VideoSense – Towards Effective Online Video Advertising

Tao Mei, Xian-Sheng Hua, Linjun Yang, Shipeng Li
Microsoft Research Asia
49 Zhichun Road, Beijing 100080, P. R. China
{tmei, xshua, linjuny, spli}@microsoft.com

ABSTRACT

With Internet delivery of video content surging to an unprecedented level, online video advertising is becoming increasingly pervasive. In this paper, we present a novel advertising system for online video service called *VideoSense*, which automatically associates the most relevant video ads with online videos and seamlessly inserts the ads at the most appropriate positions within each individual video. Unlike most current video-oriented sites that only display a video ad at the beginning or the end of a video, VideoSense aims to embed more contextually relevant ads at less intrusive positions within the video stream. Given an online video, VideoSense is able to detect a set of candidate ad insertion points based on content *discontinuity* and *attractiveness*, select a list of relevant candidate ads ranked according to *global textual relevance*, and compute *local visual-aural relevance* between each pair of insertion points and ads. To support contextually relevant and less intrusive advertising, the ads are expected to be inserted at the positions with highest discontinuity and lowest attractiveness, while the overall global and local relevance is maximized. We formulate this task as a nonlinear 0-1 integer programming problem and embed these rules as constraints. The experiments have proved the effectiveness of VideoSense for online video advertising.

Categories and Subject Descriptors

H.5.1 [**Information Interfaces and Presentation**]: Multimedia Information Systems—*video*; H.3.5 [**Information Storage and Retrieval**]: Online Information Services—*Web-based services*

General Terms

Algorithms, Experimentation, Human Factors.

Keywords

Online video advertising, contextual relevance, less intrusiveness.

MM'07, September 23–28, 2007, Augsburg, Bavaria, Germany.

1. INTRODUCTION

Driven by the coming age of the Internet and the advent of near-ubiquitous broadband Internet access, online delivery of video content has surged to an unprecedented level. Today's online users face a daunting volume of video content – be it from video sharing or blog content, or from IPTV and mobile TV. ComScore reports that in March 2006 alone consumers viewed 3.7 billion video streams and nearly 100 minutes of video content per viewer per month, compared to an average of 85 minutes in October [5]. Moreover, as reported by Online Publisher Association [22], the majority (66%) of Internet users have ever seen video ads, while 44% have taken some action after viewing ads. Accordingly, spending on online video advertising is dramatically increasing. eMarketer reports that the spending for Internet video advertising in the U.S. will nearly triple in 2007 to $640 million from 2006 year's $225 million [7]. To take the advantages of this increasing market share and effectively monetize video content, video advertising, which associates advertisements with an online video, has become a key online monetization strategy. By implementing a solid online video advertising strategy into an existing content delivery chain, content providers have the ability to deliver compelling content, reach a growing online audience, and generate additional revenue from online media.

The "effectiveness" of an online advertising was primarily defined from the advertisers' perspective, and usually measured by performance of a given ad (e.g. Click-Though Rate) or brand perception. Recent research has discovered that the deep understanding of user experience can help explore the nature and negative impacts of online advertising [26]. Thus, user experience reflects "effectiveness" from another perspective. It has been recognized that perceived intrusiveness and irrelevance have the leading negative effect on user experience [16] [20]. Intrusiveness is "a perception or psychological consequence that occurs when an audience's cognitive processes are interrupted" [16] such as television commercials during an exciting scene in a program, while contextual relevance in our study addresses the relationship or relevance of the ad content with the source video content. Therefore, an effective online advertising system designed from the viewers' perspective should take both contextual relevance and less intrusiveness into consideration.

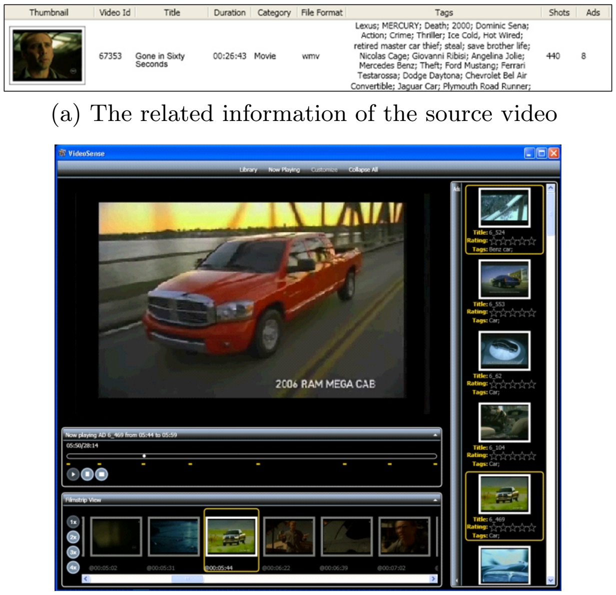
Many existing video-oriented sites, such as YouTube [34], Google Video [8], Yahoo! Video [31], Metacafe [21], and Revver [24], have tried to provide effective video advertising services. However, it is likely that most of them match the ads with online videos only based on textual information and insert ads at the beginning or the end of a video [1]. In other words, contextual relevance in these sites is only based on textual information, while less intrusive insertion points are fixed to the beginning or the end of videos. For instance, the most similar advertising system to VideoSense – Revver [24], selects one relevant ad (i.e., a static picture or video clip) for each video clip, and shows it as the last frame or segment of the corresponding video. A hyperlink to the original ad is embedded in this frame. As a result, the following problems that significantly affect advertising effectiveness and impede user experience have not been investigated.

- We believe ads should be inserted at appropriate positions within video streams rather than only at the beginning or the end of video streams. While most of the videos on the Internet consist of shorter clips, we are now seeing longer-form content, particularly as networks and film studios have started releasing top programs online. This capability will enable embedding not only a greater number of ads but also *less intrusive* ads within video content.

- We believe ads should be *contextually relevant* to online video content in terms of not only textual information but also visual and aural content. For example, when viewing an online music video, users may prefer a relevant ad with the similar editing style or audio tempo style to the video, which cannot be measured just by textual information. This capability will enable delivering the ads with more relevance.

Although the insertion of video ads within a video stream is similar to a traditional advertising spot in TV broadcasting, which temporarily interrupts programs to provide a paid sponsor's message, ads are manually inserted at a fixed interval without considering whether they are less intrusive and contextually relevant. This practice has been often criticized in the field of TV for disappointing viewers and for taking them to do zapping [13]. From this point of view, ads have not reached users positively.

Motivated by the above observations, we present a novel advertising system for online video called VideoSense, which supports more effective video advertising in terms of *contextual relevance* and *less intrusiveness*. In VideoSense, given an online source video consisting of video content and related textual information, relevant video ads will be automatically associated with the source video and seamlessly inserted into the video at appropriate positions. As a result, VideoSense generates an augmented video stream with ads embedded – selecting the most relevant ads in terms of global textual relevance, detecting the most appropriate insertion points with high content discontinuity and low attractiveness, keeping the high local visual-aural relevance between the ads and source video content around the insertion points, and uniformly distributing the ads along a timeline. We formulate the task as a nonlinear integer programming problem in which each of the above desirability is embedded as a constraint. We conducted both objective and subjective evaluations on an extensive experiment and proved the effectiveness of VideoSense for online video advertising.



Thumbnail	Video Id	Title	Duration	Category	File Format	Tags	Shots	Ads
	67353	Gone in Sixty Seconds	00:26:43	Movie	wmv	Lexus; MERCURY; Death; 2000; Dominic Sena; Action; Crime; Thriller; Ice Cold, Hot Wired; retired master car thief; steal; save brother life; Nicolas Cage; Giovanni Ribisi; Angelina Jolie; Mercedes Benz; Theft; Ford Mustang; Ferrari Testarossa; Dodge Daytona; Chevrolet Bel Air Convertible; Jaguar Car; Plymouth Road Runner;	440	8

(a) The related information of the source video

(b) VideoSense user interface

Figure 1: An example of an online source video with contextually relevant ads embedded. The yellow bars below the timeline indicate video ads being inserted at these points. The thumbnails with yellow box in a filmstrip view correspond to video ads. The candidate ads are listed in the right panel in which the highlighted ads are inserted into the source video.

An example is shown in Fig. 1. Since the content provider of this source video has tagged it "Lexus," as show in Fig. 1(a), some candidate ads listed at the right panel in Fig. 1(b) are related to "car." One of the candidate ads has been inserted into this video, i.e. the highlighted thumbnail with yellow box in the filmstrip view. Since this ad is inserted at the boundary of two scenes, as well as both the source video and ad are related to "car," we propose that the ad is less intrusive and contextually relevant.

The rest of the paper is organized as follows. Section 2 reviews research work related to online video advertising. Section 3 provides a system overview of VideoSense. The main components of VideoSense, i.e. candidate ad ranking, ad insertion point detection, and online ad insertion are described in Section 4, 5, and 6. The effectiveness is evaluated in Section 7, followed by conclusions in Section 8.

2. RELATED WORK

The research problems closely related to online video advertising include advertisement placement in sports videos and Interactive Digital Television (IDTV), as well as text-based online advertising.

To make video content more enriching, previous work in literature [17] [29] has attempted to spatially replace a specific region with product advertisement in sports videos. These regions could be locations with less information in baseball video [17], or the region above the goal-mouth in soccer video [29]. An online platform is also presented to measure the quality of product placement [11]. However,

[1] Typical examples for textual relevance matching are the keyword-targeted (e.g., Google's AdWord) and content-targeted advertising (e.g., Google's AdSense).

VideoSense is designed for general online video rather than specific video, as well as VideoSense aims at video and video segment level advertising in contrast to the object/region-level advertising. The domain-specific approaches in these applications such as the detection of line and less-information-region [17] [29], are not practical in a general case, especially in online videos. Moreover, contextual relevance, which is the main aspect influencing viewing experiences, is not taken into consideration in these systems.

While region-based product placement is challenging, the personalized ad delivery in IDTV has been a potentially hot application [12] [15] [27]. Such advertising refers to the delivery of advertisements tailored to the individual viewers' profiles on the basis of knowledge about their preferences [15] or current and past contextual information [12] [27]. However, most of these systems do not study relevance in terms of video content and the elaborate selection of ad insertion points. In other words, they focus on targeted advertising rather than contextual advertising.

In recent years, text-based contextual advertising efforts such as Google's AdWord/AdSense and Yahoo!'s Contextual Ad programs have become a substantial source of web revenue. These platforms automatically find prominent keywords from user's search query or on a web page, match these keywords against keywords associated with ads provided by advertisers, and then display contextually relevant ads to the user. The approach associating ads with user's query is referred to as keyword-targeted advertising (e.g. AdWords) [1], while associating ads with a web page is referred to as content-targeted advertising (e.g. AdSense and Contextual Ad) [14] [25] [33]. Although keyword-targeted advertising is effective, it is deemed that the use of web content information can allow more relevant ads to be displayed. Ribeiro-Neto *et al.* studied 10 strategies and evaluated their effectiveness for content-targeted advertising [25]. Lacerda *et al.* proposed a learning-based framework to apply Genetic Programming to select the most appropriate ads with respect to a given web page [14]. Another content-targeted advertising system is proposed by Yih *et al.* to learn how to extract keywords from a web page for ad targeting [33].

In summarize, most related work focuses on domain-specific ad placement and text-based web page advertising. Contextual video advertising or that associates video ads within an online video content has not yet been studied adequately. Compared with text, video has distinctive characteristics such that text-based advertising cannot be directly applied to video advertising. First, video is an information-intensive media embedding multimodal tracks. Thus, we argue that contextual relevance between ads and videos should not be measured only based on textual keywords. In addition to textual relevance, both visual and aural relevance should be taken into consideration for relevance matching. Second, video content is represented as a temporal sequence. Ad insertion positions within the video steam should be elaborately selected to support less intrusive advertising. Therefore, there is an urgent demand for an online video advertising platform in which the following three problems are effectively addressed: (1) how to select relevant video ads based on video-related information; (2) how to detect a set of less intrusive ad insertion points within the video content; (3) how to match selected ads to these insertion points to maximize overall contextual relevance.

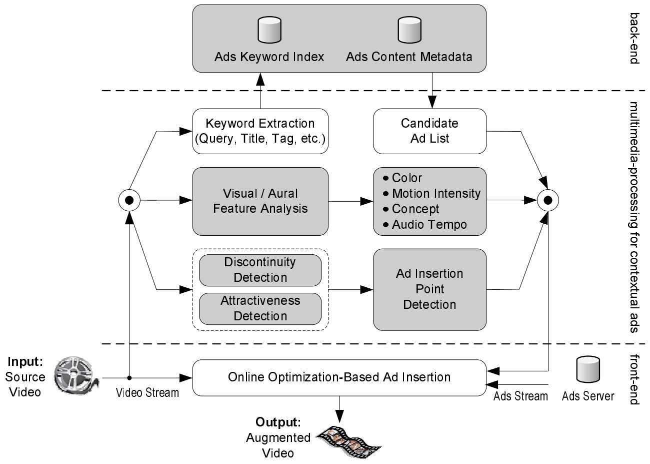


Figure 2: System overview of VideoSense. The gray blocks are performed offline, while the white ones are performed online.

3. SYSTEM OVERVIEW

3.1 Preliminaries

To clearly present the system framework of VideoSense, the following terms are clarified:

- **Video ad**: A video advertisement clip provided by advertisers that can be inserted into or associated with a source video. Although video ads will be associated with video in VideoSense, they may be in different forms (or a combination of forms) including typical ad clips in TV programs, animations, images, or text.
- **Source video**: Source video is most often produced or owned by content providers, which may be professional videographers or grassroots. Video ads will be embedded at appropriate positions in the source video.
- **Ad insertion point**: A point/position in the timeline of a source video at which one or more ad clips will be inserted, as yellow bars in the timeline in Fig. 1(b).

3.2 System Overview

Figure 2 illustrates the system overview of VideoSense. The system consists of three main parts: a back-end for building ad keyword index and extracting ad metadata; a multimedia-processing-end for selecting relevant ads, as well as detecting candidate ad insertion points and extracting visual-aural features for source videos; and a front-end for online matching ads and insertion points. The back-end builds ad keyword index, extracts a set of content features (i.e. color, motion, audio tempo, and concepts), and stores them in ad metadata database. The keywords of ads include title and tags provided by advertisers, as well as automatically recognized categories. The keyword index is used for online ranking of an ad list in the multimedia-processing-end, while the metadata is used for online ad insertion in the front-end. In the multimedia-processing-end, for a given source video, a list of candidate ads is generated and ranked according to global textual relevance, and a set of candidate ad insertion points are detected based on content discontinuity and attractiveness. Meanwhile, a set of content features

Thumbnail	Title	Tag	Tag Expansion	Categories and Confidence

	Benz	Benz car	Benz,
			Mercedes Benz
				vehicles, automobiles, p = 0.0404;
				news and magazines, automotive, p = 0.0352;
				sports and recreation, hobbies, cars, p = 0.0270;

	Mummy Return	Movie preview	movie previews, movie releases, movie news, video trailers, trailers, movie wallpaper, movie, video	arts and entertainment, performing arts, p = 0.0102;
				arts and entertainment, movies, p = 0.0099;
				arts and entertainment, digital art, p = 0.0096;

Figure 3: Ad textual information examples

is extracted. In the front-end, given a ranking list of candidate ads and a set of candidate insertion points, online ad insertion is formulated as an optimization problem, which aims at selecting a subset of insertion points and ads to maximize contextual relevance and minimize intrusiveness. Finally, a description file is generated for augmented video, in which the most relevant ads are embedded at the most appropriate positions within the source video.

4. CANDIDATE AD RANKING

In contextual advertising, ads are expected to be relevant to source videos. Given a source video, a list of candidate ads is desired to be returned from an ad database and ranked according to textual relevance. Since a video segment usually contains few textual descriptions, as well as textual information have quite a few non-informative terms, it is reasonable to rank ads based on textual information related to the whole video instead of a video segment.

In VideoSense, we consider textual information of a video (either a source video or an ad) consisting of the following parts: query (if a source video is reached by searching through a query), title, tags (a textual description provided by content providers or advertisers), and closed captions (if available) [2]. In contrast to web pages that embed enriching information, the amount of textual information related to video is usually limited. Therefore, we perform tag expansion and leverage automatic text categorization to obtain more relevant descriptions. Figure 3 shows two examples of textual information in ads.

From Figure 3, we can see that video-related textual information can be classified into two types: (1) direct text, referring to the query, title, tags, expanded tags, and closed captions; (2) indirect text, referring to categories and their corresponding probabilities obtained by automatic text categorization [3]. Thus, a video-related textual document D can be represented by (k_1, k_2, …, k_t; c_1, c_2, …, c_m), where (k_1, k_2, …, k_t) denote the set of index keywords of direct text in which each keyword k_i can be associated with a weight w_i indicating the importance of k_i, (c_1, c_2, …, c_m) denote the set of predefined categories of indirect text in which each category c_i can be associated with a probability p_i, t and m denote the number of index keywords and categories, respectively. We adopt the vector model and probabilistic model to measure the relevance based on direct and indirect text, respectively. Let D_x and D_y denote two textual documents, the textual relevance $R(D_x, D_y)$ between D_x and D_y is defined as the average of relevance from vector and probabilistic models

$$R(D_x, D_y) = \frac{1}{2}(R_{vec}(D_x, D_y) + R_{prob}(D_x, D_y)) \qquad (1)$$

where $R_{vec}(D_x, D_y)$ and $R_{prob}(D_x, D_y)$ denote the relevance from vector and probabilistic models, respectively.

4.1 Vector Model for Textual Relevance

Ribeiro-Neto *et al.* [25] have proved that matching the ads based on vector model with direct text is the best among a set of simple methods for ad ranking. In vector model, a document D is represented as a vector of weights $(w_1, w_2, \ldots, w_t)$, and the similarity between two documents D_x and D_y is evaluated by the cosine of the angle between the vectors [2]. That is

$$R_{vec}(D_x, D_y) = \frac{\omega(D_x) \cdot \omega(D_y)}{\|\omega(D_x)\| \cdot \|\omega(D_y)\|} \qquad (2)$$

where $\omega(\cdot)$ denote the weighting vector of index keywords.

In general, the weights of index keywords can be calculated in many different ways. A typical way is to use the product of *term frequency* (*tf*) and *inverted document frequency* (*idf*), based on the assumption that the more frequently a word appears in a document and the rarer the word appears in all documents, the more informative it is [2]. However, such an approach is not suitable in our scenario. First, the number of keywords related to an online video is generally much less than that in a regular textual document, which leads to a small *document frequencies* (*df*) and an unstable *idf* [4]. Second, most online content providers tend to use general keywords rather than specific keywords to describe their video content. For instance, many would like to use "car" instead of "Benz" to describe a video clip. Using *idf* will make some non-informative keywords overwhelm the informative ones. Therefore, we only use *tf* to measure the weight of a keyword. Specifically, the weight w_i of keyword k_i is measured by its *term frequency*.

4.2 Probabilistic Model for Textual Relevance

Although vector model is able to represent the keywords of a textual document, it is not enough to describe the latent semantic within a textual document related to a video. For example, a music video named "flower" may be associated with ads related to real flowers instead of more relevant ads related to music albums. This is because "flower" is an important keyword and has a high weight in vector model. To address this problem, we propose a probabilistic model to leverage the categories and corresponding probabilities. We use text categorization based on SVM [32] to automatically classify a textual document into a set of predefined category hierarchy which consists of more than $1k$ categories.

In probabilistic model, a document D is represented as a vector of probabilities (p_1, …, p_m). The predefined categories make up a hierarchical category tree, as shown in Figure 4. Let $d(c_i)$ denote the depth of category c_i in this category tree, where the depth of root is 0. For two categories c_i and c_j, we define $\ell(c_i, c_j)$ as the depth of their first common ancestor. Then for two textual documents D_x and D_y which are represented by (p_1^x, p_2^x, …, p_m^x) and (p_1^y, p_2^y, …, p_m^y), the relevance between D_x and D_y in probabilistic

[2] It is a typical scenario that a user input a query to search some online videos. It is also a commonplace that users or advertisers upload their source videos or video ads, and tag their videos or ads by a set of keywords.

[3] In our work, a document can correspond to multiple categories.

[4] For example, *idf* can be given by $\log(\frac{1}{df})$ [2].

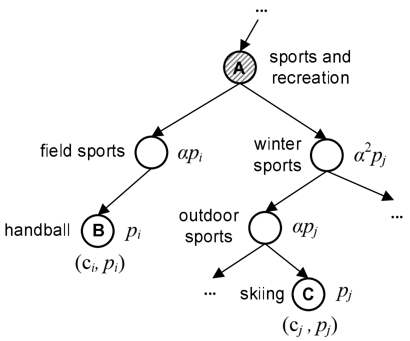


Figure 4: An hierarchical category tree. Each node denotes a category. A is the first common ancestor of B and C.

model is defined as

$$R_{prob}(D_x, D_y) = \sum\nolimits_{i=1}^{m} \sum\nolimits_{j=1}^{m} R(c_i^x, p_i^x; c_j^y, p_j^y) \qquad (3)$$

where $R(c_i^x, p_i^x; c_j^y, p_j^y) = \alpha^{(d(c_i^x)-\ell(c_i^x, c_j^y))} p_i^x \cdot \alpha^{(d(c_j^y)-\ell(c_i^x, c_j^y))} p_j^y$, if $\ell(c_i^x, c_j^y) > 0$; otherwise, 0. α is a predefined parameter to control the probabilities of upper-level categories. In our experiments, α is fixed to 0.5. Intuitively, the deeper level two documents are similar at, the more related they are. For instance, in Figure 4, for the two nodes "B" and "C" represented by (c_i, p_i) and (c_j, p_j), thus $R(c_i, p_i; c_j, p_j) = \alpha^2 p_i \cdot \alpha^3 p_j = \alpha^5 p_i p_j$.

5. AD INSERTION POINT DETECTION

Before detection of insertion points, a pre-processing step is assumed to parse source video into shots and represent each shot by a key-frame using the color-based method [35]. Since ads will be inserted into source videos, a process to elaborately detect a set of insertion points is desirable. The appropriate insertion points will reduce viewers' sensation of intrusiveness while watching augmented video content. Li *et al.* [16] examined eight factors that affect consumers' perceptions of the intrusiveness of ads in traditional TV programs. We excerpt two computable measurements based on these eight factors to intrusiveness, i.e. content *discontinuity* and *attractiveness*. Discontinuity measures the content "dissimilarity" between the two shot series at the two sides of a shot boundary, while attractiveness measures the "importance" or "interestingness" of the content in a shot. Different combinations of discontinuity and attractiveness fit the requirements of different roles. For example, it is intuitive that ads are expected to be inserted at the shot boundaries with high discontinuity and low attractiveness from the viewers' perspective. On the other hand, "high discontinuity plus high attractiveness" may be a tradeoff between viewers and advertisers (which will be part of our future investigations). Then, the detection of ad insertion points can be formulated as ranking the shot boundaries based on different combinations of content discontinuity and attractiveness.

5.1 Discontinuity Detection

The research problem close to content discontinuity detection is video segmentation such as story/scene/shot detection. Although many efforts have been conducted for video segmentation, most of them treated "discontinuity" as a hard decision, that is, whether a detected boundary is true or not. In VideoSense, we seek a soft measure of a shot boundary belonging to an ad insertion point.

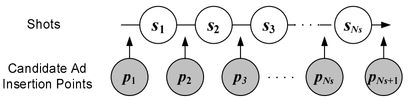


Figure 5: Candidate ad insertion points within a source video

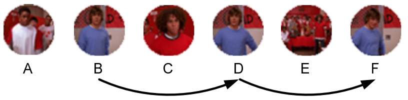


Figure 6: The interlaced repetitive pattern of shots

Suppose a source video is composed of N_s shots $\{s_i\}_{i=1}^{N_s}$, thus there are N_s+1 candidate ad insertion points $\{p_j\}_{j=1}^{N_s+1}$, say, $N_s - 1$ shot boundaries together with the beginning and the end of video, as shown in Figure 5. A degree of discontinuity $D(p_j)$ is assigned to each point p_j. The higher the discontinuity is, the more likely the corresponding insertion point is a boundary of two episodes.

Therefore, the detection of content discontinuity can be performed similarly to traditional video scene/story segmentation. Zhao *et al.* proposed Best-First Model Merging (BFMM) for scene segmentation in which visually similar shots are gradually merged [36]. We adopt BFMM for discontinuity detection since the merge order in BFMM can be regarded as the degree of discontinuity. However, BFMM cannot deal with interlaced repetitive patterns of shots as only the similarity of adjacent shots is considered. For an example in Figure 6, the shots "B" and "D" are supposed to first be merged as they depict the same person. But they are merged later because BFMM only consider the similarity between "B" and "C" or "C" and "D". To address this problem, we propose an improved BFMM (so-called iBFMM) to deal with the interlaced repetitive pattern. The novelties of iBFMM lie in: (1) A preprocess step is added to BFMM to group the interlaced repetitive shots in a multiscale manner; and (2) the normalized merge order of shots around each candidate insertion point is adopted as the measurement of its discontinuity. The algorithm for iBFMM is given in Algorithm 1. In the preprocess step, the most similar shots are merged at different scales to eliminate interlaced repetitive pattern. In the BFMM and normalization step, the merge order is recorded and normalized as the final discontinuity. Although we proposed iBFMM for discontinuity detection in this paper, any method that generates soft measure for shot boundary can be employed here.

5.2 Attractiveness Detection

In general, it is difficult to evaluate how a video clip attracts viewers' attention since "attractiveness" or "attention" is a neurobiological concept. Alternatively, a user attention model is proposed by Ma *et al.* [18] to estimate human's attention by integrating a set of visual, auditory and linguistic elements related to attractiveness, such as motion, objects (faces), static attention regions, audio and language. Another approach to exploring the attention in video sequence is averaging static image attention over a segment of

Algorithm 1 iBFMM for discontinuity detection

```
Input: $\mathcal{S} = \{s_i\}_{i=1}^{N_s}$, $\mathcal{P} = \{p_j\}_{j=1}^{N_s+1}$
Output: $\mathcal{D} = \{D(p_j)\}_{j=1}^{N_s+1}$
1: Initialize: set $D(p_{N_{s+1}}) = 1.00$, $D(p_1) = 0.99$.
2: Preprocess
     for scale $\sigma$=1 to 4 do
       compute similarity $Sim(s_i, s_{i+\sigma})$ for each pair [36]
       if $Sim(s_i, s_{i+\sigma}) < T_s$ do
         merge $\{s_k\}_{k=i+1}^{i+\sigma}$ to $s_i$, remove $\{s_k\}_{k=i+1}^{i+\sigma}$ from $\mathcal{S}$
         set $\{D(p_k)\}_{k=i+1}^{i+\sigma} = 0$, $N_s = N_s - (\sigma - 1)$
       end if
     end for
3: BFMM
     set merging order $N_m = 1$
     while $N_s > 0$ do
       compute $Sim(\cdot)$ for adjacent shots, get closest $(s_i, s_{i+1})$
       merge $s_{i+1}$ to $s_i$, remove $s_{i+1}$ from $\mathcal{S}$
       set $D(p_i + 1) = N_m$, and $N_m = N_m + 1$, $N_s = N_s - 1$
     end while
4: Normalize: set $\{D(p_j) = \frac{D(p_j)}{N_m}\}_{j=2}^{N_s}$
```

frames [19]. In VideoSense, we compute an attention value $A(s_i)$ for each shot s_i by the user attention model in [18]. The content attractiveness of insertion point p_i is highly related to the neighboring shots on both sides of p_i. Therefore, the attractiveness of p_i can be computed by weighted averaging the attention values of its neighboring shots as follows.

$$A(p_i) = \sum\nolimits_{k=-\delta}^{\delta} \alpha_{|k|} A(s_{i+k}) \qquad (4)$$

where $\sum_k \alpha_k = 1$ and $1 > \alpha_0 > \alpha_1 > \ldots > \alpha_\delta > 0$, since the nearer the neighboring shot is, the more effect it has on the attractiveness. In our implementation, we empirically set $\delta = 2$, and (α_0, α_1, α_2)=(0.4, 0.2, 0.1).

5.3 Ad Insertion Point Detection

Figure 7 gives an example of discontinuity and attractiveness detection for a feature film. One way for detecting ad insertion points can be finding peaks at the combined curve with "discontinuity minus attractiveness." However, the detection of ad insertion point should be based not only on discontinuity and attractiveness, but also on the temporal distribution of these points, as well as the global and local relevance between source video content and ad content. In other words, the selected ad insertion points should comply with contextually relevant and less intrusive advertising strategy. Thus, we integrate discontinuity and attractiveness for ad insertion point detection into an optimization framework by considering a set of rules from the viewers' perspective. This will be detailed in the next section.

6. ONLINE OPTIMIZATION-BASED AD INSERTION

The typical scenario in VideoSense can be described as follows: A user clicks an online video, then the augmented video with embedded ads is immediately returned to this user. An alternative scenario is that the ads can be associated with an online video when it is uploaded, without considering viewers' profiles and new coming ads provided by advertisers. We focus on the first scenario since it is more reasonable and extensible for current online advertising. Hence given the online video with its candidate ad insertion points and a ranking list of candidate, the optimization-based ad insertion should be performed in real time.

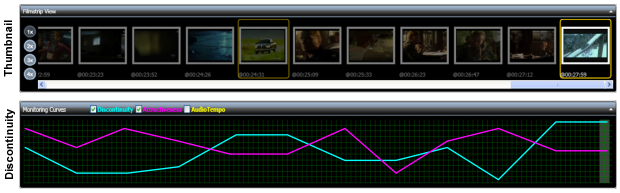


Figure 7: An example of discontinuity and attractiveness detection. Each thumbnail in the above figure corresponds to a shot. The thumbnails highlighted with yellow box indicate inserted ads. The curves in the below figure indicate discontinuity (cyan) and attractiveness (red).

The contextual relevance between source video and embedded video ads consists of two parts: (1) textual relevance between the source video and embedded ads, as described in Section 4; and (2) visual-aural relevance between each ad and the neighboring source video content on both sides of corresponding insertion point. The textual relevance is referred to as *global* relevance in this paper as it measures the correlation between a source video and an ad, while the visual-aural relevance is referred to as *local* relevance since it measures the correlation between a segment of source video and an ad at a local point within the source video. Before formulating ad insertion problem, we introduce the computation of local relevance at first.

6.1 Local Visual-Aural Relevance

As illustrated in Figure 2, the local visual relevance is measured by a set of low-level features such as motion intensity and color, as well as high-level semantic concepts, while the aural relevance is derived from audio tempo [5]. These features have proved to be effective to describe video content in many existing multimedia applications [4] [9]. More sophisticate low-level features related to visual-aural relevance can be also applied here. The local visual-aural relevance leads to non-invasive sensation when users are interrupted by relevant ads. Actually, we suggest various ways for using local relevance in VideoSense. Specifically, we can use the "positive" local relevance to keep high similarity between video and ad content, or use it in a "negative" way to gain more attention from viewers because of the high contrast. We choose the "positive" way in this paper since it is more natural for the viewers. For example, when viewing an online music video, users may feel that an ad with similar music tempo and the same concept (such as "Entertainment") does not disrupt their experiences.

Since local relevance indicates the visual-aural similarity between a source video shot and a video ad, the set of visual and aural features is computed at video level by averaging the features over all shots for ad, rather than computed at shot level for source video. Let $\mathcal{F} = \{color, motion, concept,$

[5] We select 16 concepts appearing frequently from TRECVID 2006 [28], including "Building," "Car," "Entertainment," "Face," "Government-Leader," "Meeting," "Military," "Mountain," "Office," "Person," "Road," "Sky," "Sports," "Studio," "Vegetation," and "Waterscape-Waterfront." The concept models are built based on our previous work submitted to TRECVID 2006 [10]. The 16-D concept probabilities constitute a feature vector.

tempo} denote the feature set and $R_f(s_i, a_j)$ denote the local relevance between the source video shot s_i and ad a_j in terms of feature f ($f \in \mathcal{F}$), $R_f(s_i, a_j)$ can be computed as the intersection (i.e., similarity) of f. Then the local relevance $R_\ell(s_i, a_j)$ between s_i and a_j is given by

$$R_\ell(s_i, a_j) = \max_f \{R_f(s_i, a_j)\} \tag{5}$$

It is reasonable that there is high relevance between s_i and a_j if one type of their features depict similar. Accordingly, the local relevance between an ad insertion point p_i and ad a_j is decided by the visual-aural similarity between the shots beside p_i and a_j as follows

$$R_\ell(p_i, a_j) = \lambda R_\ell(p_i^-, a_j) + (1 - \lambda) R_\ell(p_i^+, a_j) \tag{6}$$

where p_i^- and p_i^+ denote the neighboring shots before and behind p_i, and λ ($0 < \lambda < 1$) controls the strength of relevance from the both sides of p_i. Perceptually the relevance from p_i^- has more contribution to the final local relevance as the content of p_i^- is viewed before the ad is displayed. Thus the constant λ can be set bigger than 0.5. The relevance from the both sides of p_i is given by

$$\begin{aligned} R_\ell(p_i^-, a_j) &= \sum\nolimits_{k=1}^{W} w^k R_\ell(s_{i-k}, a_j) \\ R_\ell(p_i^+, a_j) &= \sum\nolimits_{k=1}^{W} w^k R_\ell(s_{i+k-1}, a_j) \end{aligned} \tag{7}$$

where w ($0 < w < 1$) is the summing weight and W is the size of neighboring window. In our implementation, $\lambda = 0.80$, $w = 2/3$, and $W = 3$.

6.2 Problem Formulation

Let $\mathbf{V}$ denote the source video which consists of N_s shots represented by $\mathcal{S} = \{s_i\}_{i=1}^{N_s}$, accordingly there are N_p ($N_p = N_s + 1$) candidate ad insertion points which is represented by $\mathcal{P} = \{p_i\}_{i=1}^{N_p}$, as discussed in Section 5.1. Each insertion point p_j has a degree of discontinuity $D(p_i)$ and attractiveness $A(p_i)$, as discussed in Section 5.2. Given $\mathbf{V}$, a list of candidate ads $\mathcal{A} = \{a_j\}_{j=1}^{N_a}$ is ranked according to global textual relevance, as discussed in Section 4. Each ad a_j has a degree of global relevance $R_g(a_j, \mathbf{V})$, which is computed by equation (1). For the sake of simplicity, we neglect $\mathbf{V}$ since it is given in the formulation. Therefore, the global relevance of a_j can be written as $R_g(a_j)$. Moreover, for a pair of (p_i, a_j), a local relevance $R_\ell(p_i, a_j)$ can be give by equation (6).

The problem of online ad insertion can be described as given a set of insertion points $\mathcal{P}$ and a list of ranked ads $\mathcal{A}$, to select N elements from $\mathcal{P}$ and $\mathcal{A}$, respectively, and to associate each $a_j \in \mathcal{A}$ with an appropriate $p_i \in \mathcal{P}$. N is the number of expected ads to be inserted, which can be given by source video content providers. To support contextually relevant and less intrusive advertising from viewers' perspective, three computable objectives can be expressed as below:

(1) The overall *contextual relevance* (including the *global* relevance from selected ads and *local* relevance from pairs of selected insertion point and ad) is maximized;

(2) The overall *attractiveness* of selected insertion points is minimized, while the *discontinuity* is maximized;

(3) The selected insertion points are *uniformly* distributed.

Suppose we introduce the following design variables $\mathbf{x} \in \mathbb{R}^{N_p}$, $\mathbf{y} \in \mathbb{R}^{N_a}$, $\mathbf{x} = [x_1, \ldots, x_{N_p}]^T$, $x_i \in \{0, 1\}$, and $\mathbf{y} = [y_1, \ldots, y_{N_p}]^T$, $y_j \in \{0, 1\}$, where x_i and y_j indicate whether p_i and a_j are selected ($x_i = 1, y_j = 1$). The above problem can be formulated as the following nonlinear 0-1 integer programming problem (NIP) [3].

Algorithm 2 The heuristic searching algorithm for Eq. (8)

```
1: Initialize: set the labels of all the elements in $\mathbf{x}$ and $\mathbf{y}$ as "0"
   (i.e. "not selected").
2: Among all the elements labeled as "0" in $\mathbf{x}$, select the maximal
   $u_i$, and set $x_i = 1$. This is to make objective (2) satisfied.
3: For each element $x_k$ in $\mathbf{x}$ falling into $[x_i - N_p/2N, x_i + N_p/2N]$,
   set $u_k = u_k - 1.0$. Thus these elements will not
   be selected in the next loop, which assures that objective (3)
   is satisfied.
4: Among all the elements (i.e. ranked ads) labeled as "0" in $\mathbf{y}$,
   select $y_j$ with the maximal $(x_i y_j r_{ij})$, and set $y_j = 1$. This is
   to make objective (1) satisfied.
5: If $\sum_{k=1}^{N_p} x_k = N$, output all the pairs of $(x_i, y_j)$; otherwise
   return to step 2.
```

$$\begin{aligned} \max_{(x_i, y_j)} f(\mathbf{x}, \mathbf{y}) &= \alpha \sum_{i=1}^{N_p} x_i (D(p_i) - A(p_i)) \\ &\quad + \beta \sum_{i=1}^{N_p} \sum_{j=1}^{N_a} x_i y_j R_g(a_i) R_\ell(a_i, p_j) + \gamma E_n(\mathbf{x}) \\ &= \alpha \mathbf{x}^T \mathbf{u} + \beta \mathbf{x}^T \mathbf{R} \mathbf{y} + \gamma E_n(\mathbf{x}) \\ s.t. \quad & \sum_{i=1}^{N_p} x_i = N, \ \sum_{j=1}^{N_a} y_j = N, \ x_i, y_j \in \{0, 1\} \end{aligned} \tag{8}$$

where $\mathbf{R} \in \mathbb{R}^{N_p \times N_a}$, $\mathbf{R} = [r_{ij}]$, $r_{ij} = R_g(a_j) R_\ell(p_i, a_j)$, $\mathbf{u}$=$[u_1, u_2, \ldots, u_{N_p}]^T$, $u_i = D(p_i) - A(p_i)$, and $E_n(\mathbf{x})$ is an entropy-like problem measuring the distribution uniformity as follows

$$E_n(\mathbf{x}) = -\frac{1}{\log N} \sum_{k=1}^{N_p - 1} (p_{\phi(x_{k+1})} - p_{\phi(x_k)}) \log(p_{\phi(x_{k+1})} - p_{\phi(x_k)})$$

where $\phi(x_k) : x_k \mapsto i \in \{1, \ldots, N_p\}$ is an index function indicating the location of k-th nonzero x_k in $\mathbf{x}$. Notice that, here the total length of $\mathbf{V}$ is 1, and $p_i \in [0, 1]$ denote the insertion point. The parameters (α, β, γ) controls the strength from different constraints, which satisfying $0 \leqslant \alpha, \beta, \gamma \leqslant 1$ and $\alpha + \beta + \gamma = 1$.

6.3 Problem Solution

It is observed that there are $C_{N_p}^N C_{N_a}^N N!$ solutions in total to equation (8). As a result, when the number of elements in $\mathcal{A}$ and $\mathcal{P}$ is large, the searching space for optimization increases dramatically. However, we can use the *Genetic Algorithm* (GA) [30] to find solutions approaching the global optimum. Alternatively, the above problem can be solved by a similar heuristic searching algorithm in practice, which is given in Algorithm 2. Although a local optimal solution to equation (8) can be achieved in Algorithm 2, the number of possible solutions can be significantly reduced to $C_{N_p}^1 C_{N_a}^1 N!$.

7. EXPERIMENTS AND EVALUATIONS

In VideoSense, the candidate ad ranking and optimization-based ad insertion can be performed on-the-fly. In our implementation, a fixed number of candidate ads (i.e. $N_a = 200$) is returned, and the number of inserted ads within the source

Table 1: The source videos used for evaluations.

Video	#	Length (min.)	# of Shots	# of Ads
Micro Video	20	47	557	40
Home Video	4	182	1037	36
Movie Clip	4	110	1784	22
Documentary	2	60	220	12

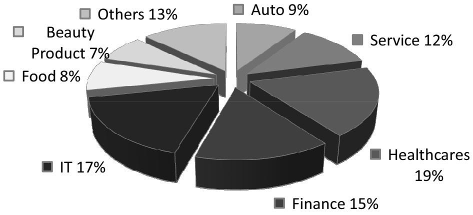


Figure 8: The distribution of ad category

video $\mathbf{V}$ is given by $N = \max(|\mathbf{V}|/5, 2)$, where $|\mathbf{V}|$ is the duration in minutes. It takes around 0.5 seconds for online ranking of 200 candidate ads and less than 0.2 seconds for ad insertion using Algorithm 2. Each embedded ad will only be displayed for up to 10 seconds as suggested by P. Horan that "the point at which a consumer has patience for an online video ad is 10 seconds" [22]. To validate the capability of VideoSense supporting contextually relevant and less intrusive advertising, we conduct an extensive objective experiments and comparisons for ad insertion point detection, and subjective experiments for evaluating contextual relevance and viewing experience of augmented videos.

7.1 Dataset

We collected more than 14k source videos which consists of more than 13k online micro videos from the most popular video sharing site, i.e. YouTube [34], and about 50 long videos such as movie clips, home videos, and documentaries. We select 32 videos for evaluations. These consist of 20 micro videos searched by top 10 representative queries from our video database, four movie clips, four home videos, and two documentaries, as listed in Table 1.The selected 10 representative queries come from the most popular queries in a video site, including "flowers," "cat," "baby," "sun," "soccer," "fire," "beach," "food," "car," and "Microsoft." For each query, only top two videos are selected for evaluations.

We have also collected 1028 unique video ads with the total duration of 547 minutes from 277 news programs in TRECVID 2006 corpus and TV programs of TNT channel. These ads cover a variety of categories defined in [6], as shown in Figure 8. The title and keywords of each ad are manually annotated in our experiment.

7.2 Objective Evaluation on Ad Insertion Point Detection

To evaluate the detection of ad insertion point, we firstly compare the results of iBFMM for video scene detection with BFMM [36] and Graph-based method [23] in terms of content discontinuity. It is easy to obtain objective benchmarks for scene detection using these methods, and there is little work on automatic detection of ad insertion points. The three methods used for scene detection can be divided into two paradigms based on processing manners: merging-based (BFMM and iBFMM), and splitting-based (Graph). The experiments were carried out on the 10 long videos. We did not use the micro videos for this evaluation since usually there are only two insertion points detected (i.e. the beginning and the end of source videos) in these videos due to their limited durations. The scene boundaries are manually labeled. The number of scenes in BFMM and iBFMM is set as same as the output of Graph. The "tolerance" [23] is adopted as the offset of detected boundaries, say, the detected boundary is regarded as true positive if the offset is less than "tolerance." As the average duration of shot in the long videos is 6.9 seconds, we compare the results in different settings of tolerance, i.e. 1 second and 9 seconds. The performance is validated by three measurements, i.e., *precision*, *recall* and F_1 ($F_1 = \frac{2 \times precision \times recall}{precision + recall}$). The results are listed in Table 2. It is observed that iBFMM achieves the best performance among the three methods, which supports the effective detection of ad insertion point in VideoSense.

Table 2: The comparisons among Graph, BFMM, and iBFMM for video scene detection.

Method	tolerance = 1 sec			tolerance = 9 sec		
	recall	*prec.*	F_1	*recall*	*prec.*	F_1
Graph [23]	0.36	0.30	0.32	0.55	0.46	0.50
BFMM [36]	0.40	0.41	0.40	0.59	0.60	0.50
iBFMM	0.47	0.48	0.47	0.63	0.63	0.63

Furthermore, we compare the results of content discontinuity detection between iBFMM and BFMM. We invited five annotators to label the confidence of scene boundary (i.e., the probability that the annotator regards a detected shot boundary as a ground truth of ad insertion point) on the set of long videos. The annotation results are averaged as the ground truth of ad insertion points. The performance is validated by a non-interpolated average precision (AP), which is widely used as a measure of retrieval effectiveness [28]. Suppose the boundaries are ranked according to content discontinuity, the AP is given by $AP(n) = \frac{1}{R_n}\sum_{j=1}^{n}(\frac{R_j}{j}I_j)$, where R_n is the number of true boundaries in a size of n, $I_j = 1$ if the j-th boundary is true and 0 otherwise. The AP corresponds to the area under the (non-interpolated) recall/precision curve, and incorporates the effect of recall when it is computed over the entire result set. The AP results among all videos are averaged as mean average precision (MAP). The different results of MAP(n) are shown in Figure 9. It is observed that iBFMM outperforms BFMM in all settings of n.

7.3 Subjective Evaluation on Ad Insertion

As objective evaluation of ad insertion performance is difficult, we conducted a subjective user study to evaluate our work. Twelve evaluators majoring in computer science were invited to participate in the user study, including four graduate students, four undergraduate students, and four researchers. All of them are familiar with several online video sites and have ever watched online videos. Each individual was assigned with eight videos and asked to get familiar with the content in advance. The eight videos consisted of five micro videos, a home video, a movie, and a documentary. In order to evaluate the effectiveness of Ad insertion from different perspectives, we rewrite equation (8) as follows:

$$\max_{(x_i, y_j)} f(\mathbf{x}, \mathbf{y}) = \lambda_1 \alpha \mathbf{x}^T \mathbf{u} + \lambda_2 \beta \mathbf{x}^T \mathbf{R} \mathbf{y} + \gamma E_n(\mathbf{x}) \qquad (9)$$

where $\lambda_i \in \{-1, +1\}$ ($i = 1, 2$) is an indicator. Clearly,

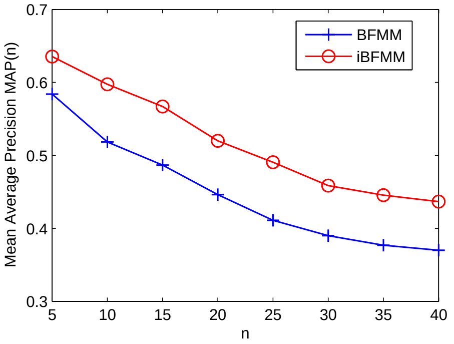


Figure 9: The comparisons of MAP between iBFMM and BFMM

the different settings of (λ_1, λ_2) will significant affect the ad insertion strategy. The following five results were randomly given to the evaluators according to different settings of (λ_1, λ_2) and r_{ij} in equation (9).

- **I: Radom** $(\lambda_1 = 0, \lambda_2 = 0)$. Randomly select N ads from candidate ad list, and randomly insert them at a fixed interval of $|\mathbf{V}|/N$, as the traditional TV *spot*, which can regarded as the baseline for comparison.
- **II: Less global relevance and more intrusiveness** $(\lambda_1 = -1, \lambda_2 = -1, r_{ij} = R_g(a_j))$. Select the most irrelevant ads from candidate ad list, and insert them at the positions with low discontinuity and high attractiveness, without considering local relevance.
- **III: More global relevance and less intrusiveness** $(\lambda_1 = +1, \lambda_2 = +1, r_{ij} = R_g(a_j))$. Select the most relevant ads from candidate ad list, and insert them at the positions with high discontinuity and low attractiveness, without considering local relevance.
- **IV: Less global and local relevance, and more intrusiveness** $(\lambda_1 = -1, \lambda_2 = -1)$. It is similar to II, except for considering local relevance here.
- **V: More global and local relevance, and less intrusiveness** $(\lambda_1 = +1, \lambda_2 = +1)$. It is similar to III, except for considering local relevance here. This is the basic scenario supported by VideoSense.

The evaluators have no knowledge of current settings. When viewing each of the five results, the evaluators were asked to give a score from 1 to 5 (higher score indicating better satisfaction) to show their satisfactions level based on the following aspects:

- **Local relevance**. For each inserted ad, how did you feel about the local relevance between the ad and its surround content?
- **Comfortableness**. For each inserted ad, did you feel comfortable as you viewed the ad?
- **Satisfaction**. For each source video, what was your level of overall satisfaction in how the ads were inserted?

Furthermore, the evaluators were required to give a score of the global relevance of candidate ad list with respect to source video. Since the returned ad ranking lists for the five results were identical, an evaluator had to do this once for each source video. As a result, the average global relevance is 2.71 (for micro videos) and 3.34 (for long videos). This is because that the online micro videos usually contain quite a little random textual information.

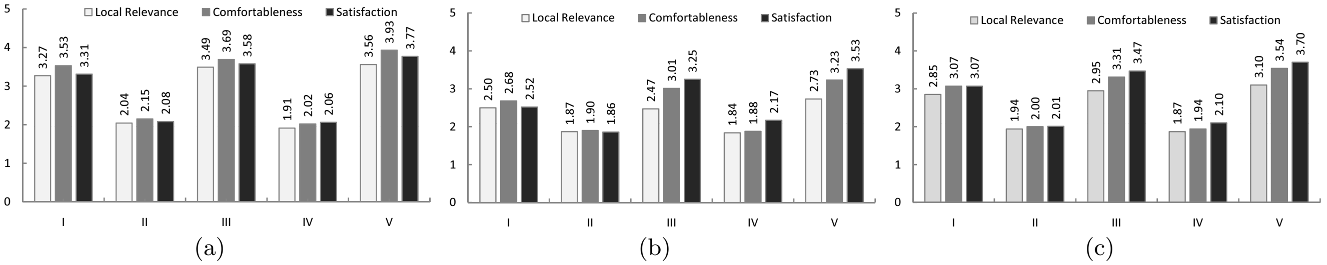
The average results of the above three questions are listed in Figure 10. In general, the evaluations of V (i.e. the results of VideoSense) achieve the best among the five results. We can see that IV achieves the worst evaluation. This observation has proved that contextual relevance (i.e. global textual relevance and local visual-aural relevance) and ad insertion points significantly influence viewers' experiences of augmented videos. In the case of IV, the evaluators felt worst when watching irrelevant ads at the inappropriate positions. The performances of II and IV are lower than I, which demonstrate that lack-of-relevance and intrusiveness can lead to worse perceptive experiences than traditional TV *spot*. The superior performance of III to I also indicates that adding relevance to traditional TV *spot* setting can result in better experience. The different settings of intrusiveness correspond to different combinations of *discontinuity* and *attractiveness* for ad insertion point detection. The lower comfortableness of II and IV than those of III and V demonstrated that the combination of "high discontinuity" and "low attractiveness" is effective for ad insertion point detection from the viewers' perspective. For micro videos, the advertising strategy in III is similar to most current video sites, such as Revver [24] and Youtube [34], in which only global textual relevance is taken into consideration and the insertion points are just the beginning or the end of videos [6]. It is also observed from Figure 10(c) that the average evaluation results of all videos comply with: V>III>I>II>IV. From this viewpoint, we can conclude that VideoSense supports more effective advertising than current video sites.

8. DISCUSSIONS AND CONCLUSIONS

In this paper, we presented VideoSense – a novel video advertising system that is able to support contextually relevant and less intrusive advertising for online video service. To support less intrusiveness, we elaborately detect a set of appropriate ad insertion points based on content discontinuity and attractiveness. To support contextual relevance, we introduce global textual relevance to find the most relevant ads, and local visual-aural relevance to find good matching between each insertion point and ad. The whole process is further formulated as an optimization problem. The objective and subject evaluations proved that VideoSense supports more effective video advertising than current video sites. Furthermore, we discussed some interesting extensions of VideoSense to support more kinds of users.

There are a number of possible improvements for VideoSense. For example, a suitable ontology for video ads will improve the textual matching between ads and videos. Although we propose two measurements, i.e. content discontinuity and attractiveness, for detecting ad insertion point, it still requires additional user studies in a typical viewer audience to

[6] As we have set $D(p_{N_s+1}) = 1.00$ and $D(p_1) = 0.99$ in Algorithm 1, in many cases the ads are inserted at the beginning and end of micro videos.



(a) (b) (c)

Figure 10: Subjective evaluations of (a) micro videos, (b) long videos, and (c) all (micro and long) videos.

know what really a good ad insertion point is. Furthermore, how we can simultaneously take both the viewers and advertisers into consideration still remains a challenging problem.

To date, user-targeted advertising is another key for online advertising in addition to contextual advertising. Targeted video advertising means that video ads will reach specified target audiences by leveraging user-provided demographic profiles. To support such advertising framework, our future work include collecting user profiles and click-through data, and studying how to deliver personalized video ads based on user interests, locations, past and current behaviors, and other data. We also aim at embedding a variety of ads into online videos, including textual, audio and image ads.

9. ACKNOWLEDGMENTS

We would like to thank Mr. Hao Wei, Lin Ma, and Bo Yang for their generous supports to build some modules. We also thank all the participants in our user study experiments.

10. REFERENCES

[1] AdWords. http://adwords.google.com/.

[2] R. Baeza-Yates and B. Ribeiro-Neto. *Modern Information Retrieval*. Addison Wesley, 1999.

[3] S. Boyd and L. Vandenberghe. *Convex Optimization*. Cambridge University Press, 2004.

[4] J.-C. Chen, W.-T. Chu, J.-H. Kuo, C.-Y. Weng, and J.-L. Wu. Tiling slideshow. In *Proceedings of ACM Multimedia*, 2006.

[5] ComScore. http://www.comscore.com/.

[6] L.-Y. Duan, J. Wang, Y. Zheng, J. S. Jin, H. Lu, and C. Xu. Segmentation, categorization, and identification of commercials from TV streams using multimodal analysis. In *Proceedings of ACM Multimedia*, pages 201–210, 2006.

[7] eMarketer. http://www.emarketer.com/.

[8] Google Video. http://video.google.com/.

[9] X.-S. Hua, L. Lu, and H.-J. Zhang. Optimization-based automated home video editing system. *IEEE Trans. on Circuit and Syst. for Video Tech.*, 14(5):572–583, 2004.

[10] X.-S. Hua, T. Mei, W. Lai, and *et al.* Microsoft Research Asia TRECVID 2006 high-level feature extraction and rushes exploitation. In *TREC Video Retrieval Evaluation Online Proceedings*, 2006.

[11] iTVx. http://www.itvx.com/.

[12] G. Kastidou and R. Cohen. An approach for delivering personalized ads in interactive TV customized to both users and advertisers. In *Proceedings of European Conference on Interactive Television*, 2006.

[13] P. Kim. *Advertisers face TV reality*. Forester Research, 2006.

[14] A. Lacerda, M. Cristo, M. A. Goncalves, and *et al.* Learning to advertise. In *Proceedings of ACM SIGIR*, 2006.

[15] G. Lekakos, D. Papakiriakopoulos, and K. Chorianopoulos. An integrated approach to interactive and personalized TV advertising. In *Proceedings of Workshop on Personalization in Future TV*, 2001.

[16] H. Li, S. M. Edwards, and J.-H. Lee. Measuring the intrusiveness of advertisements: scale development and validation. *Journal of Advertising*, 31(2):37–47, 2002.

[17] Y. Li, K. Wan, X. Yan, and C. Xu. Advertisement insertion in baseball video based on advertisement effect. In *Proceedings of ACM Multimedia*, pages 343–346, 2005.

[18] Y.-F. Ma, L. Lu, H.-J. Zhang, and M. Li. A user attention model for video summarization. In *Proceedings of ACM Multimedia*, pages 533–542, 2002.

[19] Y.-F. Ma and H.-J. Zhang. Contrast-based image attention analysis by using fuzzy growing. In *Proceedings of ACM Multimedia*, pages 374–381, Nov 2003.

[20] S. Mccoy, A. Everard, P. Polak, and D. F. Galletta. The effects of online advertising. *Communications of The ACM*, 50(3):84–88, 2007.

[21] Metacafe. http://www.metacafe.com/.

[22] Online Publishers. http://www.online-publishers.org/.

[23] Z. Rasheed and M. Shah. Detection and representation of scenes in videos. *IEEE Trans. on Multimedia*, 7(6):1097–1105, Dec. 2005.

[24] Revver. http://one.revver.com/revver.

[25] B. Ribeiro-Neto, M. Cristo, P. B. Golgher, and E. S. Moura. Impedance coupling in content-targeted advertising. In *Proceedings of ACM SIGIR*, 2005.

[26] C. Rohrer and J. Boyd. The rise of intrusive online advertising and the response of user experience research at Yahoo! In *Proceedings of ACM SIGCHI*, 2004.

[27] A. Thawani, S. Gopalan, and V. Sridhar. Context aware personalized ad insertion in an interactive TV environment. In *Proceedings of Workshop on Personalization in Future TV*, 2004.

[28] TRECVID. http://www-nlpir.nist.gov/projects/trecvid/.

[29] K. Wan, X. Yan, X. Yu, and C. Xu. Robust goal-mouth detection for virtual content insertion. In *Proceedings of ACM Multimedia*, pages 468–469, Nov 2003.

[30] D. Whitley. A genetic algorithm tutorial. *Statistics and Computing*, 4:65–85, 1994.

[31] Yahoo! Video. http://video.yahoo.com/.

[32] Y. Yang and X. Liu. A re-examination of text categorization methods. In *Proceedings of SIGIR*, 1999.

[33] W.-T. Yih, J. Goodman, and V. R. Carvalho. Finding advertising keywords on web pages. In *Proceedings of International World Wide Web Conference*, 2006.

[34] YouTube. http://www.youtube.com/.

[35] H.-J. Zhang, A. Kankanhalli, and S. W. Smoliar. Automatic partitioning of full-motion video. *Multimedia Systems*, 1(1):10–28, June 1993.

[36] L. Zhao, W. Qi, Y.-J. Wang, S.-Q. Yang, and H.-J. Zhang. Video shot grouping using best first model merging. In *Proceedings of Storage and Retrieval for Media Database*, pages 262–269, 2001.