Object-Level Video Advertising: An Optimization Framework

Haijun Zhang, *Member, IEEE*, Xiong Cao, John K. L. Ho, and Tommy W. S. Chow, *Senior Member, IEEE*

Abstract—**In this paper, we present new models and algorithms for object-level video advertising. A framework that aims to embed content-relevant ads within a video stream is investigated in this context. First, a comprehensive optimization model is designed to minimize intrusiveness to viewers when ads are inserted in a video. For human clothing advertising, we design a deep Convolutional Neural Network (CNN) using face features to recognize human genders in a video stream. Human parts alignment is then implemented to extract human part features that are used for clothing retrieval. Second, we develop a heuristic algorithm to solve the proposed optimization problem. For comparison, we also employ the Genetic Algorithm (GA) to find solutions approaching the global optimum. Our novel framework is examined in various types of videos. Experimental results demonstrate the effectiveness of the proposed method for object-level video advertising.**

Index Terms —**video advertising, content-based, object-level, optimization, in-video ads.**

I. Introduction

RECENT years have witnessed the rapid and consistently increasing popularity of online advertising amongst advertisers and publishers, coinciding with the advent of social media and the ubiquity of the Internet. The huge potential business opportunities existing in the online advertising market have attracted increasing interest in research into developing new advertising models for media, e.g., Google AdSense[1], YouTube overlay video ads[2], and Yahoo! Video[3]. With the advancement of computer technology, online video has become one of the most commonly used network services. Watching video online has moved from a niche activity to the mainstream, and Internet video, in particular, has experienced truly massive growth. According to Cisco [1], online video traffic will globally increase to 55% of all consumer Internet traffic in 2016. This growing revenue mainly came from online video advertising. The forms of online video ads are quite diverse, including flash ads, rich media ads, keyword ads and in-video ads.

One of the key objectives of video online advertising is to compose a relevant ad appealing to meet the interests of targeted customers at the right time and the right place on the screen, avoiding intrusiveness to the receivers who are not interested. In an effort to achieve this, It has been common for ads to be displayed at the beginning or the end of a video. Receivers are forced to view the ad, but they usually ignore it due to its irrelevance to their interests. In order to improve the attractiveness and lower the intrusiveness of ads to receivers, ads can be inserted at the time of playing a video, and the relevance of ads to the video content of the receivers' interests can be assured. In particular, it is desirable to make the products advertised in ads relevant to the objects occurring in a video stream. For example, a receiver may be a fan of movie star, Julia Roberts. When the receiver is watching the movie, *Notting Hill*, he or she may be attracted by a skirt worn by Julia Roberts. At that time, if an ad advertising a similar skirt pops up onto the screen, the receiver should be more possible to be attracted by the ad, instead of feeling frustrated by watching it at the beginning or in the middle of the video.

Most current studies of online video advertising primarily work on locating ads in an appropriate place in a sports video, personalized ads delivery in Interactive Digital Television (IDTV), text-based video advertising, and video segment level advertising. The earliest relevant work in advertising focused on finding an unimportant region that would not affect the theme of the scene in a sports video [7]. The region was then replaced with a product advertisement. These regions could be stationary and empty areas without any content information. Apparently, these methods are not applicable for ordinary videos because of the difficulty of detecting the unimportant region. With the widely-used devices of IDTV, personalized ads can be easily delivered in such IDTV videos [8][9]. This advertising approach works by using viewer information, or current or past users' activities. These advertising systems deliver ads according to a user's preference without considering the content-relevance of the ads and target video.

Text-based video advertising systems, such as AdSense and AdWords[4], rely on matching online textual content with keywords associated with an ad in a way that enables building the relevance between the ad and target video. Specifically, AdSense works by matching the keywords of an ad with the textual content of the webpage in which a video is embedded; whereas, AdWords aims at matching the keywords of an ad with a user's search query. Although these methods provide a common way of video advertising for online sites, the text associated with the video is usually unable to provide an accurate and precise description of the video content. In other words, text tends to be a very rough description relative

Manuscript received March 7, 2016. Accepted for publication August 24, 2016. This work was supported in part by the Natural Science Foundation of China under Grant no. 61300209 and no. 61572156, and the Shenzhen Science and Technology Program under Grant no. JCYJ20150625142543464, no. JSGG20150512145714247 and no. JCYJ20160330163900579.

Haijun Zhang and Xiong Cao are with the Shenzhen Graduate School, Harbin Institute of Technology, Xili University Town, Shenzhen 518055, P.R.China; John K. L. Ho is with the Department of Mechanical and Biomedical Engineering, City University of Hong Kong, Tat Chee Avenue, Kowloon, Hong Kong; Tommy W. S. Chow is with the Department of Electronic Engineering, City University of Hong Kong, Tat Chee Avenue, Kowloon, Hong Kong (E-mail: hjzhang@hitsz.edu.cn; jikicaxi@gmail.com; mejohnho@cityu.edu.hk; eetchow@cityu.edu.hk).

[1] www.google.com/adsense.

[2] www.youtube.com/

[3] http://video.search.yahoo.com/

[4] http://www.google.com/adwords/.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TII.2016.2605629, IEEE Transactions on Industrial Informatics

to the real video content. The effectiveness of text-based video advertising, therefore, is limited practically by the low relevance existing between the ad and target video.

Video segment level advertising, such as vADeo and VideoSense [2], aims at detecting scene changes in a video and insert a related ad in an appropriate location of the screen. Motivated by vADeo, as a simple form of video segment level advertising without considering the content-relevance of the ads and video, VideoSense was proposed to detect ad insertion positions. It relies on video content discontinuity and attractiveness to formulate an ad insertion problem as a nonlinear integer programming problem [2]. The advantage of VideoSense lies in its ability to integrate contextual relevance, i.e., global textual relevance and local visual-aural relevance, into the video advertising system. A later study further presented a contextual video advertising system (AdOn), which supports intelligent overlay in-video advertising [12]. In addition, Hong *et al.* introduced an interesting system (VideoAder) for Web video advertising [15]. It is able to leverage the well-organized media information from the video corpus for embedding visual content-relevant ads into a set of insertion points located by retrieval methods. To maximize revenues, some researchers have explored a number of other factors that may affect the video advertising system, including advertiser's bid and user interaction [13]. Moreover, Yadati *et al.* introduced an in-stream video advertising strategy by taking into account the emotional impact of the videos as well as advertisements [14]. Recently, Wang *et al.* reported an interesting system that combines online shopping information with an advertising video and directs viewers to proper online shopping places [28]. Furthermore, Hou *et al.* investigated the multi-label learning problem for advertising videos [18]. The proposed method has potential uses in solving the difficulty of concept ambiguity and diversity in ad videos.

Despite the increasingly popularity of video advertising research, OLVA has not yet been explored. This paper addresses the utilization of an informatics approach, proposing the method, model and framework to solve the crucial issues involved in the video advertising system, which can offer both content-relevance and low intrusiveness to the receivers. Given a particular unconstrained video, the proposed approach is able to automatically detect occurring objects and select ads that are relevant to the objects. Moreover, the selected ads can be inserted at the time when the related objects appear. The main contributions of this paper are twofold. First, a new framework for Object Level Video Advertising (OLVA) is proposed. Under the framework, a hybrid method for clothing retrieval is introduced, which integrates human parts alignment, human body segmentation and gender recognition into a unified framework. Second, to achieve a better relationship of video ads and receivers, an optimization model is developed. A Heuristic Algorithm (HA) is proposed to solve the optimization problem. For comparison, the Genetic Algorithm (GA) [21] is also employed to solve the global optimization problem with an appropriate encoding scheme for chromosomes. Experimental results demonstrate the effectiveness of our OLVA system, indicating that the proposed framework may offer a promising solution for online applications.

The remaining sections of this paper are organized as follows. Section II describes the overall framework and the associated implementation details. Section III presents a new optimization model for OLVA. In Section IV, the implementation of HA and the GA are described to solve the optimization problem. In Section V, the strategies of displaying video ads are discussed under the proposed framework. Experimental verifications are conducted and discussed in Section VI. Section VII concludes the paper with suggestions for future work.

II. Overview of Our Framework

A. Preliminaries

The following terms are clarified for describing our systematic framework:

- Ad insertion point: A spot in the timeline of a source video, where an ad can be inserted at a place in the video screen. It is worth noting that, for OLVA, the selected ad can be inserted around the target object. The ad insertion time point is the time when the target object appears. In addition, the ad insertion place is close to the target object, but does not cover the objects in a frame.
- Object: Objects are the main components of visual content in a video, such as humans, automobiles, beverages, etc. In comparison to keyword advertising, OLVA views an object as a keyword, and it works by retrieving the ads that are related to the object. For different objects, the retrieval methods for advertising may differ. At this stage, we focused on five types of objects, i.e. human, car, bottle, dog, and bicycle, as examples to illustrate the concept of our proposed methodology. We like to say that other types of objects can certainly be used in our framework according to the design need of different advertisement designers. It is worth pointing out that identifying what types of objects may appeal to viewers is beyond the scope of this paper, and readers may refer to [25][26][27] regarding attention, vision, and the interplay of psychology and marketing. For humans, we further processed the features extracted from the human body and used clothing retrieval for advertising (see Section III-C). For the other four types of objects, we utilized a simple category-based advertising strategy, i.e., we randomly selected an ad from the ads pool on the condition that the selected ad belongs to the same category as the target object.
- Shot: A shot is a particular sequence of pictures in a video. We segment shots from a video and insert an ad within the shot, not between two shots. Intuitively, for a shot where many objects appear, it is more probable that viewers will feel frustrated if ads are inserted into certain frames of this shot. Therefore, to lower the intrusiveness for the viewer, it is appropriate to select shots, in which only a few objects (or even one object) appear, to insert ads.
- Video ad: A video advertisement image provided by advertisers that can be inserted into a source video. In real-world applications, video ads may be in various forms, including texts, images, flash animations, video

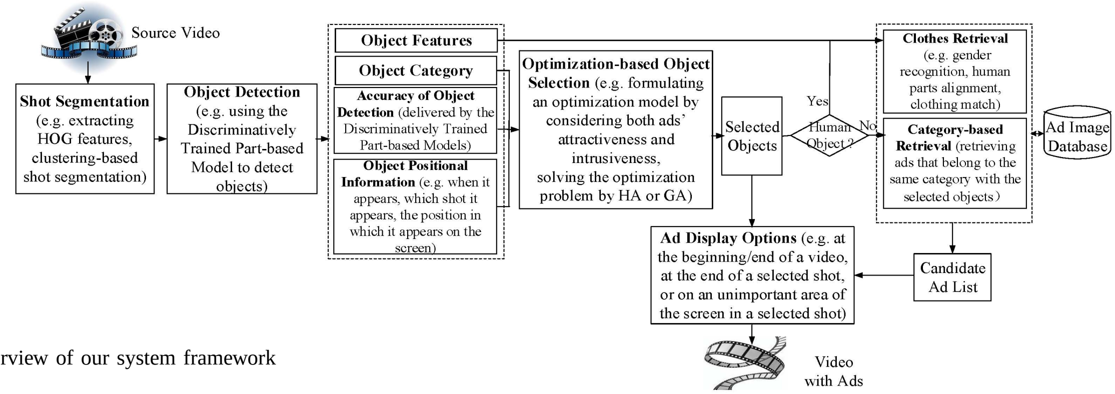

Source Video
Shot Segmentation (e.g. extracting HOG features, clustering-based shot segmentation)
Object Detection (e.g. using the Discriminatively Trained Part-based Model to detect objects)
Object Features
Object Category
Accuracy of Object Detection (delivered by the Discriminatively Trained Part-based Models)
Object Positional Information (e.g. when it appears, which shot it appears, the position in which it appears on the screen)
Optimization-based Object Selection (e.g. formulating an optimization model by considering both ads' attractiveness and intrusiveness, solving the optimization problem by HA or GA)
Selected Objects
Human Object ?
Yes
No
Clothes Retrieval (e.g. gender recognition, human parts alignment, clothing match)
Category-based Retrieval (retrieving ads that belong to the same category with the selected objects)
Ad Image Database
Candidate Ad List
Ad Display Options (e.g. at the beginning/end of a video, at the end of a selected shot, or on an unimportant area of the screen in a selected shot)
Video with Ads

Fig. 1. Overview of our system framework

clips, etc. In this paper, the ads are displayed by images that can be embedded with ads hyperlinks for online advertising.

B. Our Framework

The proposed advertising system comprises the following five components:

1) Shot segmentation using a threshold-based method that relies on the HOG (Histogram of Oriented Gradient) features;
2) Object detection using discriminatively trained part-based models;
3) Optimization-based object selection through formulating and solving an optimization problem that considers both ads' attractiveness and intrusiveness to viewers;
4) Ads retrieval including two strategies, i.e., clothing retrieval for human and category-based retrieval for other detected objects.

Fig.1 presents an overview of our framework. Given a source video, we firstly segment a video into a number of independent shots using a threshold-based method [3]. For each shot, we perform object detection using discriminatively trained part-based models provided by [4]. Through this preprocessing, we obtain certain information about each detected object, including the object features, the object category, the detection accuracy delivered by the discriminatively trained part-based models, and the object positional information that indicates when the object appears, in which shot it appears, and the position in which it appears on the screen. The achieved information is essential for the subsequent procedures. For example, the proposed optimization-based object section model relies on the identified object category, detection accuracy, and object positional information for formulating an optimization model that takes the ads' attractiveness and intrusiveness to viewers into account (see Section IV). As mentioned previously, if the object selected by the optimization model is a human body, we further process the object features for clothing retrieval by gender recognition, human parts alignment, clothes feature extraction, and matching with ad images. On the contrary, if the selected object belongs to other categories, we use simple category-based retrieval, which works by randomly retrieving an ad that belongs to the same category as the object. The rationale behind this simple strategy is that we can make use of a priori knowledge to increase the attractiveness of an ad. For example, if the selected object is a dog, embedding pet supply ads into the video is more likely to attract viewers' attention. Finally, given the selected object, which is regarded as the target object for advertising and the candidate ad that is about to be inserted into the video, we can design different ad display strategies to insert the ads into the video. The video embedded with appropriate ads is eventually transmitted to viewers or video players.

III. Implementation Details

In this section, we present the implementation details of the aforementioned framework, which contain shot segmentation, object detection, and clothing retrieval.

A. Shot Segmentation

A video typically consists of many shots. Each shot is a set of time-continuous frames, among which the background is the same. Many objects may appear in a single shot. In order to create low intrusiveness for users, at most one ad that is related to the selected object can be inserted into a shot. Before object detection and selection, a source video needs to be divided into shots. There is a large body of literature examining shot segmentation. In this paper, we used the threshold-based approach [3] to shot segmentation because of its simplicity. First, features, such as the HOG features, the LBP (Local Binary Patterns) features, and color histograms, were extracted. We calculated the distance of two adjacent frames in terms of extracted features. If the distance is larger than a predefined threshold, the two frames will be regarded as the boundaries of two shots. In our implementation, we filtered very short shots (less than 10 seconds) after shot segmentation in order to increase the ad display duration.

B. Object Detection

Object detection, an important component in our framework, enables us to accomplish object-level advertising. Many researchers have devoted much time and effort in dealing with the challenges associated with object detection. In this paper, we employ a popular method [4], which aims at discriminatively training part-based models for different objects with the HOG features. In the detection phase, it uses each model to detect a class of objects that is associated with this model. The implementation details for this can be found in [4]. The method returns a regression value, ranging from -1 to 1, which indicates the probability that the detected image block belongs to a certain type of object. In practice, the detected image block can be categorized into a type of object when the returned value is larger than a predefined threshold. This value can be regarded as a measure of detecting confidence level, which has been used in our optimization model (see Section IV). It is worth noting that each detected object is associated with a set of other information, including its category and its position in the frame. Accurately finding a region of object and regions of many objects in images or videos and efficiently dealing with background jitter, occlusion and shadow in outdoor scenes are still challenging problems. With the popularity of deep learning, recently, researchers have reported some new methods with promising results for object detection. The discriminatively trained part-based model employed under our framework can be replaced by other recently reported methods [31].

C. Clothing Retrieval

As previously mentioned, five types of objects, i.e., human, car, bottle, dog and bicycle, are detected in our system. For a human object, viewers may be interested in what he or she is wearing. Hence, if we insert a related clothes ad here, it is more likely to attract the viewers to click on the ad than if it was inserted elsewhere. To identify a related clothes ad, we further process the features extracted from the human body and use the clothing retrieval method for advertising. In this section, we design a new method for clothing retrieval, which is visually presented in Fig.2. The method contains two main components: human parts feature extraction, which is utilized to separate the human parts features from the image background, and gender recognition, which is used to recognize human gender so as to search ads in different databases. Each human body has been identified by a rectangular block using the object detection technique (see Section III-B). In practice, directly using the features of an entire human body for clothing retrieval does not perform well, because a mismatch of upper outer garment and lower garment may occur. Human parts alignment or part-based detection has been proven to be effective for matching human parts for clothing retrieval [6]. However, the extracted human parts may still contain certain background features. In particular, the background of an ad image is usually different from the background of the target human body. These background features, regarded as noise, may subsequently affect the clothes matching performance. Therefore, a human body segmentation approach is necessary for filtering out the background features in the detected human body block. On the other hand, if we only use the human body features to search for the most relevant ad in the ads database, a mismatch of male clothes and female clothes may also occur. In order to avoid this problem, we use gender recognition to classify the human of interest into male or female by face detection. Thus, if the human gender is male/female, we match the human body features with the ads in the sub-database for male/female clothes.

1) Human Feature Extraction: Feature extraction of a detected human body block includes two main procedures: human parts alignment, which aims at avoiding a mismatch of upper outer garment and lower garment in clothing retrieval, and human body segmentation, which is utilized to filter out the background features underlying a detected block.

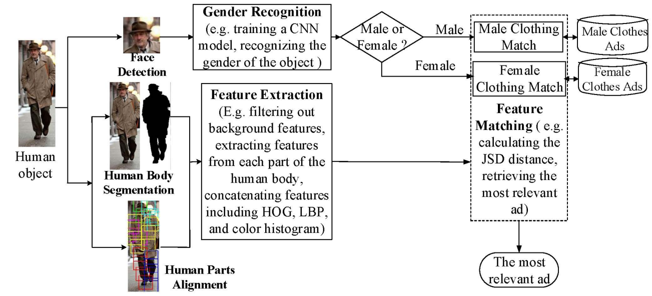

Human object
Face Detection
Human Body Segmentation
Human Parts Alignment
Gender Recognition (e.g. training a CNN model, recognizing the gender of the object)
Feature Extraction (E.g. filtering out background features, extracting features from each part of the human body, concatenating features including HOG, LBP, and color histogram)
Male or Female ?
Male
Female
Male Clothing Match
Female Clothing Match
Feature Matching (e.g. calculating the JSD distance, retrieving the most relevant ad)
Male Clothes Ads
Female Clothes Ads
The most relevant ad

Fig. 2. Framework of clothing retrieval

Human parts alignment or human parts detection [5] is capable of detecting different parts of the human body, such as the foot, hand, arm, etc. The rationale behind this method is that, regardless of how the poses of a human body change, the parts of the human body remain unchanged. After this human parts alignment, extracting features from each part is critical for clothing retrieval [6]. Similar to the work reported in [6], we use one human upper body and one human full body detector to detect 18 upper-body parts and eight lower-body parts [5]. Human parts detection results can be found in Fig.2.

Image segmentation aims at segmenting different objects in an image [10][11]. We can also use a similar approach to segment a human body from the background. In line with the work reported in [16], we utilized a dataset[5] with 5,000 human body pictures to train a model to differentiate the human body from the video background. Human segmentation results are shown in Fig.2.

After the above-mentioned two steps, we can extract the human body features by comparing the human body segmentation result with the features from each human part. Consequently, the background features underlying the detected human block can be filtered out. Three types of features including HOG, LBP and color histogram are extracted. The feature extraction of an ad picture adopts a similar process. Based on these features, we can calculate the distance between a human's feature vector and an ad picture's feature vector for clothing retrieval.

2) Gender Recognition: We have trained a six-layer deep CNN (Convolutional Neural Network) [17] for gender recognition. The first four layers in our used deep CNN are convolutional layers and the remaining two are fully-connected layers. The output of the last fully-connected layer is fed to a two-way softmax, which produces a distribution over the male and female faces. Contrast normalization layers follow the first and second convolutional layers, and max-pooling layers follow both response-normalization layers. In order to reduce overfitting, data augmentation and dropout [19] are employed. We implement data augmentation by randomly extracting a set of 96×96 patches (and their horizontal reflections) from the 128×128 images. These extracted patches are used for training our network. The implementation details of deep CNN can be found in [17].

3) Feature Matching: Three types of features, i.e., HOG, LBP and color histogram, are concatenated to form a flat vector. We extracted these features using an open library[6].

[5] http://openresearch.baidu.com/activitycontent.jhtml?channelId=411.
[6] http://www.vlfeat.org/

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TII.2016.2605629, IEEE Transactions on Industrial Informatics

The HOG features are extracted from color images. Suppose the feature vector of an ad image is denoted as $H = [h_1, h_2, ..., h_M]$, and the feature vector of a human in the video is represented as $H' = [h'_1, h'_2, ..., h'_M]$. According to empirical study [20], the Jensen-Shannon Divergence (JSD) symmetric and numerically stable when comparing two empirical distributions and performs well for color and texture. We used the JSD to calculate the distance of the two feature vectors in the form of:

$$d_{\text{JSD}}(H, H') = \sum_{m=1}^{M} h_m \log \frac{2h_m}{h_m + h'_m} + h'_m \log \frac{2h'_m}{h_m + h'_m}. \quad (1)$$

By the calculation of this distance, we can find the most relevant ad in the database with respect to the target human.

IV. Optimization Framework

To determine which shot to insert with ads and which object will be the target object in the shot for advertising, we formulate an optimization problem by considering several factors that may affect the effectiveness of our OLVA system. First, from the aspect of ad distribution over a video stream, it is expected to insert ads scattered in the video stream, because forcing users to view many ads repeatedly during a short period of time may cause them to feel quite frustrated. Second, many objects can be detected in a video shot. For different types of objects, different types of ads can be inserted corresponding to the objects. From the perspective of advertisers, it is desirable to insert as many different types of ads possible, because the more diverse the types of ads are, the more probable the viewers are to be attracted by the ads. In other words, it is better to select different categories of objects as the target objects for advertising. Third, object occurrence frequency is usually different. Some objects may frequently appear, while others appear only once. It is meaningful to insert ads associated with the frequently occurring objects, instead of the rarely occurring ones. On the one hand, frequently occurring objects are usually more important than rarely occurring objects. On the other hand, based on our advertising strategy, the time that the inserted ads corresponding to the frequently occurred objects remain on the screen is longer. Thus, it is more likely to attract the viewers' attention if we choose the frequently occurring objects as target objects for advertising, in contrast to choosing the rarely occurring objects. Fourth, from the positional aspect that an object appears on the screen, the important objects or focused objects in a shot are usually located around the center of the screen. Therefore, if a few objects are detected in a shot, it is desirable to select the objects which frequently appear around the center of the screen as the target objects. As a result, the inserted ads should be more likely to attract the attention of viewers. Fifth, the detecting confidence level of each object identified by the detection method [4] must be considered, because it is critical for an OLVA system to assure the relevance between the object and the ad. Finally, it is desirable to insert ads in a clean shot, i.e., a shot that contains only one, or very few, objects. Many objects appearing in a single shot may make the content-plot of a video clip complicated. If we insert ads into such a shot, the resulting video clips will appear even more complicated. Consequently, the intrusiveness for viewers will become high. By considering these six factors that may strongly influence the performance of an OLVA system, improving attractiveness and lowering intrusiveness of ads, from a mathematical perspective, turns out to be an optimization problem. In the following, we give the detailed problem formulation and the potential methods that provide solutions to the problem.

A. Problem Formulation

The video contains N^S shots, which can be represented by $S = \{s_i\}_{i=1}^{N^S}$. For each shot i, it contains N_i^o objects represented by $O_i = \{o_{ij}\}_{j=1}^{N_i^o}$, and the set of all the objects is represented by $O = \{O_i\}_{i=1}^{N^S}$.Our objective lies in deciding which shot should be inserted with ads and which object will be the target object in the shot for advertising.

Suppose we have the following design variables $\mathbf{x} \in R^{N^s}$, $\mathbf{y} \in R^{\sum_{i=1}^{N^s} N_i^o}$, $\mathbf{y_i} \in R^{N_i^o}$, $\mathbf{x} = [x_1, ..., x_i, ...x_{N_s}]$, $x_i \in \{0, 1\}$, and $\mathbf{y} = [\mathbf{y_1}, ..., \mathbf{y_{N^s}}]$, $\mathbf{y_i} = [y_{i1}, ..., y_{ij}, ...y_{iN_i^o}]$, $y_{ij} \in \{0, 1\}$, where x_i and $y_{i,j}$ indicate whether shot i and object j in shot i are selected, respectively. Here, we set that only one object can be selected in each shot. We will select N objects in total in different shots. Each object is associated with an ad. We designate the selected shots as $S^* = \{s_i^*\}_{i=1}^{N}$, the selected objects as $O^* = \{o_i^*\}_{i=1}^{N}$, and the ads related to the selected objects as $A^* = \{a_i^*\}_{i=1}^{N}$. Fig.3 presents an illustration of the selection process.

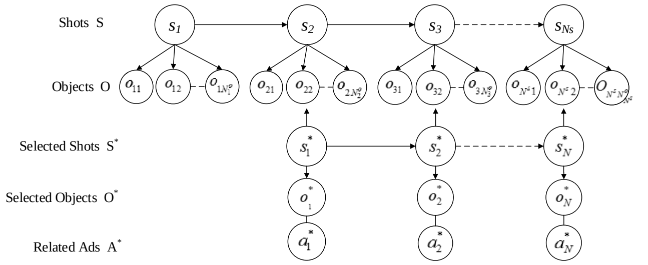


Fig. 3. Illustration of OLVA selection process

As the preceding discussion shows, several factors need to be considered for modeling. We have divided the considered factors into three categories:

- **Time Dispersion**, which indicates ad distribution over a video stream. According to [2], all of the selected ad insertion points are expected to be uniformly distributed along the timeline of the source video, so that the ads will not be densely inserted at only a few local points. It is better to insert ads that are dispersed as evenly as possible in the video stream. The optimal solution is that the time interval of any two adjacent selected shots is the same. In other words, the ad distribution is uniform, i.e., $\frac{T^v}{N}$, where T^v represents a video's playing time. In the optimization process, we can use the minimum time interval of any two adjacent selected shots as a variable to reflect the time dispersion. Apparently, the limitation of this minimum time interval is $\frac{T^v}{N}$. For simplicity, instead of using the actual time span, we use the index of a shot in the shots sequence to indicate the minimum time interval in the form of:

$$D^t = \mathop{\arg\min}_{i,k}(g(|k-i| \cdot x_i \cdot x_k)), \quad g(a) = \left\{ \begin{array}{ll} \rho, & a = 0 \\ & \\ a, & a \neq 0 \end{array} \right. , \tag{2}$$

where ρ is a very large number. Thus, for each pair of selected shots (i,k), we calculate the difference of indexes of these two shots. Then, we choose the minimum difference to reflect the degree of time dispersion.

- **Category Diversity**, which represents the diversity of the categories of selected objects. As we retrieve the relevant ads for objects, the categories of objects imply the categories of related ads. According to [29], consumers are likely to become irritated when the same ad appears too frequently. The more diverse the categories of inserted ads are, the larger probability that viewers are attracted by the ads the publisher may achieve. To make the category of ads diverse, varied types of objects can be selected. The category diversity with respect to objects can be defined by:

$$D^c = \tfrac{1}{N(N-1)} \sum_{i,j,k,l} f((C^o_{ij} - C^o_{kl}) \cdot x_i \cdot y_{ij} \cdot x_k \cdot y_{kl}), \quad f(a) = \left\{ \begin{array}{ll} 0, & a = 0 \\ & \\ 1, & a \neq 0 \end{array} \right. , \tag{3}$$

where C^o_{ij} denotes the category of object o_{ij}, i.e., the object j in shot i. Here, we use an integer to represent the category of an object.

- **Local Attractiveness and Intrusiveness**, which is only related to the selected object in a shot. It comprises four factors: 1) the detecting confidence level of a detected object; 2) the occurrence frequency of the object; 3) the distance between the object and the center of the screen; and 4) the intrusiveness brought by inserting an ad associated with the object. It is worth noting that some researchers have suggested that strengthening the visibility and animation effect of banner advertisements can attract users' attention [30]. Thus, they were naturally considered as factors in our optimization framework. We leave this interesting problem to future research regarding psychological reactions to ads using the methodologies of industrial design, marketing, and advertising. First, we denote the value of the detecting confidence level of the object j in shot i as A^d_{ij}. Second, for the calculation of the occurrence frequency of an object in a shot, we utilize different strategies for different categories of objects. For humans, we identify different persons in a shot by face clustering. We compute the occurrence frequency of a person by counting the number of frames in which the person appears. For other categories of objects, we achieve the occurrence frequency of an object by using the number of frames in which the object in the same category appears. Suppose that the number of occurrences of all of the objects in shot i is N^f_i, and the number of occurrences of the selected object j in shot i is N^f_{ij}.The frequency of object j in shot i is defined as $F_{ij} = \frac{N^f_{ij}}{N^f_i}$. Third, the positional information of a detected object can be obtained by the object detection procedure, as discussed in Section III-B. Let D^s_{ij} denote the distance between the center of object j in shot i and the center of the video screen. Finally, with respect to the intrusiveness created by inserting an ad associated with object j in shot i, we define $I_{i,j}$, the number of objects appearing at the time spot in which an associated ad is inserted, to quantitatively represent intrusiveness. Based on the above notations, local attractiveness and intrusiveness can be defined in the form of:

$$L = \tfrac{1}{N} \sum_{i,j} ((\xi_1 A^d_{ij} + \xi_2 F_{ij} + \xi_3/(D^s_{ij}+1) + \xi_4/I_{ij}) \cdot x_i \cdot y_{ij}), \tag{4}$$

where $\xi_1+\xi_2+\xi_3+\xi_4=1$, and ξ_1, ξ_2, ξ_3, ξ_4 represent the weighting parameters to balance the importance of each considered factor. In our experiment, we set $\xi_1 = 0.5$, $\xi_2 = 0.1$, $\xi_3 = 0.1$, $\xi_4 = 0.3$. In addition, the four components D^d_{ij}, $F_{i,j}$, $1/\left(D^s_{ij}+1\right)$ and $1/I_{ij}$ have been normalized in the range(0,1].

By considering these factors, we can formulate an optimization model in the form of

$$\begin{array}{l} \max h(\mathbf{x},\mathbf{y}) = w_1 \cdot D^t + w_2 \cdot D^c + w_3 \cdot L \\ = w_1 \cdot \mathop{\arg\min}_{i,k}(g(|i-k| \cdot x_i \cdot x_k)) + \\ w_2 \cdot \tfrac{1}{N(N-1)} \sum_{i,j,k,l} f((C^o_{ij} - C^o_{kl}) \cdot x_i \cdot y_{ij} \cdot x_k \cdot y_{kl})^+ \\ w_3 \cdot \tfrac{1}{N} \sum_{i,j} (\xi_1 A^d_{ij} + \xi_2 F_{ij} + \xi_3/(D^s_{ij}+1) + \xi_4/I_{ij}\} \cdot x_i \cdot y_{ij}), \\ s.t. \quad \sum x_i = N, \sum x_i \cdot y_{ij} = N, x_i \in \{0,1\}, y_{ij} \in \{0,1\}, \end{array} \tag{5}$$

where w_1, w_2 and w_3 are three predefined weighting parameters. In our experiment, we set $w_1 = w_2 = w_3 = 1$, which indicates that factors D^t, D^c and L are equally important. According to our observation, most videos under studied examples work well under this setting and performance does not vary significantly with this setting. The model parameters in our optimization model may be time-varyingly optimized according to contents of videos. It is worth noting that ad duration is also an important factor for advertising [27]. However, it is quite difficult to combine this factor into our optimization model, as the ad duration time is a continuous variable. In our implementation, an object can be selected as a candidate for advertising if its duration in the shot lasts for at least 3 seconds. If an object is selected as an advertising target in our simulated environment, the ad duration is set to, at most, 10 seconds. Thus, the ad duration varies depending on the appearance duration of the target object. If the duration of a target object is greater than 10 seconds, the associated ad duration is fixed at 10 seconds.

B. Problem Solution

Suppose that each shot contains the same number of objects N_o, then there are $C^N_{N^s} N^N_o$ solutions in total to the optimization problem as shown in Eq.(5). When the number of shots and objects is large, the searching space for optimization will dramatically increase. According to our experience, for the case of 100 shots, 10 objects in each shot, and inserting five ads (i.e. $N = 5$) in the video, it will take approximately one day to solve the problem if we use the exhaustive search. To improve efficiency, we developed a Heuristic Algorithm

(HA) to provide a solution to the problem from a local data-view. The detailed algorithm is described in **Algorithm 1**. The basic idea here is that we find the best object and shot in each search step until N objects are found. Although this HA easily becomes stuck in local minima, it exhibits a fast convergence speed. Using **Algorithm 1**, a solution can be found by searching only $NN_o\left(N^s - (N - 1/2)\right)$ steps on average.

Algorithm 1 HA algorithm:

```
Initialize: set the values of the elements in x and y to 0 (i.e., “not selected”);
repeat
    For each object in an unselected each shot, calculate the score according to the objective function if it is selected;
    Select one object in one shot with a max score, and set the according $x_i$ to 1, and $y_{ij}$ to 1;
until The number of selected objects is $N$.
Return:  solution vector x and y .
```

For comparison, we employ the Genetic Algorithm (GA) [21] to solve the problem from a global data-view. Prior to executing the main procedure of GA, we encode the chromosome, which is a solution vector $\mathbf{z} \in R^{N^s + \sum\limits_{i=1}^{N^s} N_i^o}$ in the form of

$$\begin{aligned}\mathbf{z} &= [\mathbf{x}, \mathbf{y}] = [\mathbf{x}, \mathbf{y}_1, ..., \mathbf{y}_{\mathbf{N^s}}] \\ &= [\overbrace{x_1, ..., x_{N^s}}^{\text{shot}}, \overbrace{\underbrace{y_{11}, ..., y_{1N_1^o}}_{\text{objects in shot1}}, ..., \underbrace{y_{N^s1}, ..., y_{N^sN_{N^s}^o}}_{\text{objects in shot}N^s}}^{\text{objects}}]. \end{aligned} \tag{6}$$

GA operates with a collection of these binary chromosomes, called a population. The population is randomly initialized. GA uses two operators to generate new solutions from existing ones: *crossover*, which works by combining two chromosomes, called parents, together to form new chromosomes, called offspring, and *mutation*, which aims at introducing random changes into genes of chromosomes. It is worth pointing out that the developed GA in this paper is slightly different from the conventional GA due to the exerted constraint, $\sum x_i = N, \sum x_i \cdot y_{ij} = N, x_i, y_{ij} \in \{0, 1\}$, existing in the optimization model. For initialization, we randomly select N shots and set the selected x_i to 1 so as to meet the constraint $\sum x_i = N$. On the other hand, we select one object in each shot and set its corresponding design variable, $y_{.,j}$, to 1 in order to meet the constraint $\sum x_i \cdot y_{ij} = N$. In crossover operation, we view components $\mathbf{x}$, $\mathbf{y_1}$, $\mathbf{y_2}$,...,$\mathbf{y_{N^s}}$, shown in Eq.(6), as gene units in each chromosome. Thus, the crossover operation involved in our GA is similar to the conventional GA. In the mutation operation, we separately address the components $\mathbf{x}$ and $\mathbf{y_u}(u = 1, 2, ..., N^s)$ if each of them is selected to execute the mutation according to the predefined mutation rate. Concretely, for component $\mathbf{x}$, we exchange the binary values of one previously selected shot and one of the other shots that are not selected in the current step; for each component $\mathbf{y_u}$, we randomly select one of the other objects, which is not selected in the current step, and set its value to 1. In our experiment, we set the population size to 200, the crossover rate to 0.25, the mutation rate to 0.01, and the maximal number of iterations to 200.

TABLE I
DATASET STATISTICS OF AD IMAGES

Category of ads	Number of ads images	Number of subcategories
Bicycle	365	1
Car	342	1
Dog	725	1
Drink	3258	4
Man clothing	6008	12
Woman clothing	10447	16

V. AD DISPLAY STRATEGIES

The achieved solution to the optimization model enables us to make a decision about objects and shots that are optimally selected for advertising. Moreover, each object has now been successfully associated with a relevant ad. The subsequent work focuses on determining precisely how to display the ads in a video stream. In fact, in the fields of marketing and advertising, a large body of literature exists that investigates how to attract viewers' attention [25]. However, very few papers have considered video ad placement and frequency. Recently, Hsieh and Chen discussed how different information types affect viewers' attention in Internet advertising [26]. In their experiment, a banner ad was displayed at the top of a video sample. Krishnan and Sitaraman conducted a measurement study on the effectiveness of video ads based on a real online video delivery network [27]. They analyzed the impact of pre-roll, mid-roll, and post-roll ad positions on the likelihood that a viewer would watch the ad to completion. These ad display strategies are quite common within enterprise videos, e.g., YouTube and Yahoo! Video. VideoSense [2] suggested that an ad may be embedded in an unimportant area of a video frame. In fact, there are many options for displaying ads by using our object-ad relevance enabled advertising system. Specifically, an ad can be displayed: 1) at the beginning or end of a video; 2) at the end of a selected shot; and 3) on an unimportant area of the screen in a selected shot. The first strategy has been widely employed in current online video sites, e.g., YouTube and Yahoo! Video. These sites, however, usually do not consider the relevance between displayed ads and the video content. The second strategy is quite similar to VideoSense [2], but our method does not require any textual information. Additionally, as mentioned previously, textual information tends to only provide a rough description of the video. The resulting relevance is not sufficient to bridge the semantic gap between video contents and ads. On the contrary, the object-level relevance empowered by our system provides an alternative way to maximize the value of video ads. The third strategy is a straightforward option, which is similar to the pop-up flash ads adopted by many websites. An unimportant area could be a corner of the screen. In this study, we used this strategy by embedding the ad in the right-lower corner of the screen.

VI. EXPERIMENT

A. Dataset

To evaluate the performance of our proposed framework, we firstly built a large-scale dataset containing 21,145 ad pictures, which were collected from an online shopping website[7]. We

[7] www.taobao.com

collected six categories of ad pictures: bicycle, car, dog, bottle, female clothing, and male clothing. A few keywords were used for searching ad pictures.In particular, for bottle, we used the keywords (i.e., sub-categories): “Chinese spirits”, “red wine”, “beer”, and “imported wines”; for male clothing, we used the keywords: “male T-shirt”, “male shirt”, “male wind coat”, “male jacket”, “male vest”, “male jeans”, “male fur”, “male leather wear”, “male Chinese garments”, “male suit”, “male snowsuit”, and “Chinese tunic suit”; for female clothing, we used keywords: “female T-shirt”, “suspenders”, “female shirt”, “wedding dress”, “evening dress”, “one-piece dress”, “female vest”, “female fur”, “female leather wear”, “cheong-sam”, “female wind coat”, “female Chinese garments”, “female singlet”, “chiffon shirt”, “female snowsuit”, and “female business suit”. Table I shows the dataset statistics of the six types of ad pictures.

On the other hand, 13 videos were collected[8] and used for testing videos. The videos were selected in a way that they have a representative coverage on different video genres. Also, the target objects we focused on in this paper occur in these videos frequently. We embedded a fixed number of ads, i.e., $N = 5$, in our simulated advertising environment. Table II presents the statistics of videos with respect to runtime, the number of segmented shots, and the number of detected objects. Since this work concerns the basis of object relevance, the performance of object detection delivered by the method [4] is critical to our system. Table III summarizes the results of detection accuracy over different categories in the used videos. The top three categories with respect to average detection accuracy are human, car, and bottle. Although the performance of detection accuracy depends on the types of videos, it is quite reasonable that these three categories of objects usually occur in a video. In particular, the average detection accuracy for humans reaches 96.85%. This indicates the easy implementation of building relevance between human objects and ads, especially clothing ads. As shown in Table II, 6,480 human objects were detected from all of the used videos. However, some of these human objects may be of low quality, e.g., blurry images. Fig.4 shows some failure examples when we conducted our experiments on the video “The big bang theory”. Some regions were detected and wrongly classified as human body. To increase the performance of our system, we developed a pre-processing strategy to filter out the low quality human objects. Specifically, we used the Viola-Jones face detector to perform on each block containing a human object. The detector returns a value, indicating the probability that the detected area belongs to a human face. If this value was smaller than a predefined threshold, we filtered out the corresponding human object. As a result, 3,530 human objects were filtered out, and the rest of the 2,950 human objects were used for the subsequent procedure, i.e., clothing retrieval. For gender recognition, the dataset that we utilized to train CNN is the Labeled Faces in the Wild (LFW) [22]. The dataset contains faces of 5,749 individuals including 10,256 male images and 2,977 female images. To mitigate computational load, we resized the images into a resolution of 128×128.

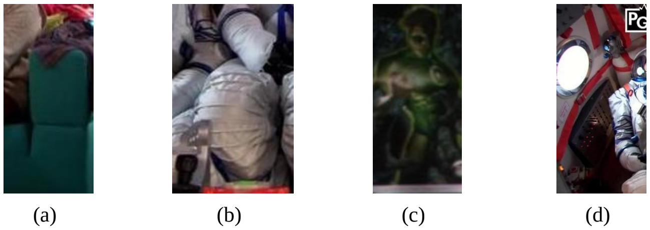


Fig. 4. Failure examples.

TABLE II
VIDEO DATASET STATISTICS

Video ID	Video name	Runtime (minute)	Number of shots	Number of detected objects
V1	3 Idiots 1	11	52	356
V2	3 Idiots 2	49	377	1884
V3	Bones	42	304	508
V4	Cycling paradise	4	33	37
V5	Do not quit drink 1	4	31	66
V6	Do not quit drink 2	6	33	169
V7	Guide dog little Q	24	208	387
V8	Hachiko’s story 1	10	65	68
V9	Prison break	43	329	646
V10	The big bang theory	21	157	740
V11	The fast and the furious 1	14	101	249
V12	The fast and the furious2	13	92	493
V13	The matrix reloaded1	46	358	877
	All	287	2140	6480
	Avg.	22	164	498

TABLE III
RESULTS OF OBJECT DETECTION IN OUR DATASET (ND AND NWD REPRESENT THE NUMBER OF DETECTED OBJECTS AND WRONGLY DETECTED OBJECTS IN CORRESPONDING CATEGORY, RESPECTIVELY)

Video ID	Object Category									
	Bicycle		Bottle		Car		Dog		Human	
	ND	NWD	ND	NWD	ND	NWD	ND	NWD	ND	NWD
V1	0	0	0	0	1	1	0	0	355	8
V2	12	12	1	1	31	6	5	5	1835	61
V3	1	1	2	1	12	6	5	5	488	10
V4	33	0	0	0	0	0	0	0	4	3
V5	0	0	0	0	1	1	0	0	355	8
V6	12	12	1	1	31	6	5	5	1835	61
V7	1	1	8	1	0	0	0	0	57	7
V8	0	0	10	0	1	1	0	0	158	9
V9	5	0	0	0	45	6	37	0	300	15
V10	3	3	0	0	4	3	9	1	52	1
V11	4	4	0	0	31	9	3	3	608	13
V12	8	8	2	0	11	11	0	0	719	14
V13	1	1	0	0	334	15	1	1	541	12
All	71	31	24	4	807	66	62	17	5516	174
Avg. Accuracy(%)	56.34		83.33		91.82		72.58		96.85	

TABLE IV
GENDER RECOGNITION RESULT (CV-5 REPRESENTS CROSS-5-VALIDATION)

Model	Feature extraction	Classifier	Test data	Accuracy(%)
Shan, 2012 [45]	LBP hist. bins	SVM-RBF	CV-5	94.81
Our Model	CNN	CNN	CV-5	97.45

B. Gender Recognition

To evaluate the performance of our designed CNN on gender recognition, we firstly tested our model on the LFW dataset, in comparison to the model reported by Shan [23]. We assigned 80% for training and 20% for test. The average accuracy based on cross-5-validation is listed in Table IV. It is clear that CNN delivers better performance than the method introduced by [23]. We also list the gender recognition result using CNN in our dataset, as shown in Table V. The average recognition rate is 94.98%, which suggests the feasibility of utilizing gender recognition in the clothes advertising framework. It is noted that there is no human object left in the video *Cycling Paradise* for gender recognition after the pre-processing strategy which is used to filter out low quality human objects as mentioned in the last section.

C. Clothing Retrieval

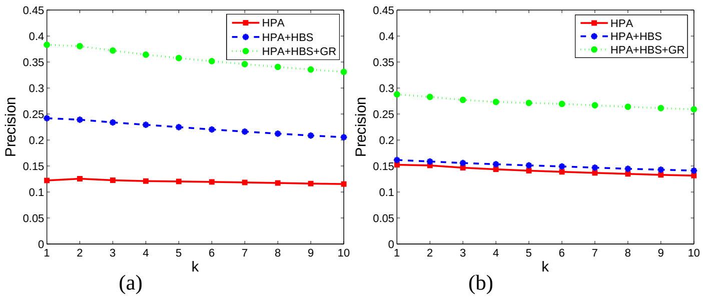
Currently, we define that two pieces of clothing are relevant if they are in a similar color and style by manual evaluation. We use a ranking-based criterion for evaluation [24],

[8] Videos were downloaded from Internet and only used for research purpose.

TABLE V
ACCURACY OF GENDER RECOGNITION USING CNN IN OUR SYSTEM

Video ID	Number of human objects	Number of wrongly recognized human genders
V1	220	4
V2	961	20
V3	250	15
V4	-	-
V5	29	11
V6	91	7
V7	101	9
V8	14	1
V9	357	18
V10	507	32
V11	96	7
V12	97	2
V13	227	22
All	2950	148
Avg. Accuracy(%)	94.98	

$Precision@k$ (a ranking of the top k retrieved images). As described in Section III-C, we introduce a new framework for clothing retrieval, which integrates human parts alignment (HPA), human body segmentation (HBS), and gender recognition (GR). For comparison, we summarize the results in Table VI in terms of precision at 10, using different techniques for clothing retrieval. Fig.5 visually illustrates the comparative results against the number of retrieved clothes. It is worth noting that the method that uses human parts alignment and human body segmentation (i.e., HPA+HBS in Table VI), is similar to the work reported in [6]. We can observe that the hybrid method outperforms the other methods to a large degree. Moreover, it is observed that the precision results produced by our system are relatively low. This is caused by the fact that the image quality has been largely affected during the video compression process and by human pose variation. The selection of high quality images for clothing retrieval still presents a challenging problem, which we leave to future work.



(a) (b)

Fig. 5. Clothing retrieval according to: (a) color relevance; (b) style relevance.

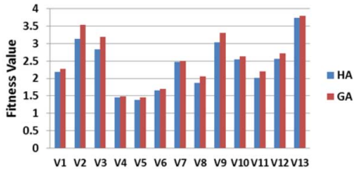


Fig. 6. Optimal fitness values over different videos using HA and GA.

D. Overall Optimization Framework

This section evaluates the overall performance of our framework. First, we compare the optimization performance of HA and GA. Second, the overall effectiveness of our system is measured.

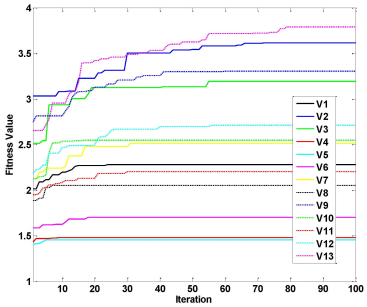
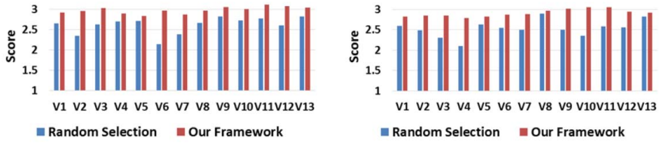


Fig. 7. Convergence curves of GA.

TABLE VI
RESULTS OF CLOTHING RETRIEVAL ON PRECISION AT 10

Method	HPA	HPA + HBS[6]	HPA + HBS + GR
Color relevance	0.12	0.21	0.33
Style relevance	0.13	0.14	0.26

1) HA vs. GA: To quantitatively evaluate the performance of our proposed optimization framework for video advertising, we summarize the average optimal fitness values, delivered by HA and GA, over all of the videos in Table VII. For comparison, the time cost of HA and GA is also listed. For clarity, the optimal fitness values searched by HA and GA over different videos are visually illustrated in Fig.6. The convergence curves of GA over different videos are plotted in Fig.7. The results of GA are based on the average result by randomly running 10 times. It is observed that GA performs better than HA over all of the videos. The proposed HA is more possible to stuck local optimum. On the contrary, the GA has better capability of searching optimal solutions in a global space. But the convergence speed of HA is around 193 times faster than GA. This suggests that HA may constitute a promising option for online applications because of its low time cost.



(a) (b)

Fig. 8. Overall performance: (a) Attractiveness; (b) Comfortableness.

2) Overall Performance: To measure the effectiveness of our optimization framework, we can compare our method with the baseline approach that simply selects ads randomly to insert into a video from the given ads pool. Here, the optimization method that we tested is HA. Similar to [2], we conducted a subjective user study. 22 graduate students from a computer science department were invited to participate in the user study. Each individual was assigned 13 videos used in this work. When viewing each of the results produced by the two ad selection methods, i.e., our optimization framework and the random selection method, the evaluators were asked to give a score from 1 to 5 (with a higher score implying better satisfaction) in terms of the attractiveness and comfortableness created by the inserted ads in a video:

1) *Attractiveness*: for each ad, did you think the ad was

TABLE VII
COMPARATIVE RESULTS USING HA AND GA

Method	HA	GA
Optimal fitness value	2.38	2.53
Time cost (sec)	0.16	30.89

TABLE VIII
AVERAGE SCORES OF DIFFERENT VIDEO ADVERTISING METHODS

Method	Random Selection	Optimized Selection (our method)
Attractiveness	2.61	2.98
Comfortableness	2.53	2.91

attractive as you viewed the video?

2) *Comfortableness*: for each ad, did you feel comfortable as you viewed the video?

The average scores given by the evaluators for different videos are visually illustrated in Fig.8(a)-8(b), and the overall averaged results are summarized in Table VIII. The results clearly reveal that the optimization framework can improve performance on both the attractiveness and comfortableness dimensions. This indicates that the proposed framework may constitute a promising solution to real online applications for video advertising. Furthermore, in order to evaluate the effect of ad display strategies on user's viewing experience, we develop an ad insertion point detection method by formulating another optimization model for ad display. The detailed model description and performance evaluation can be found in the supplementary document. Our subjects for evaluating the performance of our system were limited to a small group of college-aged students. The response results may be slightly deviated from real online users' attributes and behaviors. We need to conduct a large-scale user study on more videos in future work.

In addition, regarding our optimization framework, we asked participants the question, "Would you have continued to watch this video?" Most of the participants indicated that when they knew that the same video sources were provided in other online channels, they would change to these video channels to avoid the ads. This finding is in line with much extant research on ad irritation [29]. Then, under the assumption that there were no other channels on which to watch these videos, the participants were asked to give a score from 1 to 3 (a score of 1 meant "No, immediately stop continuing to watch", a score of 2 meant "I could bear these ads, and I would continue to watch", and a score of 3 meant "The ads had not affected my viewing, and I would continue to watch"). The average score in our experiment was around 2.02, which suggested that participants could tolerate the ads and would continue to watch a video associated with ads using our method.

In our proposed framework, the computational cost mostly comes from object detection, object selection based on our optimization model, feature extraction from detected objects, ads retrieval. In our experiment, the object detection process using the Discriminatively trained Part-based Model requires around 1.8s for each selected image. The time costs of object selection based on HA and GA are 0.16s and 30.89s (see Table VII), respectively. The time cost of feature extraction from detected objects is around 0.019s for each object. The time cost of ads retrieval relies on the size of ad dataset. Retrieving one relevant ad from 21,145 ad images requires around 20.58s.

VII. CONCLUSIONS

This paper presents new models and algorithms for OLVA under an optimization framework. First, we develop an OLVA system, which involves numerous techniques, such as shot segmentation, object detection, and clothing retrieval. A new clothing retrieval method is also introduced. Second, we propose a mathematical modeling method for the optimization of video advertising by taking into account several factors, including the ad's time dispersion, category diversity, and local attractiveness and intrusiveness to users. Third, we develop an HA to provide solutions to the optimization model using local information. For comparison, we also employ the GA, a stochastic global optimization algorithm, to optimize the model with an appropriate encoding scheme for chromosomes and using global information. Experimental results clearly demonstrate the effectiveness of our proposed framework. In future work, it would be interesting to further investigate what types of video ad placements and what designs of ads in videos are optimal using well-established theories in industrial design, marketing, and advertising. The view from different angles of objects has not been considered under our framework. It is worth a further examination. In addition, we did not include certain factors in our optimization framework, such as shot length, ad duration, and ad frequency. These may also affect the effectiveness of video ads, and are worthy of further study.

REFERENCES

[1] Cisco. Visual Networking Index: Forecast and Methodology, 2013-2018. URL: www.cisco.com/c/en/us/solutions/collateral/service-provider/ip-ngn-ip-next-generation-network/white_paper_c11-481360, May 2014.

[2] T. Mei, X. S. Hua, S. Li, "VideoSense: A contextual in-video advertising system," *IEEE Trans. on Circuits Syst. Video Technol.*, vol. 19, no. 12, pp. 1866-1879, Dec. 2009.

[3] G. Pal, *et al.*, "Video shot boundary detection: a review," in *Proc. IEEE Int. Conf. Electrical, Computer and Communication Technologies*, 2015, pp. 1-6.

[4] D. Forsyth, "Object detection with discriminatively trained part-based models," *Computer*, vol. 47, no. 2, pp. 6-7, 2014.

[5] Y. Yang and D. Ramadan, "Articulated human detection with flexible mixtures of parts," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 35, no. 12, pp. 2878-2890, Dec. 2013.

[6] S. Liu, Z. Song, G. Liu, "Street-to-shop: cross-scenario clothing retrieval via parts alignment and auxiliary set," in *Proc. IEEE Conf. Comput. Vision Pattern Recogn. (CVPR)*, 2012, pp. 3330-3337.

[7] M. López-Nores, Y. Blanco-Fernández, J. Pazos-Arias, "Cloud-based personalization of new advertising and e-commerce models for video consumption," *The Computer Journal*, vol. 56, no. 5, pp. 573-592, 2013.

[8] R. Redondo, *et al.*, "Bringing content awareness to web-based IDTV advertising," *IEEE Trans. Syst. Man Cybern. C, Appl. Rev.*, vol. 42, no. 3, pp. 324-333, May 2012.

[9] D. Véras, *et al.*, "A literature review of recommender systems in the television domain," *Expert. Syst Appl.*, vol. 42, no. 22, pp. 9046-9076, 2015.

[10] D. Mukherjee, Q. M. Jonathan Wu, and T. M. Nguyen, "Gaussian mixture model with advanced distance measure based on support weights and histogram of gradients for background suppression," *IEEE Trans. Ind. Informat.*, vol. 10, no. 2, pp. 1086-1096, May 2014.

[11] H. Liu, *et al.*, "Intelligent video systems and analytics: a survey," *IEEE Trans. Ind. Informat.*, vol. 9, no. 3, pp. 1222-1233, Aug. 2013.

[12] T. Mei, *et al.*, "AdOn: toward contextual overlay in-video advertising," *Multimedia Syst.*, vol. 16, no. 4, pp. 335-344, 2010.

[13] J. Wang, *et al.*, "Interactive ads recommendation with contextual search on product topic space," *Multimedia Tools Appl.*, vol. 70, no. 2, pp. 799-820, 2014.

[14] K. Yadati, H. Katti, M. Kankanhalli, "CAVVA: Computational affective video-in-video advertising," *IEEE Trans. Multimedia*, vol. 16, no. 1, pp. 15-23, Jan. 2014.

[15] R. Hong, *et al.*, "Advertising object in web videos," *Neurocomputing*, vol. 119, pp. 118-124, 2013.

[16] J. Shotton, *et al.*, "TextonBoost for image understanding: multi-class object recognition and segmentation by jointly modeling texture, layout, and context," *Int. J. Comput. Vision*, vol. 81, no. 1, pp. 2-23, 2009.

[17] A. Krizhevsky, I. Sutskever, G. Hinton, "ImageNet classification with deep convolutional neural networks," *Adv. Neural Informat. Proc.*, pp. 1097-1105, 2012.

[18] S. Hou, *et al.*, "Multi-label learning with label relevance in advertising video," *Neurocomputing*, vol. 171, pp. 932-948, 2016.

[19] T. Sainath, *et al.*, "Dropout: a simple way to prevent neural networks from overfitting," *J. Mach. Learn. Res.*, vol. 15, no. 1, pp. 1929-1958, 2014.

[20] Y. Rubner, *et al.*, "Empirical evaluation of dissimilarity measures for color and texture," Comput. Vis. Image Und., 2001, 84(1): 25-43.

[21] S. Oreski, G. Oreski, "Genetic algorithm-based heuristic for feature selection in credit risk assessment," *Expert Syst. Appl.*, vol. 41, no. 4, pp. 2052-2064, 2014.

[22] G. B. Huang, E. Learned-Miller, "Labeled faces in the wild: Updates and new reporting procedures," *Tech. Rep.*, Dept. Comput. Sci., Univ. Massachusetts Amherst, Amherst, MA, USA, 14-003, 2014.

[23] C. Shan, "Learning local binary patterns for gender classification on real-world face images," *Pattern Recogn. Let.*, vol. 33, no. 4, pp. 431-437, 2012.

[24] J. Deng, A. C. Berg, F. F. Li, "Hierarchical semantic indexing for large scale image retrieval," in *Proc. IEEE Conf. Comput. Vis. Pattern Recogn. (CVPR)*, 2011, pp. 785-792.

[25] H. Banga and B. W. Wojdynskib, "Tracking users' visual attention and responses to personalized advertising based on task cognitive demand," *Comput. Hum. Behav.*, vol. 55, Part B, pp. 867-876, 2016.

[26] Y. C. Hsieh and K. H. Chen, "How different information types affect viewer's attention on internet advertising," *Comput. Hum. Behav.*, vol. 27, pp. 935-945, 2011.

[27] S. Krishnan and R. Sitaraman, "Understanding the effectiveness of video ads: a measurement study," in *Proc. ACM Internet Measurement Conf. (IMC)*, 2013, pp. 149-162.

[28] J. Wang, *et al.*, "ActiveAd: A novel framework of linking ad videos to online products," *Neurocomputing*, vol. 185, pp. 82-92, 2016.

[29] H. Li, Steven M. Edwards, and Joo-Hyun Lee, "Measuring the intrusiveness of advertisements: scale development and validation," *J. Advert.*, vol. XXXI, no. 2, pp. 37-47, 2002.

[30] W. C. Chiou, C. C. Lin, C. Perng, "A strategic framework for website evaluation based on a review of the literature from 1995-2006," *Informat. & Manage.*, vol. 47, no. 5, pp. 282-290, 2010.

[31] R. Girshick, *et al.*, "Region-based convolutional networks for accurate object detection and segmentation," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 38, no. 1, pp. 142-158, 2016.

Haijun Zhang (M'13) received the B.Eng. and Master's degrees from Northeastern University, Shenyang, China, and the Ph.D. degree from the Department of Electronic Engineering, City University of Hong Kong, Hong Kong, in 2004, 2007, and 2010, respectively. He was a Post-Doctoral Research Fellow with the Department of Electrical and Computer Engineering, University of Windsor, Windsor, ON, Canada, from 2010 to 2011. Since 2012, he has been with the Shenzhen Graduate School, Harbin Institute of Technology, China, where he is currently an Associate Professor of Computer Science. His current research interests include neural networks, multimedia data mining, machine learning, computational advertising, and evolutionary computing.

Xiong Cao received his BS and MS degrees from the Harbin Institute of Technology in 2012 and 2015, respectively. He was a Master candidate in Computer Engineering of the Harbin Institute of Technology Shenzhen Graduate School, under this research performed. He is currently working on TEG recommendation systems at Tencent Computer System CO., LTD. His research interests include multimedia data mining, computational advertising, and deep learning.

John K. L. Ho received his BSc & MSc degrees in computer, control engineering from Coventry University and PhD degree from University of East London, UK. He has many years design experience in the field of automation when was working in GEC Electrical Projects Ltd in UK. Currently, Dr. Ho is the chairman of Control, Automation & Instrumentation Discipline Advisory Panel of the Hong Kong Institution of Engineers. He is an associate professor in the Department of Mechanical and Biomedical Engineering, City University of Hong Kong. His research interest is in the fields of data mining, control engineering, green manufacturing, enterprise automation and product design.

Tommy W. S. Chow (M'94 - SM'03) received the B.Sc. (1st Hons) degree and the Ph.D. degree from the Department of Electrical and Electronic Engineering, University of Sunderland, U.K. He is currently a Professor in the Department of Electronic Engineering at the City University of Hong Kong. He has been working on different consultancy projects with the Mass Transit Railway, Kowloon-Canton Railway Corporation, Hong Kong. He has also conducted other collaborative projects with the Hong Kong Electric Co. Ltd, the MTR Hong Kong, and Observatory Hong Kong on the application of neural networks for machine fault detection and forecasting. He is an author and co-author of over 170 Journal articles related to his research, 5 book chapters, and 1 book. His main research has been in the area of neural networks, machine learning, pattern recognition, and documents analysis and recognition. He received the Best Paper Award in 2002 IEEE Industrial Electronics Society Annual meeting in Seville, Spain.