in *Proc. SIGCHI Conf. Human Factors Comput. Syst.*, May 2012, pp. 579–588.
[10] M. Kaytoue, A. Silva, L. Cerf, W. Meira, and C. Raïssi, "Watch me playing, I am a professional: A first study on video game live streaming," in *Proc. 21st Int. Conf. World Wide Web*, Apr. 2012, pp. 1181–1188, doi: 10.1145/2187980.2188259.
[11] J. Deng, F. Cuadrado, G. Tyson, and S. Uhlig, "Behind the game: Exploring the twitch streaming platform," in *Proc. Int. Workshop Netw. Syst. Support Games (NetGames)*, Dec. 2015, pp. 1–6, doi: 10.1109/NetGames.2015.7382994.

[12] W. A. Hamilton, O. Garretson, and A. Kerne, "Streaming on twitch: Fostering participatory communities of play within live mixed media," in *Proc. 32nd Annu. ACM Conf. Human Factors Comput. Syst.*, Apr. 2014, pp. 1315–1324.
[13] *Google AdSense*. Accessed: Feb. 14, 2024. [Online]. Available: https://www.google.com/adsense
[14] B. Geddes, *Advanced Google AdWords*, 3rd ed. Hoboken, NJ, USA: Wiley, 2014.
[15] B. Ribeiro-Neto, M. Cristo, P. B. Golgher, and E. S. de Moura, "Impedance coupling in content-targeted advertising," in *Proc. 28th Annu. Int. ACM SIGIR Conf. Res. Develop. Inf. Retr.*, Aug. 2005, pp. 496–503.
[16] A. Lacerda, M. Cristo, M. A. Gonalves, W. Fan, N. Ziviani, and B. Ribeiro-Neto, "Learning to advertise," in *Proc. 29th Annu. Int. ACM SIGIR Conf. Res. Develop. Inf. Retr.*, 2006, pp. 549–556.
[17] S. H. Sengamedu, N. Sawant, and S. Wadhwa, "VADeo: Video advertising system," in *Proc. 15th ACM Int. Conf. Multimedia*, Augsburg, Germany, Sep. 2007, pp. 455–456.
[18] R. Hong, L. Tang, J. Hu, G. Li, and J.-G. Jiang, "Advertising object in web videos," *Neurocomputing*, vol. 119, pp. 118–124, Nov. 2013.
[19] T. Lindeberg, "Feature detection with automatic scale selection," *Int. J. Comput. Vis.*, vol. 30, no. 2, pp. 79–116, Nov. 1998.
[20] D. G. Lowe, "Object recognition from local scale-invariant features," in *Proc. 7th IEEE Int. Conf. Comput. Vis.*, vol. 2, Jul. 1999, pp. 1150–1157, doi: 10.1109/ICCV.1999.790410.
[21] J. Wang, B. Wang, L.-Y. Duan, Q. Tian, and H. Lu, "Interactive ads recommendation with contextual search on product topic space," *Multimedia Tools Appl.*, vol. 70, no. 2, pp. 799–820, May 2014.
[22] J. Wang, M. Xu, H. Lu, and I. Burnett, "ActiveAd: A novel framework of linking ad videos to online products," *Neurocomputing*, vol. 185, pp. 82–92, Apr. 2016.
[23] G. Wang, L. Zhuo, J. Li, D. Ren, and J. Zhang, "An efficient method of content-targeted online video advertising," *J. Vis. Commun. Image Represent.*, vol. 50, pp. 40–48, Jan. 2018.
[24] B. MOCANU, R. TAPU, and T. ZAHARIA, "Content targeted ad insertion in temporal structured video documents," in *Proc. IEEE Int. Conf. Consum. Electron. (ICCE)*, Jan. 2021, pp. 1–4, doi: 10.1109/ICCE50685.2021.9427664.
[25] R. Tapu, T. Zaharia, and F. Prêteux, "A scale-space filtering-based shot detection algorithm," in *Proc. IEEE 26th Conv. Electr. Electron. Engineers Isr.*, Nov. 2010, pp. 000919–000923.
[26] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 770–778.
[27] B. Lucas and T. Kanade, "An iterative technique of image registration and its application to stereo," in *Proc. ICAI*, 1981, pp. 1–9.
[28] R. Tapu and T. Zaharia, "Salient object detection based on spatiotemporal attention models," in *Proc. IEEE Int. Conf. Consum. Electron. (ICCE)*, Jan. 2013, pp. 39–42.
[29] B. Zhou, A. Lapedriza, A. Khosla, A. Oliva, and A. Torralba, "Places: A 10 million image database for scene recognition," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 40, no. 6, pp. 1452–1464, Jun. 2018.
[30] S. Ren, K. He, R. Girshick, and J. Sun, "Faster R-CNN: Towards real-time object detection with region proposal networks," in *Proc. Adv. Neural Inf. Process. Syst.*, 2015, pp. 1137–1149.
[31] H. Jiang and E. Learned-Miller, "Face detection with the faster R-CNN," in *Proc. 12th IEEE Int. Conf. Autom. Face Gesture Recognit. (FG)*, May 2017, pp. 650–657, doi: 10.1109/FG.2017.82.
[32] B. Mocanu, R. Tapu, and T. Zaharia, "Single object tracking using offline trained deep regression networks," in *Proc. 7th Int. Conf. Image Process. Theory, Tools Appl. (IPTA)*, Nov. 2017, pp. 1–6, doi: 10.1109/IPTA.2017.8310091.
[33] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," in *Proc. ICLR*, 2015, pp. 1–14.
[34] W.-T. Chu and Y.-H. Liu, "Thermal facial landmark detection by deep multi-task learning," in *Proc. IEEE 21st Int. Workshop Multimedia Signal Process. (MMSP)*, Sep. 2019, pp. 94–108.
[35] R. Yin, H. Bredin, and C. Barras, "Speaker change detection in broadcast TV using bidirectional long short-term memory networks," in *Proc. Interspeech*, Aug. 2017, pp. 3827–3831.
[36] W. Xie, A. Nagrani, J. S. Chung, and A. Zisserman, "Utterance-level aggregation for speaker recognition in the wild," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, May 2019, pp. 5791–5795.
[37] C.-Y. Wang, A. Bochkovskiy, and H.-Y. Mark Liao, "YOLOv7: Trainable bag-of-freebies sets new state-of-the-art for real-time object detectors," 2022, *arXiv:2207.02696*.
[38] Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollar, and C. L. Zitnick, "Microsoft COCO: Common objects in context," in *Proc. ECCV*, 2014, pp. 740–755.
[39] O. M. Parkhi, A. Vedaldi, and A. Zisserman, "Deep face recognition," in *Proc. Brit. Mach. Vis. Conf.*, vol. 1, 2015, p. 6.
[40] H. Fan and E. Zhou, "Approaching human level facial landmark localization by deep learning," *Image Vis. Comput.*, vol. 47, pp. 27–35, Mar. 2016.
[41] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "ImageNet: A large-scale hierarchical image database," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2009, pp. 248–255.
[42] J. S. Chung, A. Nagrani, and A. Zisserman, "VoxCeleb2: Deep speaker recognition," in *Proc. Interspeech*, Sep. 2018, pp. 1086–1090.
[43] P. Ji, L. Cao, X. Zhang, L. Zhang, and W. Wu, "News videos anchor person detection by shot clustering," *Neurocomputing*, vol. 123, pp. 86–99, Jan. 2014.
[44] D. M. Chen, P. Vajda, S. S. Tsai, M. Daneshi, M. C. Yu, H. Chen, A. F. Araujo, and B. Girod, "Analysis of visual similarity in news videos with robust and memory-efficient image retrieval," in *Proc. IEEE Int. Conf. Multimedia Expo. Workshops (ICMEW)*, Jul. 2013, pp. 1–6.
[45] M. Broilo, A. Basso, and F. G. B. De Natale, "Unsupervised anchorpersons differentiation in news video," in *Proc. 9th Int. Workshop Content-Based Multimedia Indexing (CBMI)*, Jun. 2011, pp. 115–120.
[46] P. De Pelsmacker, M. Geuens, and P. Anckaert, "Media context and advertising effectiveness: The role of context appreciation and context/Ad similarity," *J. Advertising*, vol. 31, no. 2, pp. 49–61, Jun. 2002.
[47] Y. Wang, "Analyzing advertising effectiveness across media channels: A case study of TV news and social media platforms," in *Proc. IEEE Int. Conf. Marketing Advertising (ICMA)*, Aug. 2021, pp. 1–6.
[48] A. Gupta, D. Chugh, Anjum, and R. Katarya, "Automated news summarization using transformers," 2021, *arXiv:2108.01064*.
[49] A. F. Smeaton, P. Over, and W. Kraaij, "TRECVID—An overview," in *Proc. TRECVID*, 2006, pp. 1–40.

BOGDAN MOCANU (Member, IEEE) received the B.S. degree in electronics, telecommunications, and information technology and the Ph.D. degree in electronics and telecommunication from the University Politehnica of Bucharest, Romania, in 2008 and 2011, respectively, and the Ph.D. degree in informatics from University Paris VI–Université Pierre et Marie Currie, Paris, France. Since 2012, he has been a Researcher with the ARTEMIS Department, Institut Mines-Telecom, Télécom SudParis, France. His research interests include computer application technology, such as 3D model compression and algorithm analysis in image processing.

RUXANDRA TAPU (Member, IEEE) received the B.S. degree in electronics, telecommunications and information technology (valedictorian) and the Ph.D. degree in electronics and telecommunication from the University Politehnica of Bucharest, Romania, in 2008 and 2011, respectively, and the Ph.D. degree (Hons.) in informatics from University Paris VI-Université Pierre et Marie Currie Paris, France. Since 2012, she has been a Senior Researcher with the ARTEMIS Department, Institut Mines-Telecom, Télécom SudParis, France. Her research interests include content-based video indexing and retrieval, pattern recognition, and machine learning techniques.