High-quality Activity-Level Video Advertising

Mu Yuan, Lan Zhang, Zhengtao Wu, Daren Zheng
University of Science and Technology of China
ym0813@mail.ustc.edu.cn, zhanglan@ustc.edu.cn, wzt@mail.ustc.edu.cn, zdr123@mail.ustc.edu.cn

***Abstract*—Online video advertising is a billion dollar business, but the current low CTR reveals the huge potential for improvement in the ad serving quality. In this work, we present a novel activity-level video advertising system named ActVA. Different from existing systems that assume a fixed scope of ad keywords, ActVA enables advertising targeted to non-predefined activities in a highly efficient way requiring no training data for diverse activities. To achieve this goal, a general and extensible graphical representation of both video content and advertising demand is proposed to embed multimodal content at the activity level. Our ad-content relevancy measurement can achieve 10,000 FPS retrieval speed. We model the ads assigning task as an optimization problem taking content relevance, ads revenue as well as viewer experience into consideration. A non-maximal suppression based algorithm is designed to significantly reduce the algorithm complexity for online ad serving. Our extensive objective and subjective experimental results show the effectiveness and efficiency of ActVA. ActVA can effectively uncovers numerous high-quality (content-relevant) advertising opportunities and delivers ads to viewers in a profitable and user-friendly way.**

***Index Terms*—Online video advertising, High-quality Advertising, Activity recognition, Graphical model**

I. Introduction

Nowadays, watching videos online has become the daily activity of almost everyone. As an example, the statistics results show that the number of YouTube users has reached 1.58 billion [23], and on average 5 billion videos were watched on YouTube per day in 2018 [19]. Programmatic video advertising is grow at an unprecedented speed and has became the dominate source of revenue for online video ad publisher. Market research company eMarketer reported that YouTube generated 3.36 billion dollars by video advertising in 2018 [5]. Common video ad formats include video ads before, during or after the main video and semi-transparent overlay ads. Though these days have witnessed the improvement of ads quality, e.g., attractive movie-like video ads, the click-through rate (CTR) of video ads is much lower than that of other types of ads. According to the report of AdStage, in the first-quarter of 2019 the average click-through rate (CTR) on YouTube was **0.3%** [1], meanwhile the Google AdWords CTR is around 5% [8]. Therefore, there is still a large improvement potential for video advertising in terms of ad opportunity quantity and ad serving quality. Since users have a high probability to ignore ads irrelevant to the content they are watching, Since users have a higher probability to ignore ads irrelevant to the content they are watching, an intuitive direction is to improve the relevance between the inserted ads and the video content.

Previous work [18], [29]–[31] explore different approaches to support content-based video advertising. The content of the video is represented in a variety of ways, including objects [30], affective factors [29] and human poses [31]. Though those work achieve effective advertising for the specific types of content they focused on, there are still several critical challenges hindering the quality improvement of video advertising.

- **Highly diverse activities in the video.** When measuring relevance between the video content and ads, most previous approaches focus on object/person recognition and matching but ignore the activities of persons in the video. According to our market research in Section III-A, 72.5% advertising demands are relevant to activities. For instance, a jewelry company may want to insert ads of wedding ring into video shots of marriage proposal. Activity-level video advertising requires comprehensive understanding of the video. Facing massive video data and highly diverse activities, traditional deep neural network based activity recognition methods [7], [24], [25] are inefficient due to they heavy depends on extensive training samples and expensive computation.
- **Unpredictable advertising demands.** Existing systems generate a set of pre-defined labels for videos using pre-trained models. In real-world advertising systems, advertisers may have a great variety of advertising demands, which are highly complex and hard to predict. Facing each newly arrived non-predefined demand, for example the untrained activity "making a marriage proposal", a straight-forward solution that collects training data first, then retrains and re-executes recognition models is obliviously infeasible.
- **Viewer experience facing multiple advertisements.** A series of methods [18], [30], [31] formalize the video ad assignment task as an optimization problem and attempt to improve the viewer experience from different perspectives. [18] proposes to insert ads into the main video where the attractiveness of the current content is relatively low. [30] regards the number of detection objects as the measurement of intrusive perception. Most of them consider the insertion of each advertisement separately, however, the intrusiveness effect among successive advertisements has not been studied adequately, which is significant for improving the overall user experience.

In this work, we propose a novel activity-level video adver-

Lan Zhang is the corresponding author.

Part of work was done during Zhengtao Wu's internship at Tencent Marketing Solution.

This paper has been accepted for publication at IWQoS'20. This version is a preprint identical in content to the version to be published.

tising system named **ActVA** to deal with the aforementioned challenges. The contributions of this work are summarized as follows:

1) To best of our knowledge, we propose the first activity-level video advertising system, which enables advertising targeted to non-predefined activities in a highly efficient way requiring no training data for diverse activities. To achieve this goal, we design a multi-modal information based key frame detection method, a graphical activity-level multimodal content embedding approach and the corresponding ad-content relevancy measurement.
2) We formalize the ad serving procedure as an optimization problem, taking the content relevance, the potential revenue and the intrusiveness to the viewers as the optimization objectives. For the intrusiveness, we take the effect of multiple assigned advertisements into consideration. An efficient non-maximum suppression based heuristic algorithm is designed to solve this challenging task. Based on the proposed advertising algorithm, ActVA maximizes the possible ads revenue while ensuring the intrusiveness to the viewers is under a certain constraint.
3) Both objective and subjective evaluations are conducted extensively to evaluate our proposed video advertising system, and the experimental results show the effectiveness and efficiency of ActVA. ActVA can be adopted to create numerous content-relevant video ad opportunities for flexible advertising demands to attractive viewers in a more friendly way.

The rest of this paper is organized as follows. Section II reviews previous work of two related areas, content-target video advertising and activity recognition. An overview of the proposed system is introduced in Section III. Section IV demonstrates detailed design and implementation of activity-level video advertising. Evaluation results are shown in Section V and finally Section VI concludes this paper.

II. Related Work

To the best of our knowledge, there is no previous work directly addressing the activity-level video advertising problem. Our work is related to the following two areas: content-targeted video advertising and activity recognition. Our graphical representation in this work is inspired by the notion of scene graph, and we also briefly introduce it in the end of this section.

A. Content-Targeted Video Advertising

Content-targeted video advertising aims to place advertisements based on the analysis of the video content and [17] provides a broad survey of this area. [18] proposes a video advertising system that takes the content discontinuity, attractiveness and content-ad relevance into consideration when inserting ads into videos. By assessing emotional impact of video content and advertisements, [29] models the content relevance in a computational effective way. [30] attempts to support object-level advertising and utilizes deep neural networks to extract semantic features including the faces, human bodies and objects. [31] focuses on clothing advertisements and proposes a deep learning framework consists of human body detection, pose selection, cloth localization and ads retrieval. [4] explores the combination of the personalized product recommendation and the video advertising, by utilizing the correlation among the users, videos and products. These efforts have greatly enriched the opportunities for video advertising as well as improved the ad serving quality. However, most of them assume that a fixed set of semantic labels is predefined. They cannot deal with the diverse activities in videos or non-predefined advertising demands. For the intrusiveness, they usually ignore the effect of multiple successive ads. In this work, we attempt to elevate the video advertising to the activity level which will create a large number of attractive ad opportunities. Our system aims to improve the quality of video ad serving in terms of flexible targeting, high content relevancy and friendly view experience.

B. Activity Recognition

Recognizing activities in images and videos has attracted a lot of attention from industry and academia in the machine learning field. The main objective of most work is exploring the spatiotemporal representation abilities of different modeling approaches. [13] explores to use the Fisher vector for encoding dense trajectory feature and proposes a SVM-based classification model. Recent work focus on learning spatiotemporal structure of activities based on the convolutional neural networks (CNN). [24] extends the dimension of convolutional layers from 2D to 3D and utilize the 3D-CNN model to learn the spatiotemporal structure of activities. A two-stream CNN is proposed by [7], which models the temporal information by applying CNN to the optical flow and fuses spatiotemporal features at convolutional layers. In order to obtain the long-range temporal structure, [25] propose the temporal segment networks, which combines a sparse temporal sampling method and a CNN-based spatial feature extractor. Two-stream inflated 3D model is proposed in [3] which fuses the previous technologies and achieves very high recognition accuracy.

The proliferation of activity recognition lays the foundation for this work to study activity-level video advertising. Unfortunately there are still many challenges impeding us to apply these technologies. First, to train such a recognition model requires a large amount of labeled videos, which would be infeasible for the advertising application due to the variety of advertising demands. Second, the recognition ability of a pretrained model is fixed and non-trivial to extend. Moreover, the high computation cost incur a serious challenge for online activity-level video advertising.

C. Scene Graph

Scene graph is proposed by [12], which is utilized to model the semantic relationships and attributes of objects in images. A conditional random fields model is proposed to retrieve images using scene graphs as queries. Many applications have been benefited from the scene graph representation, including description-based image generation [11], image caption auto-evaluation [2], few-shot action recognition [9] and event detection in videos [22]. The recent work [28] proposes to use RNNs to automatically generate scene graphs in an end-to-end manner. Scene graphs provide a powerful method to model the detailed structural semantics of objects in images. The principles of the scene graph inspire us to represent activities in a graphical way.

III. System Overview

A. Market Research and Motivation

In order to better understand the video advertising demands, we conducted a market research according to the real business requirements in a well-known online-video company. 150 types of content-targeted advertising demands were collected from the advertisers that distributed across 11 main domains, i.e. vehicle, food and beverage, cosmetic products, clothing, jewelry, electronic products, household products, estate, medicine, gaming and network services. We summarized the interested semantic information of advertisers into five categories:

1) **Celebrity identity:** Many companies, especially those who have their own spokespersons, tend to target their ads to where those celebrities appear.
2) **Object:** A lot of videos are about real-life scenes, so many companies hope to insert their ads when relevant objects occur in the videos.
3) **Scene of the video frames:** Some product advertisements are suitable for certain scenes. For example, a video about the gym may attract viewers who are interested in fitness equipment and online gym courses.
4) **Human behavior:** Another type of essential but usually ignored semantic information is the behaviors performed by people in the videos. For example, the jewelry company or the wedding dress brand would like to insert ads in video frames where a man is making a proposal to his girlfriend.
5) **Subtitles keywords:** The subtitles of videos also contain rich information which is a good complement to the visual data. For example, when people in the videos are talking about "buying a car", it would be a good point to insert an advertisement of a certain car brand.

Those advertising demands require a comprehensive understanding of online videos. Traditional solutions need a set of pre-trained models to generate as comprehensive labels as possible, which is inefficient due to the heavy demands for both labelled training data and computation resources, as well as non-predefined advertising demands. Furthermore, among the 150 types of advertising demands, 109 of them (72.6%) require information of people's behaviors, which cannot be satisfied with currently mature business solutions. The market research results show an urgent demand to elevate the video advertising system from object-level to activity-level.

B. System Framework

To effectively meet the continuously arriving highly diverse advertising demands, we propose an activity-level video advertising system, ActVA. Figure 1 illustrates the framework of ActVA, which consists of the multimodal content embedding module and the activity-level video advertising module.

Multimodal Content Embedding. The system inputs include the video data, subtitle files (which is optional) and the *advertising query*. The reason of leveraging multimodal data for content embedding is to satisfy the diverse activity-level requirements of advertisers. Neither video frames nor subtitles can cover all the requisite semantics. To enable the advertiser to freely express their demands for non-predefined activities, in our system, we define the advertising query as a combination of sample images containing the targeted activities and interested keywords of subtitles. For the video and subtitle data, a preprocessing step is detecting key frames from the entire videos to reduce computation complexity for relevancy calculation and ads assignment. Then each video data will be represented as a set of key frames and the corresponding subtitle text. To represent the content in video as well as to support ad-content relevancy measuring, atom features of both the detected key frames and advertising queries will be extracted. Here, we use the concept **atom feature** to represent the interested visual and textual features. According to the market research results above, five types of visual and textual atom features are considered to support comprehensive video understanding:

- **Face identity:** In order to identify the celebrities, we need the features of human faces. Face detection models could locate the faces and the encoding model embeds the face regions into a fixed-length feature vector [14].
- **Object class:** The features of objects are extracted by a pre-trained object detector [20]. We use the bounding box location, object class and the classification confidence as the features of detected objects.
- **Scene class:** For the scene feature, we deploy the CNN-based scene classifier [32] to classify the scene of images. The scene features of each image are embedded as the top-K scene classes and the corresponding classification confidence.
- **Human behavior:** Human behavior is too various to pre-train an action classifier for modeling it. So instead of deploying action recognition models, we utilize human pose estimation technologies [6], [27] to extract the human body keypoints from video frames. Figure 4 illustrates some examples of human body keypoints. The pose features of each person are represented as the location and confidence of body keypoints.
- **Subtitle keyword:** To model the textural features, we extract the keywords from subtitles and find their synonyms by natural language processing techniques [16].

The atom features describe the multimodal content in the video from different perspectives and they are semantically related. So simply concatenating these feature vectors will

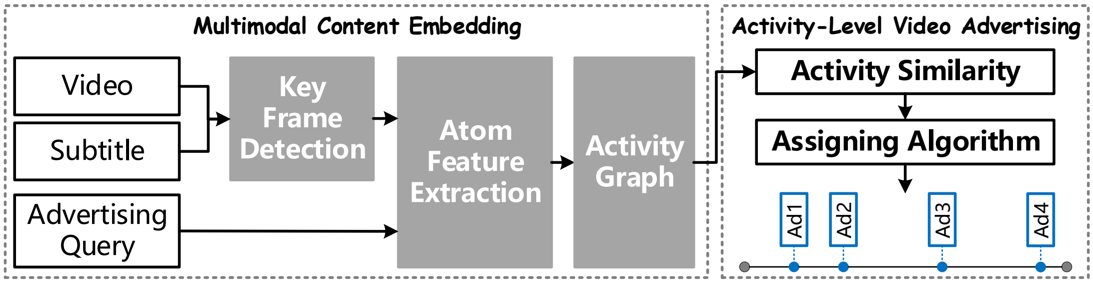


Fig. 1. Framework of the proposed activity-level video advertising system ActVA. The three gray blocks are performed offline and the two white blocks are performed online.

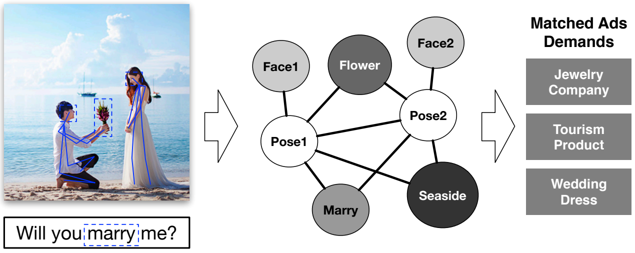


Fig. 2. An example of activity graph.

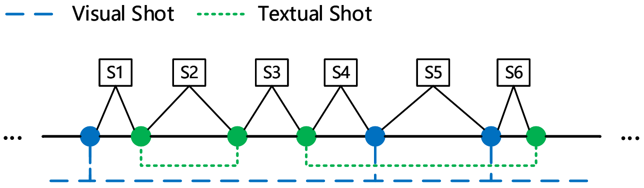


Fig. 3. Illustration of key frame detection: blue nodes are boundaries of visual shots and green nodes refer to the start and end of textual shots. S1-S6 denote the final segments.

lead to the loss of their semantic connections, which play an important role in activity recognition. For example, a person who is driving a car and a model sitting next to a car are totally different activities. Some existing work [9], [11], [12], [15], [22], [26], [28] explore graph-based methods to model the semantic relationships among objects (and persons) in images. For online video advertising, due to the great variety of unpredictable ad demands, only analyzing the person and objects information is not enough to understand the complex activities. CNN-based activity recognition methods [7], [24], [25] need adequate training samples and expensive computation, which impedes adopting them to the video advertising task. Facing these challenges, we propose a graphical structure, named **activity graph (ActGraph)**, to model the correlations among multimodal activity-level visual and textual atom features instead of only considering relationships among objects. Figure 2 illustrates an example of how the ActGraph represents the content. We will give details of ActGraph construction in Section IV-C. ActGraph is able to conveniently express rich activity-level content and contains comprehensive information for future recognition of diverse activities.

Activity-Level Video Advertising. Based on the ActGraphs of both videos and advertising queries, we propose a method to measure their relevancy at the activity level. We formulate the task, assigning advertisements to key frames, as an optimization problem, which aims to maximize the ads revenue while ensuring a good user experience. By giving a highly efficient solution to this optimization problem, our system assigns a set of advertisements to the videos to be displayed. Note that, although we stipulate the format of advertising queries, the formats of displayed ads are flexible, like video, image, inside/outside layout and so on.

IV. Design of ActVA

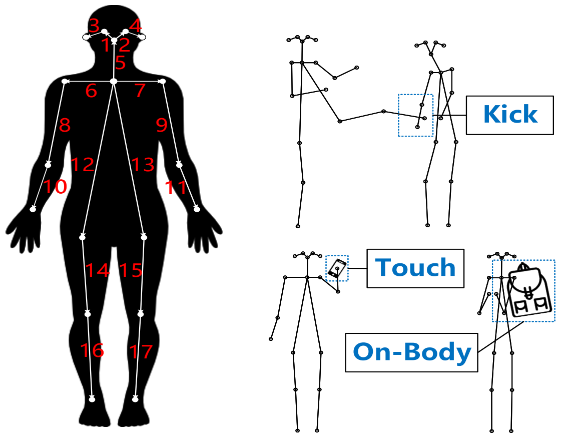
A. Key Frame Detection

Analyzing videos frame by frame is infeasible and unnecessary due to the expensive computation and content redundancy. The objective of the key frame detection is to detect frames that contain rich visual and textual information, which are able to satisfy more possible advertising demands. Based on existing work, we can segment a video into several *visual shots* in many ways [21]. In this work, we propose to utilize the textual information from the subtitles to help determine the key frames. For the visual information, we use the color histogram feature and calculate the correlation between adjacent frames to determine *visual shots*. If the correlation value is lower than a predefined threshold, the two frames will be set as the boundaries of two shots. Common subtitle files organize subtitles by the start and end time, so we can divide the videos into multiple *textual shots*. As Figure 3 shows, through adopting the boundaries of both visual and textual shots, the entire video is divided into semantic segments. Then we select k ($k = 1$ in our implementation) key frames uniformly within each segment.

B. Atom Feature Extraction

As we introduced in Section III, five types of visual and textual atom features need to be extracted for activity-level content embedding. Here we give the specific implementation of each atom feature extraction.

(1) Face detection and encoding The Dlib toolkit [14] is used for detecting and encoding human faces in images. The



(a) Human body keypoints encoding. **(b) Interaction samples.**

Fig. 4. Pose-related illustration.

output consists of the bounding box, the fixed-length (128-length in the Dlib implementation) encoding vector and the model confidence [1]. Let F denote the set of face encoding and model confidence.

(2) Object detection We use the Yolo-V3 [20] model to detect objects in images. We define O as the set of object features which composes of the discrete class index (totally 80 classes in our implementation) and the corresponding confidence.

(3) Scene classification We use the ResNet50 [10] model which is trained on the Places-365 [32] dataset as the scene classifier. Let S denote the scene features of an image which contains the top-K (K=5 in our implementation) classification results and the model confidence.

(4) Human pose encoding AlphaPose [6], [27] model is used to extract human body keypoints in images. The original location information is not suitable for representing the human behaviors, due to the variability of size and angle. We propose to encode the locations (the estimated x-y values) of body keypoints (18 keypoints in our implementation) by computing the vector angles of the skeletons shown in Figure 4(a). The 17 arrows refer to the skeleton vectors and the cosine values of calculated angles are used as the encoding. Let P denote the pose information in an image and it consists of the pose encoding and the corresponding confidence of each keypoint.

(5) Keyword extraction For analyzing the subtitles, we first extract keywords from the text and then utilized the WordNet in NLTK [16] to obtain their synonyms. We define K as the textual feature of an image and it composes of the keywords and their synonyms.

Note that, the keypoint locations and bounding boxes of faces and objects are only used for analyzing the correlations among atom features, which will be explained in the next subsection. In our implementation, we select the above five types of atom features to characterize the activity-level content, but the proposed framework is flexible and can be easily extended for other types of atom features.

C. Activity Graph Construction

We propose a graph structure to model the semantic correlations among atom features, named activity graph (ActGarph). To support matching between diverse video content and flexible advertising demands, in an ActGraph, we consider four types of semantic correlations, which are analyzed as follows.

(1) Pose-pose interaction. We first analyze the interactions between human poses. Based on the human pose encoding, for an image we obtain a set of pose features P. For $\forall p_i, p_j \in P$, we define a set of interactions I_{pp}. Let $R_{pp} : P \times P \rightarrow I_{pp}$ denote the mapping function which classifies a pair of poses as one type of interaction. In our implementation, we define five classes of interactions: face-to, touch, kick, close-to and away-from, which could be classified based on the spatial locations of body keypoints of two persons. Figure 4(b) shows an example of pose-pose interaction.

(2) Pose-object interaction. The interactions between human and objects could also be estimated by the spatial relations between object bounding boxes and human body keypoints. Based on the human pose encoding and object detection results, we have the pose set P and the object set O. Similarly, we define a set of pose-object interactions I_{po}. Then the mapping function is formulated as $R_{po} : P \times O \rightarrow I_{po}$. In our implementation, five different interactions are considered: on-face, on-body, touch, kick and away-from.

(3) Pose-face matching. The interactions between human poses and faces only have two possibilities: matched or unmatched. Formally, we use R_{pf} denote the mapping function that maps a pair of pose and face into a binary value. In our implementation, we use the condition whether the eyes, nose and ears keypoints fall in the bounding box of the face to determine the pose-face interactions.

(4) Pose-scene correlation. Not only the human behaviors effect the activity, but also the correlation between the person and the scene matters. The visual proportion of a person significantly effect his/her importance. We propose to use the visual proportion of the detected human poses in an image as the correlation between the poses and the scene. And we use function R_{ps} to output the value of the pose-scene correlation.

We organize the atom features as nodes and the above four relations as edges between nodes. Then we obtain a graph structure $G(V, E)$ where $V = \{F, O, S, P, K\}$ and the edges are analyzed by the mapping functions $R_{pp}, R_{po}, R_{pf}, R_{ps}$. The activity graph contains rich visual and textual information and is designed to be flexible and extensible, which enables video advertising supporting non-predefined activities.

D. Similarity/Relevancy Calculation.

ActGraph provides a graph-based embedding for key frames of online video data and ad queries. The final objective of our system is to serve ads to advertising positions of most relevant key frames. So the next step is to measure the

[1] The value that measures how much confidence the model holds for its classification or detection results

similarity/relevancy between the activity graphs of ad queries and key frames. We design the similarity measurement to serve two main tasks of the video ad publisher:

- **Advertising position inquiry:** when an advertiser makes a demand, the video ad publisher needs to inquiry the amount of matched ad key frames (ad opportunities) in all video data. The objective is to find as many relevant key frames as possible for the specific advertising demands.
- **Online ad assignment:** when a viewer actually starts to play a video, the video ad publisher needs to determine which ads should be assigned to the ad positions in this video in real time. The objective is to assign the most relevant ads to the video while ensuring the click through rate and experience of the viewer.

Different from typical relevancy metrics, our proposed similarity measurement $s : G \times G \rightarrow \mathbb{R}$ does not have the symmetry, i.e. $s(G_i, G_j) \not\equiv s(G_j, G_i)$. A high similarity value of $s(G_i, G_j)$ indicates that G_j includes the content similar with G_i. For example, given an image with very simple content G_i and another image with rich content G_j including the simple one, then $s(G_i, G_j)$ is high while $s(G_j, G_i)$ could be very low. The detail of the similarity function is demonstrated as follows.

(1) Atom-feature (node) similarity. We divide the five types of atom features into two embedding forms: numerical encoding with confidence and categorical encoding with confidence. The face and pose features are the first type embedding while the object, scene and keyword features are the second type embedding. Let enc denote the set of encoding and $conf$ denote the corresponding model confidence. We propose to use the weighted Euclidean norm metric to calculate the distance of the numerical encoding. Given two nodes $x_1 = (enc_1, conf_1)$ and $x_2 = (enc_2, conf_2)$, the similarity is measured by:

$$s_1(x_1, x_2) = 1 - \frac{1}{|enc_1|} \sum_{\substack{e_i \in enc_1 \\ c_i \in conf_1}} \min_{\substack{e_j \in enc_2 \\ c_j \in conf_2}} \frac{\theta_1 ||e_i - e_j||_2}{(1 + \min(c_i, c_j))} \tag{1}$$

We use the reciprocal of $1 + \min(c_i, c_j)$ as the weight of each encoding pair, which could filter the low-confidence but similar features. And for the information encoded by category and model confidence, we apply a discrete metric ρ (if $x = y$, $\rho(x, y) = 1$, otherwise $\rho(x, y) = 0$) to calculate the similarity. Let cls denote the set of categories and $conf$ denote the mode confidence. Given $x_1 = (cls_1, conf_1)$ and $x_2 = (cls_2, conf_2)$, the similarity is calculated by:

$$s_2(x_1, x_2) = \frac{1}{|cls_1|} \sum_{\substack{e_i \in cls_1 \\ c_i \in conf_1}} \max_{\substack{e_j \in cls_2 \\ c_j \in conf_2}} (\theta_2 \min(c_i, c_j) \rho(e_i, e_j)) \tag{2}$$

The θ_1 and θ_2 are parameters used to balance the effects of encoding and confidence. For a pair of activity graphs, the proposed similarity functions produce five different similarity scores for faces, object, scenes, poses and keywords respectively.

(2) Interaction (edge) similarity. In order to analyze the similarity of interactions, we first match the atom features by a threshold-based approach. For each atom feature in x_1, we can find the most similar atom feature in x_2. Then we can determine whether these two features matched or not by defining a threshold of similarity. The unmatched nodes will be dismissed and we can efficiently measure the similarity of the edges attached to remaining nodes. Similarly, we divide the four types of interactions into two forms: numerical and categorical. The pose-scene correlation is numerical while the other three interactions are represented by categorical values. For any edge e_i in $G1$, if the connected nodes of e_i have matched nodes in G_2, i.e. the potentially matched edge e_j in G_2 exists, then we calculate thei similarity as follows:

$$s_3(e_i, e_j) = \begin{cases} 1 - ||e_i - e_j||_2, & \text{numerical} \\ \rho(e_i, e_j), & \text{categorical} \end{cases} \tag{3}$$

In this way, we obtain the similarity scores of pose-pose, pose-object, pose-face and pose-scene interactions.

Combining node similarity and edge similarity, given two graphs, we have a nine-dimension similarity measurement function $s : G \times G \rightarrow \mathbb{R}$ and the advertiser can flexibly adjust the weights for different dimensions according to his/her preference. For example, some advertisers may think the celebrity is more important while some advertisers may consider the behavior as a more important factor.

E. Ad Assignment Problem Formulation

Given a video, let $X = \{x_i\}_{i=1}^{N_x}$ denote the set of N_x embedded key frames. And each key frame has its time stamp t_i where $1 \leq i \leq N_x$. Suppose we have N_y embedded advertising demands: $Y = \{y_i\}_{i=1}^{N_y}$. Each advertising demand has a revenue value r_i where $1 \leq i \leq N_y$. A pair of a key frame and an ad demand refers to an ad assignment and let $P = \{(x_i, y_j) | x_i \in X, y_j \in Y\}$ denote the universe set of all possible ad assignments. Note that, for one key frame, multiple ads can be assigned to it. Based on the proposed similarity function, we define the evaluation function $f : 2^P \rightarrow \mathbb{R}$ as follows:

$$f(S) = \sum_{(x_i, y_j) \in S} r_j s(x_i, y_j) \tag{4}$$

where $S \subseteq P$. Assigning ads not only brings revenue, but also may cause discomforts to the viewers. To model the user experience, a function g is designed to measure the perception of intrusiveness brought by the ad assignments. Previous work [18], [29], [30] explore different task-specific ways to model perception of viewers. But they treat each ad separately without considering the effect among multiple assigned ads. In this work, we design the intrusiveness function based on the following principles to improve the user experience globally:

- Frequency: the viewers would feel more uncomfortable if two successive ads are displayed within a short interval.
- Relevancy: the relevancy between an ad and the video content effects the duration of the intrusive perception.
- Duplication: duplicated ads would cause bad user experience.

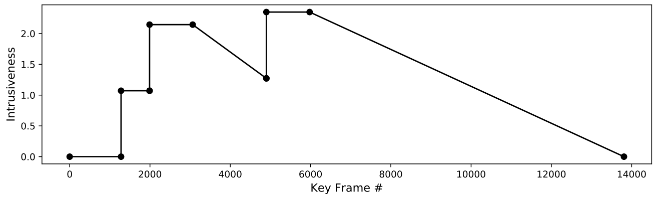


Fig. 5. Intrusiveness distribution examples.

Given a time-ordered list of N_s ad assignments $S = \{p_i\}_{i=1}^{N_s} \subseteq P$, the intrusiveness is modeled as a function of time t. Assume that the video starts at $t = 0$ and ends at $t = t_{max}$. Then the intrusiveness distribution is modeled in a piece-wise manner. From the beginning of the video till the first ad assignment, the intrusiveness is zero.

$$g(t|S) = 0 \quad ,0 \le t \le t_1 \tag{5}$$

For the following ad assignments p_i that starts at t_i. If $t_{i+1} \le t_i + \frac{\beta}{s(p_i)}$:

$$g(t|S) = g(t_i) + \frac{\alpha}{s(p_i)} \quad ,t_i < t \le t_{i+1} \tag{6}$$

Let $t^* = t_i + \frac{\beta}{s(p_i)} + \frac{g(t_i)+\alpha/s(p_i)}{\gamma}$. If $t_i + \frac{\beta}{s(p_i)} < t_{i+1} \le t^*$:

$$g(t|S) = \begin{cases} g(t_i) + \frac{\alpha}{s(p_i)} & , \\ & t_i < t \le t_i + \frac{\beta}{s(p_i)} \\ g(t_i) + \frac{\alpha}{s(p_i)} & -\gamma(t - t_i - \frac{\beta}{s(p_i)}), \\ & t_i + \frac{\beta}{s(p_i)} \le t < t_{i+1} \end{cases} \tag{7}$$

If $t^* < t_{i+1}$:

$$g(t|S) = \begin{cases} g(t_i) + \frac{\alpha}{s(p_i)} & , \\ & t_i < t \le t_i + \frac{\beta}{s(p_i)} \\ g(t_i) + \frac{\alpha}{s(p_i)} & -\gamma(t - t_i - \frac{\beta}{s(p_i)}), \\ & t_i + \frac{\beta}{s(p_i)} \le t \le t^* \\ 0, & t^* < t \le t_{i+1} \end{cases} \tag{8}$$

We believe the incremental intrusiveness brought by the ad assignment p_i should be inversely proportional to the similarity $s(p_i)$. And the duration of high intrusive perception is also inversely proportional to the similarity score. As the time goes on, the intrusiveness will decline and γ controls the descent speed. The three non-negative parameters α, β, γ control the effect of the mentioned factors. Figure 5 shows an example of intrusiveness distribution, given $\alpha = 5$, $\beta = 2000$ and $\gamma = 0.0004$. Finally, our objective is to maximize the evaluation function under the constraint of intrusiveness to the viewers, which is formalized as follows:

$$\begin{aligned} &\max_{S \subseteq P} f(S) + \delta E(S) \\ \text{s.t.} \quad &\int_0^{t_{max}} g(t|S)\mathrm{d}t \le B \end{aligned} \tag{9}$$

where B denotes the intrusiveness budget. $E(S)$ is an entropy-like function that measures balance of served ad demand distribution:

$$E(S) = -\sum_{(x_i,y_j)\in S} p_{y_j} \log_2 \frac{p_{y_j}}{N_y} \tag{10}$$

where p_{y_j} indicates the number of times the ad is served and δ controls the influence.

Algorithm 1 NMS-Greedy algorithm for Equation 9.

```
Require: key frame set $X$, ad set $Y$, intrusiveness budget $B$
Ensure: assignments $S$
Initialize the value matrix $M_{N_y \times N_x}$, where $m_{ij} = r_i s(x_j, y_i)$.
Apply the kernel $K_{1_{1 \times N_x}}$ to $M$.
Apply the kernel $K_{2_{N_y \times N_k}}$ to $M$.
Greedily select assignments with $\arg\max m_{ij}$ into $S$ until $\int_0^{t_{max}} g(t|S)\mathrm{d}t$ exceeds the intrusiveness budget $B$.
return $S$
```

F. Solution Algorithm

The optimization problem (Equation 9) has $2^{N_x N_y}$ possible solutions in total. In a real advertising system, the searching space for this problem is too large to applying enumeration-based algorithm. According to the design of intrusiveness function, we propose a non-maximum suppression based algorithm. Given a set of elements and a measurement of value of each element, non-maximum suppression keeps the maximal value unchanged while suppresses the value of the other elements. For our ad assignment task, we can calculate the value of each assignment and organize them as a value matrix $M_{N_y \times N_x} = (m_{ij})$ where $m_{ij} = r_i s(x_j, y_i)$. In the Algorithm 1, we apply two non-maximum suppression kernels to the value matrix.

- $1 \times N_x$ kernel K_1: In order to keep the balance of the assignments of each ad, we apply K_1 to keep the maximal assignment value of each ad and suppress the others;
- $N_y \times N_k$ kernel K_2: Since successive ads within short interval would cause high intrusiveness to the viewers, K_2 is applied to suppress the assignments except the one with the maximal value within the k neighboring key frames.

Figure 6 visualizes the non-maximum suppression process. After applying these two kernels to the value matrix, we greedily select assignments until the intrusiveness exceeds the budget. It is possible that Algorithm 1 achieves a locally optimal solution, but the computation complexity is greatly reduced to $N_x N_y$.

V. Experiments

In this section, we conduct both objective and subjective evaluations on diverse video data to show the efficacy and efficiency of our design.

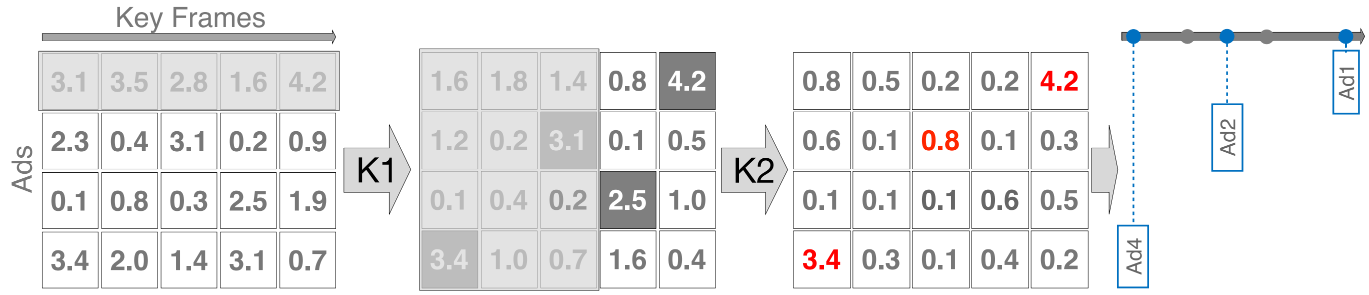
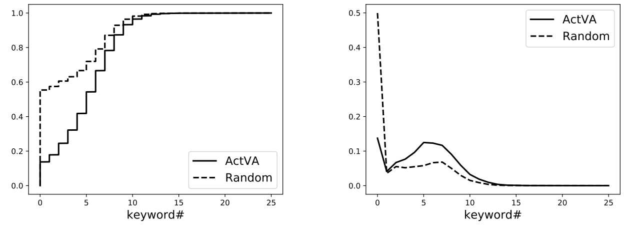


Fig. 6. Illustration of non-maximum suppression.



(a) Keyword# CDF. **(b) Keyword# distribution.**

Fig. 7. Comparison of the number of keywords extracted from each key frame selected by ActVA and the random method.

A. Dataset

Existing video advertising and image retrieval datasets are unable to meet our evaluation needs, since the proposed activity graph integrates both visual and textual features. Advertisers are interested in the keywords that characters talk about in the videos, which can be extracted from the subtitle data instead of image description words. So we collected a dataset of 140 videos and the matched subtitle files, including 100 trending videos from YouTube, 10 movies and 30 tv-series videos. In activity level, totally 150 advertising demands are collected from a various categories, including vehicle, food and beverage, cosmetic product, clothing, jewelry, electronic product, household product, estate, medicine, gaming and network services. It is worth reemphasizing that there is no classification of activities in our system and each advertising demand is represented as a set of images and selected keywords.

B. Key Frame Detection

Instead of randomly selecting key frames within each visual shots, we propose to utilize the subtitle to help segment each visual shot into multiple textual shots, in order to detect key frames with both rich visual and textual information. To evaluate the effectiveness of the proposed method, we analyze the number of different keywords of each selected frame. We compare the results with a random approach which randomly select K frames within each visual shot while ensuring that the number of total key frames selected by two methods are the same. As shown in Figure 7, more than 50% key frames selected by the random algorithm do not contain any textual information, while around 60% key frames detected by ActVA contain more than 5 different keywords. At the same time, the introduced textual shot computation has little effect on detection efficiency. Running on a machine with four Inter Core i7-5930K CPU, the key frame detection process achieves 500 to 600 FPS, which is almost the same as the random approach.

C. Advertising Position Retrieval

To truly simulate the various demands in advertising, we provide 150 different demands for testers to freely choose from, to simulate the demand arrivals in the real-world scenario. For each ad demand q, we calculate the similarity score between the demand and every key frame. The similarity calculation process is efficient which achieves 10000 to 13000 FPS for the 150 different demands. We set a similarity threshold (3.0 in our implementation) to filter out the irrelevant results and ask for people to select “Good” or “Bad” if they think the retrieved key frames are activity-level related with the demand or not. Setting the similarity threshold is a subjective task, since the visual relatedness of activities has no absolute definition. By analyzing the experimental results of adjusting the similarity threshold, a good threshold is in the range of 2.0 to 4.0 for the 150 demands. Figure 8 shows some samples of retrieved key frames and we can see the results are quite satisfactory. We invited 48 people for the subjective evaluation and in all collected 30204 ratings, consists of 22748 “Good” and 7456 “Bad” (**75.3%** Good rate).

D. Video Advertising

Our final objective is to assign advertisements with high revenue and activity-level relevancy while ensuring the good user experience at the same time. The intrusive perception of viewers is modeled in Section IV which is effected by three parameters α, β and γ. For the following experiments, $\alpha = 5$, $\beta = 2000$ and $\gamma = 0.0004$ and Figure 5 shows an example of intrusiveness distribution. The setting of these parameters is empirical and does affect the experimental results. But modeling intrusiveness of audiences is inevitably subjective, manually finetuning some parameters should be regarded as necessary labor work. For the formalized optimization problem in Equation 9, we test four different ads assignment algorithms:

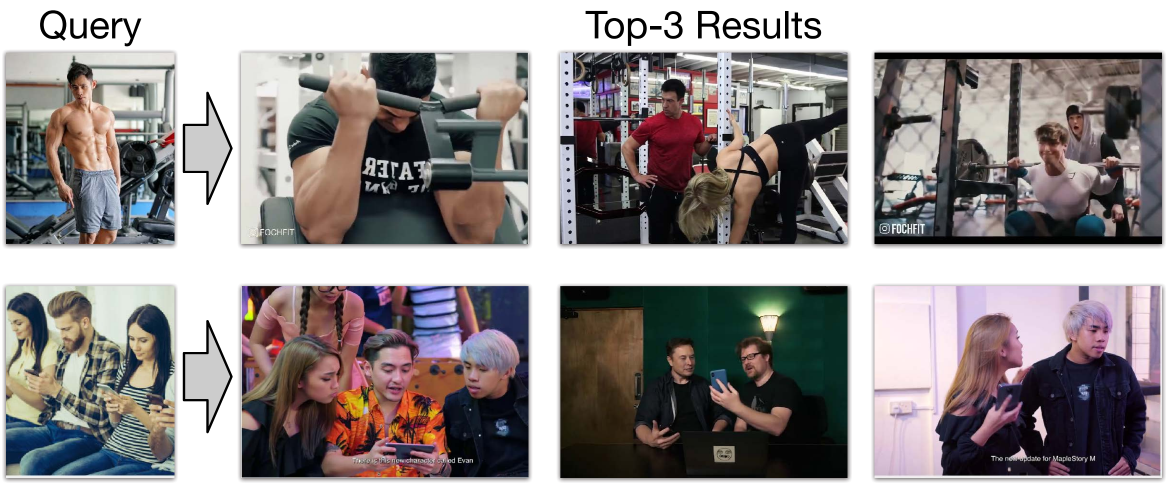
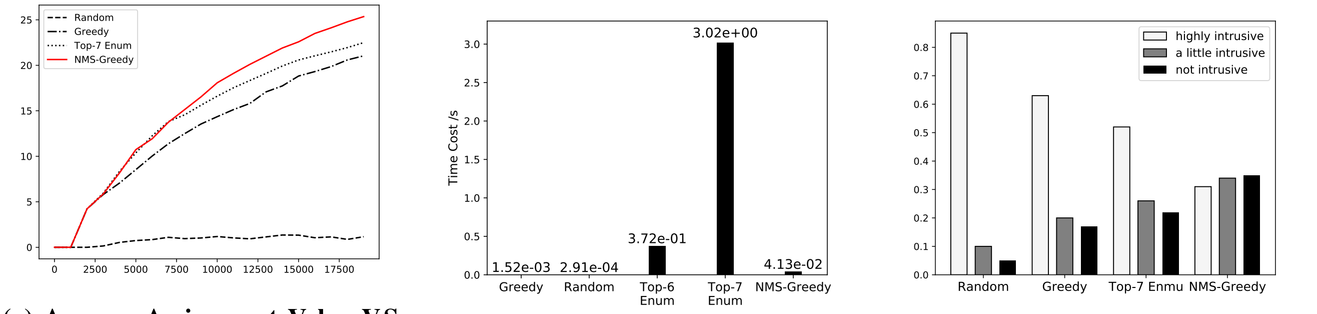

Query
Top-3 Results
FOCHFIT
FOCHFIT
There is this new character called Evan
The next update for MapleStory M

Fig. 8. Advertising position retrieval samples.



(a) Average Assignment Value V.S. Intrusiveness Budget

(b) Average Time Cost

(c) Subject Intrusiveness Evaluation

Fig. 9. The effectiveness and efficiency of the different assignment algorithms.

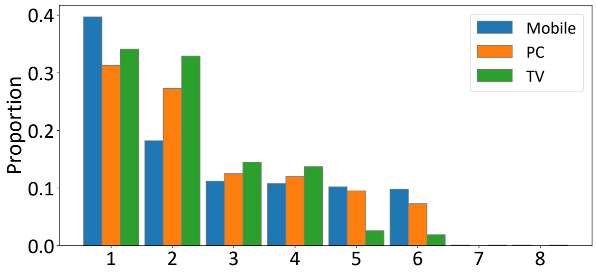


Fig. 10. The number of actual ad positions in each online video on different platforms.

- **Random** Randomly assign ads to the key frames until the intrusiveness exceeds the budget.
- **Greedy** After calculating similarity score of every pair of advertising query and key frame, the greedy algorithm select the assignment with maximal score at each iteration.
- **Top-K Enum** This algorithm enumerating all combinations of assignments with top-K similarity score and select the solution with maximal evaluation value and the intrusiveness is under the budget.
- **NMS-Greedy** our algorithm explained in Algorithm 1.

We evaluate the four algorithms on the 140 videos and Figure 9.(a) shows the average assignment value of different algorithms. When the intrusiveness budget is relatively low ($<$ 3000 in our implementation), the effectiveness of the proposed NMS-Greedy is not obvious since the number of possibly selected assignments is around one and two. With the increased intrusiveness budget, the NMS-Greedy outperforms the others significantly when there are about 3 more assignments to each video. We conduct a large-scale analysis (of tens of thousands of online videos) to count the number of actual ad positions in each online video in today's popular online video websites. Figure 10 presents the statistical result and shows that for more than 30% videos there are at least three ad positions. When considering activity-level video advertising in the future, the number of ad positions will increase significantly, which stresses the importance of our assignment algorithm. Since the ads assigning process should be finished online, the time cost is an important factor. From Figure 9.(b) we can see that NMS-Greedy algorithm is comparable to the greedy algorithm on execution speed while

Top-K Enum is too expensive to utilize when K is larger than 6. And we did a subjective evaluation which shows videos with assigned advertisements by the four algorithms and asks the viewers to rate it as ”highly intrusive”, ”a little intrusive” or ”not intrusive”. We asked 48 people to watch the videos with assigned advertisements and give their ratings. Figure 9.(c) shows the evaluation results of the four algorithms. We can see that the NMS-Greedy algorithm also performs best with more than 30% ratings are “not intrusive”.

VI. Conclusion

In this paper, we present an activity-level video advertising system, which provides a flexible video content representation and assigns advertisements with high relevance in an highly efficient and user-friendly way. A graphical model is proposed for embedding multimodal content of videos and demands and measuring ad-content relevancy. The ad serving process is formalized as an optimization task which takes ads revenue, content relevance and user experience into consideration. We also design a non-maximal suppression based heuristic algorithm to address the optimal assignment problem with low computation cost. However, designing an assignment algorithm with theoretical performance guarantee is still a very challenging task, which is our future work. We believe our system can bring a large number high-quality video ads to advertisers and viewers, thus a significant revenue growth to publishers with low computation cost.

VII. Acknowledgements

This research is supported by the National Key R&D Program of China 2017YFB1003003/2018YFB0803400, NSF China under Grants No.61822209, 61932016, 61751211, China National Funds for Distinguished Young Scientists with No.61625205, the Fundamental Research Funds for the Central Universities. Part of this work is supported by the Tencent Marketing Solution’s “Advanced Technology of Computing Advertising” project.

References

[1] Adstage, 2019. [Online]. Available: https://blog.adstage.io/youtube-benchmarks-cpc-cpm-and-ctr-q1-2019-archive

[2] P. Anderson, B. Fernando, M. Johnson, and S. Gould, “Spice: Semantic propositional image caption evaluation,” in *European Conference on Computer Vision*. Springer, 2016, pp. 382–398.

[3] J. Carreira and A. Zisserman, “Quo vadis, action recognition? a new model and the kinetics dataset,” in *proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017, pp. 6299–6308.

[4] Z.-Q. Cheng, Y. Liu, X. Wu, and X.-S. Hua, “Video ecommerce: Towards online video advertising,” in *Proceedings of the 24th ACM international conference on Multimedia*. ACM, 2016, pp. 1365–1374.

[5] eMarketer, 2018. [Online]. Available: https://www.emarketer.com/

[6] H.-S. Fang, S. Xie, Y.-W. Tai, and C. Lu, “RMPE: Regional multi-person pose estimation,” in *ICCV*, 2017.

[7] C. Feichtenhofer, A. Pinz, and A. Zisserman, “Convolutional two-stream network fusion for video action recognition,” in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 1933–1941.

[8] Google, 2019. [Online]. Available: https://instapage.com/blog/google-ads-industry-benchmarks

[9] M. Guo, E. Chou, D.-A. Huang, S. Song, S. Yeung, and L. Fei-Fei, “Neural graph matching networks for fewshot 3d action recognition,” in *Proceedings of the European Conference on Computer Vision (ECCV)*, 2018, pp. 653–669.

[10] K. He, X. Zhang, S. Ren, and J. Sun, “Deep residual learning for image recognition,” in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 770–778.

[11] J. Johnson, A. Gupta, and L. Fei-Fei, “Image generation from scene graphs,” in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 1219–1228.

[12] J. Johnson, R. Krishna, M. Stark, L.-J. Li, D. Shamma, M. Bernstein, and L. Fei-Fei, “Image retrieval using scene graphs,” in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2015, pp. 3668–3678.

[13] V. Kantorov and I. Laptev, “Efficient feature extraction, encoding and classification for action recognition,” in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2014, pp. 2593–2600.

[14] D. E. King, “Dlib-ml: A machine learning toolkit,” *Journal of Machine Learning Research*, vol. 10, pp. 1755–1758, 2009.

[15] H. Koppula and A. Saxena, “Learning spatio-temporal structure from rgb-d videos for human activity detection and anticipation,” in *International conference on machine learning*, 2013, pp. 792–800.

[16] E. Loper and S. Bird, “Nltk: The natural language toolkit,” in *Proceedings of the ACL-02 Workshop on Effective Tools and Methodologies for Teaching Natural Language Processing and Computational Linguistics - Volume 1*, ser. ETMTNLP ’02. Stroudsburg, PA, USA: Association for Computational Linguistics, 2002, pp. 63–70. [Online]. Available: https://doi.org/10.3115/1118108.1118117

[17] T. Mei and X.-S. Hua, “Contextual internet multimedia advertising,” *Proceedings of the IEEE*, vol. 98, no. 8, pp. 1416–1433, 2010.

[18] T. Mei, X.-S. Hua, L. Yang, and S. Li, “Videosense: towards effective online video advertising,” in *Proceedings of the 15th ACM international conference on Multimedia*. ACM, 2007, pp. 1075–1084.

[19] Omnicore, 2018. [Online]. Available: www.omnicoreagency.com

[20] J. Redmon and A. Farhadi, “Yolov3: An incremental improvement,” *arXiv*, 2018.

[21] A. SenGupta, D. M. Thounaojam, K. M. Singh, and S. Roy, “Video shot boundary detection: A review,” in *2015 IEEE International Conference on Electrical, Computer and Communication Technologies (ICECCT)*. IEEE, 2015, pp. 1–6.

[22] D. Shao, Y. Xiong, Y. Zhao, Q. Huang, Y. Qiao, and D. Lin, “Find and focus: Retrieve and localize video events with natural language queries,” in *Proceedings of the European Conference on Computer Vision (ECCV)*, 2018, pp. 200–216.

[23] Statista, 2018. [Online]. Available: www.statista.com

[24] D. Tran, L. Bourdev, R. Fergus, L. Torresani, and M. Paluri, “Learning spatiotemporal features with 3d convolutional networks,” in *Proceedings of the IEEE international conference on computer vision*, 2015, pp. 4489–4497.

[25] L. Wang, Y. Xiong, Z. Wang, Y. Qiao, D. Lin, X. Tang, and L. Van Gool, “Temporal segment networks: Towards good practices for deep action recognition,” in *European conference on computer vision*. Springer, 2016, pp. 20–36.

[26] X. Wang and A. Gupta, “Videos as space-time region graphs,” in *Proceedings of the European Conference on Computer Vision (ECCV)*, 2018, pp. 399–417.

[27] Y. Xiu, J. Li, H. Wang, Y. Fang, and C. Lu, “Pose Flow: Efficient online pose tracking,” in *BMVC*, 2018.

[28] D. Xu, Y. Zhu, C. B. Choy, and L. Fei-Fei, “Scene graph generation by iterative message passing,” in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017, pp. 5410–5419.

[29] K. Yadati, H. Katti, and M. Kankanhalli, “Cavva: Computational affective video-in-video advertising,” *IEEE Transactions on Multimedia*, vol. 16, no. 1, pp. 15–23, 2014.

[30] H. Zhang, X. Cao, J. K. Ho, and T. W. Chow, “Object-level video advertising: an optimization framework,” *IEEE Transactions on Industrial Informatics*, vol. 13, no. 2, pp. 520–531, 2017.

[31] H. Zhang, Y. Ji, W. Huang, and L. Liu, “Sitcom-star-based clothing retrieval for video advertising: a deep learning framework,” *Neural Computing and Applications*, pp. 1–20, 2018.

[32] B. Zhou, A. Lapedriza, A. Khosla, A. Oliva, and A. Torralba, “Places: A 10 million image database for scene recognition,” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2017.