

Contextual Internet Multimedia Advertising

By TAO MEI, *Member IEEE*, AND XIAN-SHENG HUA, *Member IEEE*

ABSTRACT | The advent of media-sharing sites has led to the unprecedented Internet delivery of community-contributed media like images and videos. Those visual contents have become the primary sources for online advertising. Conventional advertising treats multimedia advertising as general text advertising by displaying advertisements either relevant to the queries or the Web pages, without considering the potential advantages which could be brought by media contents. In this paper, we summarize the trend of Internet multimedia advertising and conduct a broad survey on the methodologies for advertising which are driven by the rich contents of images and videos. We discuss three key problems in a generic multimedia advertising framework. These problems are: *contextual relevance* that determines the selection of relevant advertisements, *contextual intrusiveness* which is the key to detect appropriate ad insertion positions within an image or video, and *insertion optimization* that achieves the best association between the advertisements and insertion positions so that the effectiveness of advertising can be maximized in terms of both contextual relevance and contextual intrusiveness. We show recently developed *MediaSense* which consists of image, video, and game advertising as an exemplary application of contextual multimedia advertising. In the *MediaSense*, the most contextually relevant ads are embedded at the most appropriate positions within images or videos. To this end, techniques in computer vision, multimedia retrieval, and computer human interaction are leveraged. We also envision that the next trend of multimedia advertising would be game-like advertising which is more impressionative and thus can promote advertising in an interactive, as well as more

compelling and effective way. We conclude this survey with a brief outlook on open research directions.

KEYWORDS | Computer vision; contextual advertising; multimedia advertising; survey

I. INTRODUCTION

The proliferation of digital capture devices and the explosive growth of online social media (especially along with the so called Web 2.0 wave) have led to the countless private image and video collections on local computing devices, such as personal computers, cell phones, and personal digital assistants (PDAs), as well as the huge yet increasing public media collections on the Internet [8]. For example, the amount of digital images captured worldwide in 2011 will increase from 50 billion in 2007 to 60 billion according to IT Facts' report [44]. The most popular photo sharing site, i.e., Flickr, reached three billion photo uploads at the end of 2008 and 3–5 million new photos uploaded daily [53], [85], while Youtube drew 5 billion U.S. online video views in July 2008 [118]. On the other hand, we have witnessed a fast and consistently growing online advertising market in recently years. Jupiter Research forecasted that online advertising spending will surge to \$18.9 billion by 2010-up, which is about 59 percent from an estimated \$11.9 billion in 2005 [50]. Motivated by the huge business opportunities in the online advertising market, people are now actively investigating new Internet advertising models. To take the advantages of the visual form of information representation, multimedia advertising, which associates advertisements with an online image or video, has become an emerging online monetization strategy.¹

¹Please note that “multimedia advertising” and “advertising multimedia” are two different concepts. By multimedia advertising, we refer to the process of associating advertisements with multimedia, while advertising multimedia indicates using multimedia as the form of advertisement.

In an advertising system, there is usually an intermediary commercial ad-network entity (i.e., a service provider between the publisher and advertiser) in charge of optimizing the ad selection and displaying with the twin goal of increasing the revenue (shared between the publisher and ad-network) and improving user experience. With these goals, it is preferable to the publishers and profitable to the advertisers to have ads relevant to media content rather than generic ads. By implementing a solid multimedia advertising strategy into an existing content delivery chain, both the publishers and advertisers have the ability to deliver compelling content, reach a growing online audience, and eventually generate additional revenue from online media.

Advertising has embarked on a dramatic evolution, which will be rapid, fundamental, and permanent. Although this evolution is still underway in advertising in terms of objectives, strategy, and solutions, we can summarize the trends of Internet advertising into two generations in terms of methodologies: conventional advertising

and contextual advertising. Fig. 1 shows the evolution of advertising using text and media (such as image, video, and audio) as information carriers for advertising, respectively. The conventional text-based advertising, i.e., the first generation in the leftmost of Fig. 1, is characterized by delivering ads at certain positions on Web pages which are relevant to either the queries or Web page content. In this generation, paid search and display advertising are the main strategies which support a “long-tail” and “head” business model, respectively. For example, Google’s AdWords [2] and AdSense [1] are successful paid search advertising platforms, while DoubleClick [24] and Yahoo! [112] have predominantly focused on the latter. From the perspective of research, the rich research in the first generation has proceeded along three dimensions from the perspective of what the ads are matched against: 1) keyword-targeted advertising (also called “paid search advertising” or “sponsored search”) in which the ads are matched against the originating queries [49], [75], [106], 2) content-targeted advertising (also called “contextual







	Text as carrier	Media as carrier
Conventional advertising	<ul style="list-style-type: none"> ○ Problem: embed ads at fixed positions ○ Solution: <ul style="list-style-type: none"> ○ Textual relevance matching ○ Preserved ad block ○ Examples: <ul style="list-style-type: none"> ○ DoubleClick ○ AdSense ○ AdWords 	<ul style="list-style-type: none"> ○ Problem: embed ads at fixed positions ○ Solution: <ul style="list-style-type: none"> ○ Textual relevance matching ○ Post-roll, pre-roll, mid-roll ○ Examples: <ul style="list-style-type: none"> ○ AdSense ○ Revver ○ Yahoo!  
Contextual advertising	<ul style="list-style-type: none"> ○ Problem: embed ads within text ○ Solution: <ul style="list-style-type: none"> ○ Textual relevance matching ○ Advertising keyword detection ○ Examples: <ul style="list-style-type: none"> ○ Vibrant 	<ul style="list-style-type: none"> ○ Problem: embed ads within media ○ Solution: <ul style="list-style-type: none"> ○ Multimodal relevance matching ○ Nonintrusive ad position detection ○ Examples: <ul style="list-style-type: none"> ○ MediaSense  

Fig. 1. Trend of Internet advertising in the text and media domain. The online advertising can be summarized as two generations in different information carries of advertising: the first generation is conventional advertising which embeds relevant ads at fixed positions, while the second is contextual advertising embeds ads at automatically detected positions within page and media.

advertising”)² in which the ads are associated with the Web page content rather than the keywords [4], [11], [54], [83], [89], and (3) user-targeted advertising (also called “audience intelligence”) in which the ads are driven based on user profile and demography [37], comments [6], or behaviour [14], [16], [22], [90]. The advertisements in the first generation are typically embedded at certain preserved blocks in the Web pages. While conventional advertising primarily embeds ads around the content, the second generation of text advertising—contextual advertising, aims to deliver relevant ads inside the content. For example, Vibrant Media [100] associates relevant ads with certain keywords or paragraphs within a Web page and embeds these ads in-text.

Using text-based advertising as reference, we can figure out online advertising using media as carrier as two generations, including conventional and contextual advertising [40], shown in the rightmost of Fig. 1. Similar to text, the first generation of multimedia advertising directly applies text-based advertising approaches to media and embeds relevant ads at certain preserved positions on the Web pages. For example, Yahoo! [112] and BritePic [10] provide relevant ads around images. In video domain, Revver [88] and Youtube [118] which subscribe advertising service from Google’s AdSense [1] employ pre-roll or post-roll advertising (i.e., embed ads or related videos at the very beginning or end of videos), or overlay the textual ads on certain video frames (e.g., on the bottom fifth of videos 15 seconds in).

It is observed that the first generation of multimedia advertising primarily uses text rather than visual content to match relevant ads. In other words, multimedia advertising has been treated as general text advertising without considering the potential advantages which could be brought by media contents. There are very few systems in this generation to automatically monetize the opportunities brought by individual images and videos. As a result, the ads are only generally relevant to the entire Web page containing images or videos rather than specific to the images or videos it contains. Moreover, the ads are embedded at a predefined position in a Web page adjacent to the image or video, which normally destroys the visually appealing appearance and structure of the original Web page. It could not grab and monetize users’ attention aroused by these compelling contents.

It has proved not suitable to treat multimedia advertising as general text advertising. The following distinctions between media (i.e., image and video) and text advertising motivate a new advertising generation dedicated to media.

- Beyond the traditional media of Web pages, *images and videos can be powerful and effective carriers of*

²Please note here, “contextual relevance” mainly indicates that the relevance is derived from the entire web content. In a broader view, contextual relevance includes not only the relevance, but also the position where the advertisements are inserted in a Web page.

online advertising. Compared with text, image and video have some unique advantages which consequently make them become the most pervasive media formats on the Internet: they are more attractive than plain text, and they have been found to be more salient than text, thus they can grab users’ attention instantly [29]; they carry more information that can be comprehended more quickly, just like an old saying, “a picture is worth thousands of words.” Media like image and video are now used almost as much as text in Web pages and have become powerful information carriers for online advertising. There is a new advertising model using media as the carriers for advertising, in which ads can leave a much deeper impression due to the salience of visual signal in human perception.

- The ads are expected to be *locally relevant to media content and the surrounding text*, rather than globally relevant to the entire Web page. Compelling media content naturally would become the region of interest (ROI) in a Web page. The most effective way to advertise would be putting ads information precisely relevant to the media content, hoping audience who are interested in this image or video would have similar interests to the relevant product or service advertised in it. It is likely that the text in a Web page is either too much (e.g., using whole page text), or too few and/or too noisy (e.g., image and video sharing sites), to accurately describe an embedded image or video. On one hand, ads picked only based on the whole page content may not be contextually relevant enough to the image or video in that page. On the other, conventional ad-networks like AdSense [1] and Adwords [2] cannot work well for the very few or noisy textual contexts. Therefore, it is reasonable to assume that the media content and its surrounding text should have much more contributions to the relevance matching than the whole Web page.
- The ads are expected to be *dynamically embedded at the appropriate positions within each individual image or video (i.e., in-media)* rather than at a predefined position in the Web page. In conventional advertising, publishers have to reserve certain predefined blocks (please refer to Fig. 1) in the Web page for advertising—being banners or other forms of ads. Such advertising strategy has proved intrusive to Internet users [74], as the ad blocks have significantly broken the page structure and visual appearance, as well as they are unattractive or boring to users. Now that users’ attention is on the media, by embedding the ads within an image or video, the ads will in turn get more attention. Meanwhile, the publisher no longer needs to



Fig. 2. Screenshots of ImageSense, VideoSense, and GameSense. The highlighted areas in (a) and (b) correspond to the ads, while a missing block in (c) corresponds to the ad position. The ads are inserted into the non-salient spatial within images or temporal positions in video streams via visual saliency analysis. The ads are relevant to both the visual content and Web page rather than only relevant to the entire Web page. (a) ImageSense [78]; (b) VideoSense [79], [80]; (c) GameSense [59], [60].

worry about the reserved blocks. By putting ads only in the non-salient portions in an image or video, it reduces the intrusiveness of display ads and the user experience will be improved at the same time.

Motivated from the above observations, we go one step further from the first generation of multimedia advertising and propose in this paper the second generation which supports contextual multimedia advertising by associating the most relevant ads to an online medium (image or video) and seamlessly embedding the ads at the most appropriate positions within this medium. As show in the rightmost of Fig. 1, the ads are selected according to multimodal relevance, i.e., the ads are to be globally relevant to the Web page containing images or videos, as well as locally relevant to the content and surrounding text of each suitable medium. Meanwhile, the ads are embedded at the most non-salient positions within the medium. By leveraging computer vision and multimedia retrieval techniques, we are on the positive side to better solve two challenges in the Internet multimedia advertising, i.e., ad *relevance* and ad *position*. We demonstrate MediaSense which includes ImageSense [78] and VideoSense [79], [80] as two exemplary applications in the new generation, dedicated to image and video, respectively. We also envision that the next trend of multimedia advertising would be game-like advertising which would be more impressionative and thus can promote the advertisements in an interactive way. We show GameSense [59], [60] as an example of the next advertising platform. Fig. 2 shows the screenshots of ImageSense, VideoSense, and GameSense. It is also worth noticing that many metrics have been adopted to evaluate the performance of a multimedia advertising system. We review the performance evaluation in different domains.

The rest of the paper is organized as follows. Section II provides a system overview of contextual multimedia

advertising, as well as the key problems. Sections III–V address how we can leverage computer vision and multimedia retrieval techniques to solve these problems in details. Section VI shows the implementations of ImageSense, VideoSense, and GameSense. Section VII discusses how to evaluate the performance of multimedia advertising systems. Section VIII concludes this paper and outlooks future challenges.

II. SYSTEM AND KEY PROBLEMS

A. Terminology

To clearly present the system framework of the two generations of multimedia advertising, we will adopt a standard vocabulary to describe many of the common aspects and terms across each of the exemplary systems.

- **Advertisement (ad):** Advertisement is a public notice or announcement for calling something to the attention of the public, especially by paid announcements. In multimedia advertising, advertisements take a variety of forms, including text banner, image, video (i.e., traditional TV commercial), animation, or a combination of forms. Advertising is a form of communication that typically attempts to persuade potential customers to purchase or to consume more of a particular brand of product or service. In this paper, we mainly discuss two types of ads, i.e., image and video ads.
- **Image ad:** An image advertisement is a static image or banner provided by advertisers that will be inserted into or associated with a source image or video. An image ad could be a product logo [59], [60], [66], [69], [78] or a banner composed of a product logo, product name, description, and link [31], [76], [101], [118].

- **Video ad:** A video ad is a video clip about a product. Although video ads will be associated with a video in MediaSense [25], [80], they may be in different forms (or a combination of forms), including typical commercials in TV programs [19], [25], [39], [67], as well as a clip composed by text, animations, or images.
- **Source media:** Source media are most often produced or owned by content providers (or partially by publishers who help to distribute the contents), which may be images, videos, or audio clips, captured and provided by professional photographers, videographers, or grassroots. The advertisements will be embedded at certain positions around, within, or overlay the source media.
- **Ad insertion point:** A point/position where one or more advertisements will be associated. Ad insertion point could be a spatial region around the medium in a Web page, a region on an image, a spot on the timeline of a video, a spatiotemporal patch in a video, or even a position out of the Web page. For example, the highlight rectangle in Fig. 1(a), the yellow spots on the timeline in Fig. 1(b), and the center missing block in Fig. 1(c) are ad insertion points.
- **Contextual advertising:** Contextual advertising refers to the placement of commercial advertisements within the content of a generic Web page based on similarity between the content of the target page and the ad description provided by the advertiser [11], [83]. If the advertisements will be

associated with media, then this type of advertising is contextual multimedia advertising.

- **Multimodal relevance:** A modality is defined as any source of information about media contents that can be leveraged algorithmically for analysis in [52]. In multimedia advertising applications, the modality can be decomposed into various modalities which measure various low-level aspects of visual data (such as the color and textures in an image, the motions in a video sequence, as well as the tempos in an audio stream), some mid-level visual concept or object categories (such as people, location, and objects), as well as high-level textual descriptions associated with the data (such as user-provided tags on an image or video, transcripts associated with a video stream, automatically recognized captions on a video frame). The relevance can be measured by the similarity between two media files in terms of certain types of modalities. For example, there are textual, visual, and aural relevance. Accordingly, the multimodal relevance is a combination of the results from various modalities between two media.

B. General Framework

Fig. 3 shows the general framework of multimedia advertising. It also summarizes the distinctive between the two generations of multimedia advertising, as we mentioned in Section I. Given an online medium which could be an image within a Web page, a collection of image search results, a video sequence, or even an audio stream, a list of candidate ads are selected from an ad inventory and

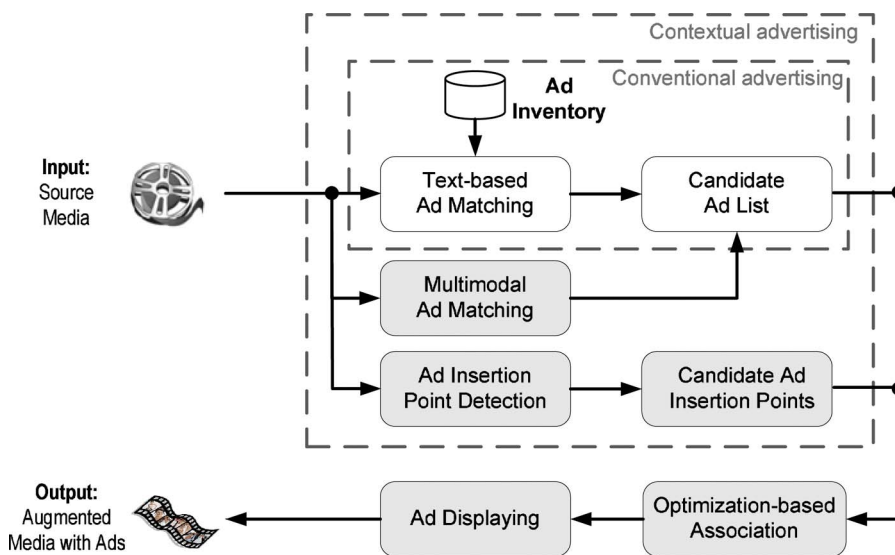


Fig. 3. A general framework of multimedia advertising. Conventional advertising only focuses on text-based ad relevance matching, while contextual advertising like MediaSense [40] considers not only textual relevance but also multimodal (textual, visual, and aural) relevance, as well as automatic detection of appropriate ad insertion points within media.

ranked according to the relevance between the source medium and the ads. The multimodal relevance can be derived from textual information, such as user-provided tags, video transcripts, automatically recognized captions, and also can be derived from the low-level similarity in terms of color and textures in images, camera or object motions in video sequences, or tempo and beat in audio streams, as well as semantic level in terms of automatically detected visual concepts and object categories [27], [77], [84], [95]. Meanwhile, a set of candidate ad insertion points are automatically detected on the basis of spatio-temporal visual saliency analysis [15], [64], [69], [73], [78]–[80], [103]. Intuitively, the ads would be inserted into the most non-salient positions within the media contents so that the ads would be nonintrusive to the viewers. Moreover, the ads associated with the media would be the most relevant to the contents. By minimizing contextual intrusiveness and maximizing contextual relevance simultaneously, the effectiveness of advertising can be maximized [74]. Given a list of candidate ads and ad insertion points, an optimization-based association module will associate each candidate ad with the best insertion point.

We can see from Fig. 3 that conventional multimedia advertising only focuses on ad relevance matching, referred to as “text-based ad matching” and “candidate ad list” modules, while the new generation of advertising not only considers “multimodal ad matching” for ad selection but also investigates the problem of “ad insertion point detection.” Finally, given a set of candidate ads and ad insertion points, the “optimization-based ad delivery” module will associate the most relevant ads with the most appropriate insertion points by maximizing the overall contextual relevance while minimizing the overall intrusiveness.

C. Key Problems

In general, there are four key problems in an effective contextual multimedia advertising system: *contextual relevance*, *contextual intrusiveness*, *insertion optimization*, and *rich displaying*.

- **Contextual relevance**—Which ads should be selected for a given image or video? Since relevance increases advertising revenue [57], [74], contextual multimedia advertising performs multimodal relevance matching by considering both global textual relevance from the entire Web page and local relevance from textual information associated with

the media content, as well as low-level visual and high-level semantic similarity between the ads and ad insertion points.

- **Contextual intrusiveness**—Where should the selected ads be inserted so that the contextual intrusiveness will be minimized? Ad position will certainly affect user experience when an image or a video is viewed [74]. In contextual multimedia advertising, the selected ads are to be inserted into the most non-intrusive positions within the media.
- **Insertion optimization**—Given a ranked list of candidate ads and ad insertion points, how to associate each ad with the ad best insertion point? The objective is to maximize the effectiveness of advertising by simultaneously minimizing the contextual intrusiveness to viewers and maximizing the contextual relevance between the ads and media.
- **Rich displaying**—How the selected ads are displayed or rendered? The rich displaying includes the duration of each ad, the way the ad is rendered, the support of interaction between the ad and users, the rich information associated with ads. An effective displaying will make the advertising not have an intrusive experience to users.

In this paper, we mainly focus on the first three problems while leave the fourth an open issue. The comparisons between conventional and the proposed contextual multimedia advertising in terms of contextual relevance and contextual intrusiveness are listed in Table 1.

III. CONTEXTUAL RELEVANCE

One of the fundamental problems in contextual advertising is “relevance” which in studies detracts from user experience and increases the probability of reaction [57], [74]. By contextual relevance, we refer to the fact that ads are expected to be relevant both to the entire source media and the local ad insertion points within the media. The contextual relevance for each pair of ad and ad insertion point is a multimodal relevance consisting of textual, visual, conceptual, and user relevance. In this section, we will review the relevance from different modalities.

A. Textual Relevance

The major effort in advertising has focused on text domain. There exists rich research in the literature on textual relevance that can be leveraged or applied to multimedia

Table 1 Comparisons Between Conventional and Contextual Multimedia Advertising, in Terms of Contextual Relevance and Contextual Intrusiveness

Advertising generation	Contextual relevance	Contextual intrusiveness	Exemplary systems
Conventional advertising	Textual relevance, solely relying on Web page	Predefined ad positions by publishers	AdSense [1], Yahoo! [112], BritePic [10], TV commercial
Contextual Advertising	Multimodal relevance (textual, visual, and aural relevance)	Automatically detected ad positions within media	MediaSense [40], VideoSense [80] ImageSense [78], GameSense [76]

advertising. The literature review on text relevance in this paper will focus on two key issues: 1) ad keyword selection, i.e., how to pick suitable keywords, Web pages, or images for advertising so that the relevance can be improved, and 2) ad relevance matching, i.e., how to select relevant ads according to a set of selected keywords or a Web page.

Typical advertising systems analyze a Web page or query to find prominent keywords or categories, and then match these keywords or categories against the words for which advertisers bid. If there is a match, the corresponding ads will be displayed to the user through the web page. Yih *et al.* has studied a learning-based approach to automatically extracting appropriate keywords from Web pages for advertisement targeting [117]. Instead of dealing with general Web pages, Li *et al.* propose a sequential pattern mining-based method to discover keywords from a specific broadcasting content domain [58]. In addition to Web pages, queries also play an important role in paid search advertising. In [94], the queries are classified into an intermediate taxonomy so that the selected ads are more targeted to the query. The works in [56], [59], [60], [78]–[80] present the methods for detecting potential suitable images or videos in a Web page for advertising, by analyzing the structure of Web page and the visual appearance of images or videos.

As we have mentioned in Section I, research on ad relevance has proceeded along three dimensions from the perspective of what the ads are matched against: 1) keyword-targeted advertising (“paid search advertising” or “sponsored search”), 2) content-targeted advertising (“contextual advertising”), and 3) user-targeted advertising (“audience intelligence”). Although the paid search market develops quicker than contextual advertising market, and most textual ads are still characterized by “bid phrases,” there has been a drift to contextual advertising as it supports a long-tail business model [57]. For example, a recent work examines a number of strategies to match ads to Web pages based on extracted keywords [89]. A follow-up work applies Genetic Programming (GP) to learn functions that select the most appropriate ads, given the contents of a Web page [54]. To alleviate the problem of exact keyword match in conventional advertising, Broder *et al.* propose to integrate semantic phrase into traditional keyword matching [11]. Specifically, both the pages and ads are classified into a common large taxonomy, which is then used to narrow down the search of keywords to concepts. Most recently, Hu *et al.* propose to predict user demographics from browsing behavior [37]. The intuition is that while user demographics are not easy to obtain, browsing behaviors indicate a user’s interest and profile.

When applying textual relevance to visual domain, in addition to the techniques discussed above in text domain, the characteristics of media should be taken into account from the following perspectives: 1) The entire texts in the Web page are too noisy and broad to describe the media

embedded in the page, while the surrounding texts which are spatially close to the media can better describe the contents and lead to better ad relevance. 2) The surrounding texts associated with an image or video are sometimes too few for selecting relevant ads. The hidden texts (e.g., expanded words, visual concepts, object categories, or events) which are automatically recognized from visual signals can more precisely describe the media contents. Using the surrounding and hidden texts can yield better textual relevance.

Given a Web page containing images or videos, it is desirable to first segment it into several blocks with coherent topic, detect the blocks with suitable images or videos for advertising, and extract the semantic structure such as the surrounding texts from these blocks. The Vision-based Page Segmentation (VIPS) algorithm [12], [13] is adopted to extract the surrounding texts associated with a medium in [59], [60], [78], [79]. The VIPS algorithm makes full use of page layout structure. It first extracts all the suitable blocks from the Document Object Model (DOM) tree in html, and then finds the separators between these blocks. Based on these separators, a Web page can be represented by a semantic tree in which each leaf node corresponds to a block. In this way, contents with different topics are distinguished as separate blocks in a Web page. Fig. 4(a) and (b) illustrate the vision-based structured of a sample page. It is observed that this page has two main blocks and the block named “VB-1-2-1-1” is detected as the video block. Specifically, after obtaining all the blocks via VIPS, the images or videos which are suitable for advertising in the Web page are elaborately selected. Intuitively, the images or videos with poor visual qualities, or belonging to the advertisements (usually placed in certain positions of a page) or decorations (usually are too small), are first filtered out. Then, the corresponding blocks with the remaining images or videos are selected as the advertising page blocks. The surrounding texts (e.g., title and description) are used to describe each image or video.

Based on the surrounding texts, the expansion text can be obtained by leveraging query expansion based on user log [21], while the hidden texts are obtained by automatic text categorization [116] and video concept detection [77], [87]. Specifically, we use text categorization based on Support Vector Machine (SVM) [116] to automatically classify a textual document into a set of predefined category hierarchy which consists of more than 1000 categories. The concept texts are selected by using the top concepts with the highest confidence scores in a specific ontology like LSCOM [77]. The hidden texts can inform the direct text as the surrounding text related to video is usually not quality-controlled. Fig. 4(c) shows these two types of texts with the corresponding probabilities derived from (a) and (b).

Given the surrounding texts, the vector space model (VSM) [7] can be adopted to measure the textual relevance

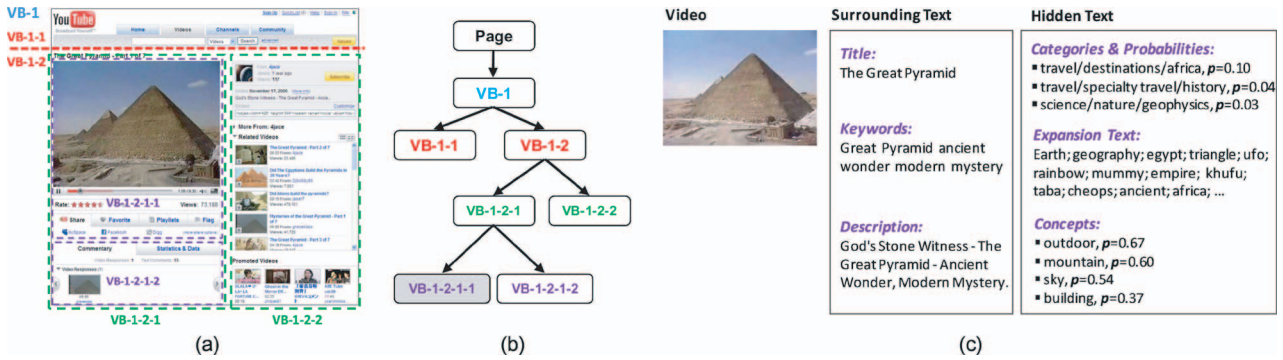


Fig. 4. Text extraction from a Web page. “VB” indicates the visual block. A source video is extracted from block “VB-1-2-1-1,” while the textual information is extracted from block “VB-1-2-2.” (a) The segmented Web page, (b) DOM tree, (c) Textual information including surrounding and hidden texts.

between a source medium and a candidate ad [78], [89]. All the texts associated with the ad is supposed to be provided or bid by advertisers. In the VSM, each textual document is represented as weighted vectors in an n -dimensional space. The weight can be computed using the product of the term frequency (tf) and inverted document frequency (idf), reflecting 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. The ranking of the ads with regard to the document is computed by the cosine similarity between their surrounding texts [7].

B. Visual Relevance

Different from textual relevance which is globally defined on the entire image or video, the visual relevance is regarded local as it is usually defined on the basis of a local ad insertion point. The visual relevance measures the local similarity between an ad insertion point and ad, in terms of visual appearance. Intuitively, the inserted ads are expected to have similar appearance with the local insertion points, so that viewing experience may not be degraded as much as possible [15], [31], [69], [76], [78], [91]. Likewise, in a video advertising scenario, the ads are expected to have similar visual-aural style with the video content around the insertion point [79], [80], [96]. Such visual relevance can be adjusted according to different advertising strategies. For example, if the ad-network cares the advertisers more than the viewers, it can make the visual relevance as low as possible, so that the ads can attract viewers’ attention as much as possible. However, which strategy for visual relevance would be the best for contextual multimedia advertising still remains an open issue.

In contextual in-image advertising [59], [60], [65], [78], the product logos are embedded in non-salient image blocks. The ads are assumed to have similar appearance with the neighboring blocks around the insertion posi-

tions, so that the users may perceive the ads as a natural part of original image. To measure the visual relevance between the ad and ad insertion position, the L_1 distance in a HSV color space between the ad block and the neighboring image blocks is adopted. This distance has been widely adopted in visual retrieval and search systems [23], [77]. Instead of global feature, local features such as the scale-invariant feature transform (SIFT) descriptors [71] can be used for finding the visually exact match between a video frame and a product logo [66].

In contextual in-video advertising (or in-stream) [31], [76], [79], [80], the visual relevance is measured by a set of visually perceptual features such as motion and color, as well as high-level concept text [as shown in Fig. 4(c)]. The visual relevance is usually combined with the aural relevance which is derived from audio track. Specifically, the motion intensity is used to characterize motion which is computed by averaging the frame differences within a shot [38]. The color is represented by a 16-dimensional dominant color histogram in HSV space [79]. The audio tempo is used to describe the rhythm and energy in audio track [38]. These features have proved to be effective to describe video content in many existing multimedia applications [38], [114]. More sophisticated low-level features related to visual-aural relevance can be also applied here. Actually, the authors suggest various ways for using local visual relevance in [31], [76], [79], [80]. For example, we can use the “positive” local visual relevance to keep high similarity between the video and ad content, or use it in a “negative” way to gain more attention from viewers because of the high contrast. Typically, the “positive” way is chosen 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 does not degrade 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. The visual relevance between the insertion point and video ad is the linear fusion of the visual similarities between the ad and the source video shots around the insertion point (i.e., before and after the shot boundary). Please note that the insertion point in the in-video advertising system is a shot boundary or a story break. Different from the in-video advertising in [76], [79], [80], the vADeo system attempts to solve visual relevance by finding the same person in a video segment via face detection and recognition [96].

C. Conceptual Relevance

Given the hidden texts which can be obtained by the methods described in Section III-A, a probabilistic model [79], [80], [114] is adopted to measure the conceptual relevance between a source medium and a candidate ad. The probabilistic model represents the categories or concepts in a hierarchical tree, in which each node corresponds to a category or concept. The relevance between two documents is the sum of the relevance between all possible pairs of categories or concepts. The relevance between two nodes in this tree is measured by their probabilistic distance.

Another way to compute conceptual relevance is representing the hidden text associated with a medium by a normalized vector, with each element corresponding to the probability of certain category or concept [34], [36]. Then, the relevance between the ad and medium can be measured by the L_1 distance or other distance metrics between the text of ads (provided by advertisers) and hidden texts associated with the media. Most recently, Wu *et al.* propose to measure the textual relevance between two concept words via Flickr distance, which is built upon the visual language models from a collection of images searched by the concepts [109]. Flickr distance is a new textual relevance designed from the perspective of visual domain.

D. User Relevance

With social networks are becoming more and more pervasive, delivering user-targeted ads for multimedia advertising has become an emerging issue. Based on user relevance, the ads are expected to be relevant to user profile, interest, location, click-through, historical behavior (e.g., travel traces), and so on. For example, the personalized ad delivery in Interactive Digital Television (IDTV) has been a potentially hot application [51], [55], [97], [99]. Such advertising refers to the delivery of advertisements tailored to the individual viewers' profiles on the basis of knowledge about their preferences [55], current and past contextual information [51], [97], or sponsors' preference [99]. In these systems, the users are assumed to be grouped into a set of predefined interest groups or provide their profile in advance, and then the ads falling into the users' interest will be delivered based on text matching and classification techniques described in Section III-A. However, most of these systems do not study relevance in terms of visual content.

IV. CONTEXTUAL INTRUSIVENESS

Contextual relevance deals with the selection of relevant ads according to a given image or video, while contextual intrusiveness answers the question where the selected ads are embedded. Detecting ad insertion point within a medium is on the contrary to traditional commercial detection in TV programs [25], [39], [67]. Similar to the visual relevance described in Section III-B, there are various strategies for finding ad insertion points via contextual intrusiveness. One way is to find the most interesting or highlighting segments in a video or the most salient region in an image as ad insertion point, so that the ad impression will be maximized [57], [74]. The other is contrary, i.e., to seek the most non-salient parts as ad insertion points, so that users may not feel intrusive when browsing the augmented media with ads. Finding suitable insertion points is actually a kind of trade-off between ad impression and viewing experience, which deserves a deep study on the effectiveness of advertising. Technically, contextual intrusiveness is the key to automatically detect ad insertion point in a given medium. Traditional advertising adopts the preserved blocks in a Web page as ad positions, while in the new generation of multimedia advertising, computer vision techniques such as visual saliency detection and video content analysis can benefit the automatic detection of such positions. We next describe how to detect ad insertion points in the image and video domain via contextual intrusiveness analysis, respectively.

A. Contextual Intrusiveness in Image Domain

In contextual image advertising, the relevant ads are embedded at certain spatial positions within an image. The image ad-network finds the non-intrusive ad positions within an image and selects the ads whose product logos are visually similar or have similar visual style to the image, so as to minimize intrusiveness and improve user experience. Specifically, in [78], the candidate ad insertion positions (usually image blocks) are detected based on the combination of image saliency map, face detection, and text detection, while visual similarity is measured on the basis of HSV color feature.

In ImageSense [78], given an Image I which is represented by a set of blocks, a saliency map S representing the importance of the visual content is extracted by investigating the effects of contrast in human perception [73]. Fig. 5(d) shows an example of the saliency map of an image. The brighter the pixel in the saliency map, the more salient or important it is. However, saliency map predominantly focuses on modeling visual attention while neglects face and embedded text (i.e., caption) in images. In fact, face and text usually present informative content in an image. A product logo should not overlay the areas with face or text. Based on this assumption, we can perform face [61] and text detection [17], [41] for each image and obtain the face and text areas. Then, the saliency map S is overlaid on the detected face and text areas by a max operation. In this way,

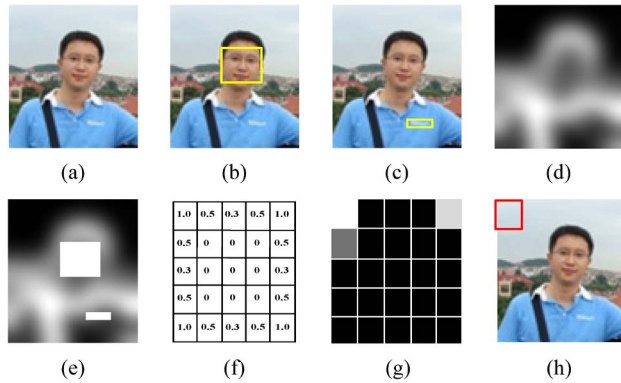


Fig. 5. An example of detecting candidate ad positions in an image in [78]. (a) I : original image, (b) face detection result, (c) text detection result, (d) S : image saliency map, (e) C : combined saliency map, (f) W : weight map, (g) final saliency map combining (e) and (f), (h) the least nonintrusive candidate ad position. The brighter the pixel in the saliency map (d) and (e), the more important or salient it is; while the brighter each block in (g), the more likely it is a candidate ad position. The highlighted rectangle in (h) is obtained from (g). We use 5×5 for illustration in this figure.

a combined saliency map C is obtained in which the value of each pixel indicates the overall salience for ad insertion. Fig. 5(b)–(d) show face, text, and saliency maps of the original image in (a), while (e) shows the combined saliency map.

Intuitively, the ads should be embedded at the most non-salient regions in the combined saliency map C , so that the informative content of the image will not be occluded and the users may not feel intrusive. Meanwhile, the image block set B is obtained by partitioning image I into grids. Each grid corresponds to a block b_i and also a candidate ad insertion point. For each block, a combined saliency energy c_i ($0 \leq c_i \leq 1$) is computed by averaging all the normalized saliency energies of the pixels within b_i . As the combined saliency map C does not consider the spatial importance for ad insertion, a weight map $W = \{w_i\}$ is designed to weight the energy c_i , so that the ads will be inserted into the corners or sides rather than center blocks. Fig. 5(f) and (g) show an example of the weight map and $w_i \times (1 - c_i)$, respectively. Therefore, $w_i \times (1 - c_i)$ indicates the content non-intrusiveness of block b_i for embedding an ad. As a result, the top-left block is selected as the ad insertion point in Fig. 5(h).

Rather than overlay the ads at non-salient regions within images, Li *et al.* propose to overlay the ads on the center of the image before the full-resolution images are downloaded [65]. Fig. 6 shows the basic idea of delivering ads on the image. From this perspective, the contextual intrusiveness depends on the time when images appear on the Web pages.

B. Contextual Intrusiveness in Video Domain

The ads in contextual video advertising can be inserted in a preserved page block around the video

[66], as an overlay video on certain frames [31], [76], in the story or scene breaks in a video [79], [80], or even into a spatiotemporal portion in a video [15], [62], [64], [69], [103].

The detection of in-stream ad insertion point is based on contextual intrusiveness [79], [80]. A pre-processing step is assumed to parse the source video into shots and represent each shot by a key-frame using the color-based method [120]. Each shot boundary is naturally a candidate ad insertion point. Li *et al.* investigate eight factors that affect consumers’ perceptions of the intrusiveness of ads in traditional TV programs [57]. Two computable measurements based on these eight factors, i.e., content discontinuity and attractiveness are excerpted from [57]. The content discontinuity measures content “dissimilarity” between two video shots, while content attractiveness measures the “importance” or “interestingness” of the content within a shot. The higher the discontinuity, the more likely the corresponding insertion point is a boundary of two stories. In fact, 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, “high discontinuity plus high attractiveness” may be a tradeoff between viewers and advertisers. Then, the detection of ad insertion points can be formulated as ranking the shot boundaries based on different combinations of content discontinuity and attractiveness.

In [79], [80], the authors seek a soft measure of a shot boundary to be an ad insertion point. As shown in Fig. 7, a degree of discontinuity is assigned to each insertion point. The higher the discontinuity, the more likely the corresponding insertion point is a boundary of two stories. An improved BFMM (so-called “iBFMM”) is proposed to deal with the interlaced repetitive pattern problem in BFMM and assign each boundary a soft discontinuity value [122]. 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.

In general, it is difficult to evaluate how a video clip attracts viewers’ attention since “attractiveness” or

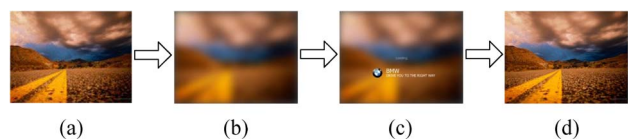


Fig. 6. The ad is overlaid on the image before the full-resolution image is downloaded in [65]. The image in (a) is the original full-resolution image, while the images and ads in (b)–(d) appear according to time.



Fig. 7. Candidate ad insertion point detection in [79], [80]. Suppose there are N_s shot or scene boundaries in a video sequence, then the number of candidate points is $N_s + 1$ (including the beginning and the end of the video), each boundary indicating one candidate insertion point.

“attention” is a neurobiological concept. Alternatively, a user attention model is proposed by Ma *et al.* to estimate human attention by integrating a set of visual, auditory, and linguistic elements related to attractiveness [72]. Another approach to exploring the attention in video sequence is to average static image attention over a segment of frames [73]. In [79], [80], an attention value is computed for each shot by the user attention model in [72]. The content attractiveness of an insertion point is highly related to the neighboring shots on both sides of this point. Therefore, the attractiveness of ad insertion point is computed by weighted averaging the attention values of its neighboring shots.

Different from video advertising for general video contents in [79], [80], to make video contents more enriching, researchers have attempted to spatially replace a specific region with product advertisement in sports videos [15], [64], [69], [103]. These regions could be locations with less information in baseball [64] and tennis video [15], the region above the goal-mouth in soccer video [103], or the smooth regions in the video highlights [69]. Fig. 8 shows some examples of advertisements in sports video. An online platform is also presented to measure the quality of product placement [45]. The domain-specific approaches in these applications, such as the detection of line and less-information-region [64], [103], are not practical in a general case, especially in online videos. Li *et al.* propose to find the most non-salient space-time portions in the video [62]. They formulate the problem as a Maximum a Posterior (MAP) problem which maximizes the desired properties related to less intrusive viewing experience, i.e., *informativeness, consistency, visual naturalness, and stability.*

V. INSERTION OPTIMIZATION

Given a list of ads ranked according to their contextual relevance and a list of ad insertion points ranked according to their contextual intrusiveness, how to associate each ad with one of insertion points so that the effectiveness of contextual advertising could be maximized? As we have mentioned in Section II-C that an effective advertising system is able to maximize contextual relevance while minimizing contextual intrusiveness at the same time. This problem is formulated as a non-linear 0–1 integer programming problem (NIP) in [78], [80], with each of the above rules as a constraint.

For example, without of loss of generality, the task of insertion optimization can be defined as the association of ads with insertion points in an online medium which might be an image or a video. Suppose we have an ad database \mathcal{A} which contains N_a ads, represented by $\mathcal{A} = \{a_i\}_{i=1}^{N_a}$, and we also have a set of candidate ad insertion points \mathcal{P} which contains N_p points, represented by $\mathcal{P} = \{p_j\}_{j=1}^{N_p}$. The contextual relevance $R(a_i, p_j)$ between each ad a_i and point p_j can be obtained by the approaches in Section III (i.e., the linear combination of textual, visual, conceptual, and user relevance), while the contextual intrusiveness $I(a_i, p_j)$ can be obtained in Section IV. The objective of contextual advertising is to maximize the overall contextual relevance $R(\mathcal{A}, \mathcal{P})$ while minimizing the overall contextual intrusiveness $I(\mathcal{A}, \mathcal{P})$. The following design variables can be introduced for problem formulation, i.e., $\mathbf{x} \in \mathbb{R}^{N_a}$, $\mathbf{y} \in \mathbb{R}^{N_p}$, $\mathbf{x} = [x_1, \dots, x_{N_a}]^T$, $x_i \in \{0, 1\}$, and $\mathbf{y} = [y_1, \dots, y_{N_p}]^T$, $y_j \in \{0, 1\}$, where x_i and y_j indicate whether a_i and p_j are selected ($x_i = 1, y_j = 1$) in \mathcal{A} and \mathcal{P} . Given the number of

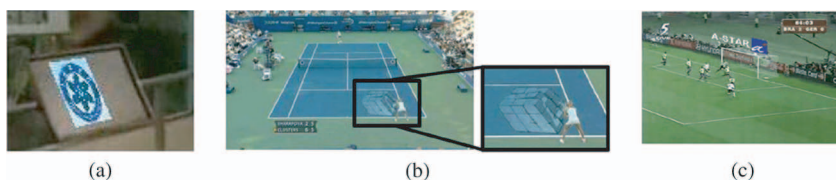


Fig. 8. Ad insertion point detection in sports videos. Virtual contents are inserted in the smooth areas by leveraging domain-specific knowledge in sports video. (a): [69], (b): [15], (c): [64], [103].

ads N to be inserted in a medium, the NIP formulation [9] is

$$\begin{aligned} \max_{(x_i, y_j)} f(\mathbf{x}, \mathbf{y}) &= \alpha \sum_{i=1}^{N_a} \sum_{j=1}^{N_p} x_i y_j R(a_i, p_j) \\ &\quad - \beta \sum_{i=1}^{N_a} \sum_{j=1}^{N_p} x_i y_j I(a_i, p_j) \\ &= \alpha \mathbf{x}^T \mathbf{R} \mathbf{y} - \beta \mathbf{x}^T \mathbf{I} \mathbf{y} \\ \text{s.t.} \quad &\sum_{i=1}^{N_a} x_i = N, \\ &\sum_{j=1}^{N_p} y_j = N, \quad x_i, y_j \in \{0, 1\} \end{aligned} \quad (1)$$

where $\mathbf{R} \in \mathbb{R}^{N_a \times N_p}$, $\mathbf{R} = [R_{ij}]$, $R_{ij} = R(a_i, p_j)$, and $\mathbf{I} \in \mathbb{R}^{N_a \times N_p}$, $\mathbf{I} = [I_{ij}]$, $I_{ij} = I(a_i, p_j)$, α and β are two weights for linear combination. Other constraints such as the uniform distribution of ad insertion points in the source video [79], [80] can be added in (1).

It is observed that there are $C_{N_a}^N C_{N_p}^N N!$ solutions in total to (1). As a result, when the number of elements in \mathcal{A} and \mathcal{P} is large, the searching space for optimization increases dramatically. However, the Genetic Algorithm (GA) [107] can be employed to find solutions approaching the global optimum. Alternatively, the above problem can be solved by a similar heuristic searching algorithm in practice [78]–[80]. In this way, the number of possible solutions can be significantly reduced to $C_{N_a}^1 C_{N_p}^1 N!$. Note that (1) is a general formulation for advertising. In fact, it can be easily extended to various advertising strategies from different perspectives. The authors in [79] have given detailed discussions on supporting diverse advertising scenarios based on this framework.

VI. EXEMPLARY SYSTEM: MEDIASENSE

We will show the implementations of recently developed exemplary application of contextual multimedia advertis-

ing, called MediaSense. MediaSense includes ImageSense [78], VideoSense [79], [80], and GameSense [59], [60] which are dedicated to online image, video, and game, respectively. Specifically, we introduce how the relevant ads are selected and how the ad insertion positions are automatically detected, as well as how the ads and these insertion points are associated in these applications.

A. ImageSense

A snapshot of ImageSense is shown in Fig. 2(a). ImageSense is able to automatically decompose a Web page into several coherent blocks, select the suitable images from these blocks for advertising, detect the nonintrusive ad insertion positions within the images, and associate the relevant advertisements (i.e., product logos) with these positions [78]. The ads are selected based on not only textual relevance but also visual relevance, so that the ads yield contextual relevance to both the text in the Web page and the image content. The ad insertion positions are detected based on image saliency, as well as face and text detection, to minimize intrusiveness to the user. An example of image tagged with “domestic” and “kitten” which is associated with the product logo of “Animal Planet” is shown in Fig. 9. We can see that the ads ranked only by textual relevance are not closely related to this image. However, by contextual relevance, we can know this image is about “cat” and “animal” as we can build specific visual models for recognizing “cat” and “animal.” Therefore, the ads related to animal can be ranked higher. Furthermore, a suitable position is selected by image saliency detection so that the visual similar ad is recommended to be embedded on the top-right corner in this image.

B. VideoSense

A snapshot of VideoSense is shown in Fig. 2(b). VideoSense is an in-video advertising system that is able to elaborately detect a set of appropriate ad insertion points (i.e., shot breaks) based on content discontinuity and attractiveness, and associate the most relevant video ads to these points, according to not only global textual relevance but also local visual-aural relevance [79], [80].



Fig. 9. A sample image with ad embedded in ImageSense [78]. (a) The source image and image with ads embedded, (b) the ranked ad list by only textual relevance and contextual relevance.

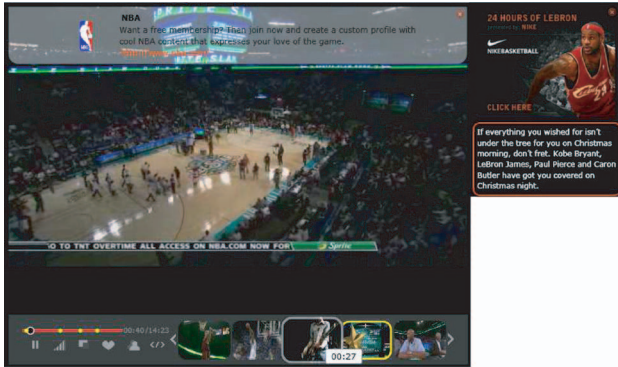


Fig. 10. An example of a video subscribing overlay ads service [31], [76]. The highlighted yellow rectangle indicates that an ad is overlaid on this video shot, while the yellow spots on the timeline indicate the overlay ads in the video stream. The relevant ads are embedded at the non-salient positions (bottom or top) on the suitable frames, while the corresponding accompany ads appears beside the video player (on the right).

In addition to in-video advertising, the overlay video advertising can be supported based on the same framework of VideoSense [31], [76]. Fig. 10 shows the example of a video subscribing intelligent overlay video advertising service. Unlike most of current ad-networks such as Youtube [118] and Revver [88] that overlay the ads at fixed positions in the videos (e.g., on the bottom fifth of videos 15 seconds in), the ads in [31], [76] are automatically overlaid on the frames based on highlight detection and visual saliency analysis. An accompany ad with more information about the product or service is displayed on the right.

C. GameSense

A snapshot of GameSense is shown in Fig. 2(c). GameSense is a game-like advertising system built upon ImageSense framework [59], [60]. Given a Web page which typically contains images, GameSense is able to select a suitable image to create online in-image game and associates relevant ads within the game. The contextually relevant ads (i.e., product logos) are embedded at appropriate positions within the online games. The image blocks in the ad positions are used as missing blocks in the “sliding puzzle” game. The ads are selected based on not only textual relevance but also visual similarity. During the game, the ads can alternate once the user has proceeded one step. The game is able to provide viewers rich experience and thus promote the embedded ads to provide more effective advertising. Furthermore, users participating the games can get incentives from the multiple-player mode. On the other hand, advertisers can achieve more ad impression during the game. The idea is similar to ESP game which tackles the problem of creating difficult meta-data via human power and computer game [3], [30]. The GameSense represents one of the first attempts towards impressionate advertising.

VII. PERFORMANCE EVALUATION

In order to obtain a fair empirical evaluation of multimedia advertising systems, it is important to use standard and representative data sets and performance metrics. However, there are no benchmark data sets in the research communities. Most experimental results were reported using different data sets. Therefore, we only review the performance metrics which are commonly used in the typical advertising systems. These metrics are categorized as follows in terms of the preliminary domains in which they are adopted.

- Online advertising. The performance of online advertising is commonly measured by ads Click-Through Rate (CTR) or the revenue from advertisers [74]. A CTR is obtained by dividing the “number of users who clicked on an ad” on a web page by the “number of times the ad was delivered” (impressions) [108]. The commonly used pricing models, such as pay-per-click (PPC), cost-per-thousand (CPT), and cost-per-mille (CPM), etc., are based on the estimation of CTR. However, CTR is usually difficult to obtain without a long-term investigation in a real advertising system.
- Information retrieval. The key problem for advertising is the relevance between advertisements and landing pages or programs (i.e., images or videos). The relevance is usually judged by human on several scales (e.g., “relevant,” “somewhat relevant,” and “irrelevant”). Based on these judges, the relevance can be measured by the Precision-Recall (P-R) curve, Average Precision (AP), Normalized Discounted Cumulative Gain (NDCG), and so on [47], [102]. For example, *recall* is defined as fraction of relevant advertisements (in the whole data set) which has been retrieved for a given document or program, while *precision* is the fraction of the retrieved advertisements (in the returned subset) which is relevant [7]. Then, a P-R curve can be generated by plotting the curve of precision versus recall. The AP corresponds to the area under a non-interpolated P-R curve, which is widely adopted in the TRECVID [98]. NDCG is a commonly adopted metric for evaluating a search engine’s performance. Given a query q , the NDCG score at the depth d in the ranked documents is defined by

$$NDCG@d = Z_d \sum_{j=1}^d \frac{2^{r^j} - 1}{\log(1 + j)} \quad (2)$$

where r^j is the rating of the j -th document, Z_d is a normalization constant and is chosen so that a perfect ranking’s $NDCG@d$ value is 1.

- Multimedia advertising. In addition to the relevance, user experience is an important problem in this domain. The evaluation of user experience often depends on a series of subjective tests or user studies [91]. However, there are no commonly used metrics for the subjective evaluations. Different researchers used different kinds of evaluations and metrics. For example, the authors in [76], [78], [80] let subjects judge “satisfaction on ad location” and “overall satisfaction on advertising effect” on a 1 to 5 scale. The authors in [69], [91], [103], [111] designed a survey to collect user feedback (e.g., “positive” or “negative”) after they viewed the advertisements.
- Human-computer interaction. Researchers in this domain employed eye-movement tracking as the measurement of the effectiveness of advertising [46], [86]. The metrics include fixation (e.g., number of fixations, fixations per area of interest, fixation duration), saccade (e.g., number of saccades, saccade amplitude), scanpath (e.g., scanpath duration, scanpath length), and so on [86].

VIII. CONCLUSIONS AND FUTURE CHALLENGES

Image and video have become the most pervasive formats and compelling contents on the Internet. With the right strategy and the right technology for advertising, we can start leveraging the power of visual contents to build value in any Web site, page, image, video, and even game. In this paper, we have outlined the trend of multimedia advertising as two generations and discussed the key issues in contextual multimedia advertising, including contextual relevance, contextual intrusiveness, and insertion optimization. We conclude that an effective advertising for Internet multimedia should maximize the content relevance while minimizing contextual intrusiveness at the same time. We have observed that the key techniques for advertising come from text domain, that is, textual relevance is the basis of selecting relevant advertisements to a given image or video. However, we have witnessed that the visual distinctive characteristics of multimedia have inspired the employment of the techniques in visual domain. The advanced techniques in computer vision (e.g., visual saliency analysis, face detection, text detection, object categorization, and so on) and multimedia retrieval (e.g., video structuring, video concept detection, highlight detection, and so on) have proved effective to improve contextual relevance and reduce contextual intrusiveness. We further described an exemplary system for contextual multimedia advertising, called MediaSense, which consists of ImageSense, VideoSense, and GameSense, dedicated for monetizing the compelling contents of online image, video, and game, respectively.

Multimedia advertising is a vibrant area of research. There are a lot of emerging topics deserving deep investigation and research. We can summarize the future challenges as follows.

How can we improve contextual relevance? As we have mentioned in Section III, contextual relevance comes from textual, visual, conceptual, and user relevance. Each type of relevance is a challenging problem in the corresponding research community, ranging from information retrieval, computer vision, human computer interface, and so on. From the perspective of vision, there exist some possible investigations towards better ad relevance.

- 1) Visual categorization in image and video domains can be used for better and more precisely describing visual contents and in turn improve contextual relevance. For example, the advanced techniques for objective recognition [27], [104] and video concept detection [34], [36], [87] can benefit conceptual relevance. While fully automatic categorization or annotation still achieved limited success [35], [42], Internet-based annotation which is characterized by collecting crowd-sourcing knowledge, as well as combining human and computer for active tagging (i.e., ontology free annotation), is a promising direction for improving ad relevance. For example, a recent work presents an active tagging approach to combine the power of human and computer for recommending tags to images [115]. The research on social tagging has proceeded along another dimension which aims to differentiate the tags with various degrees of relevance [63], [68]. The tags with different relevance can benefit visual search performance and in turn improve the relevance for advertising.
- 2) Finding advertising visual keywords in images or videos can make the advertisements more visually relevant. Intuitively, the ROIs in an image or video are naturally the most suitable regions that might attract a lot of eyeballs and thus better information carriers for advertising. The techniques for detecting ROI can help to promote the ad relevance and impression [18], [70]. Quality assessment can be used for finding advertisable video content with high visual quality [81].
- 3) To improve advertising relevance, in addition to categorize source media, the advertisements are expected to be automatically classified to a predefined categories, so that it would be easy to deliver specified and relevant ads to specified users. To this end, similar to the LSCOM ontology for video domain [84], [119], we need to build an ontology for advertisements. Some preliminary works have focused on this issue [19], [25], [67].
- 4) To improve user relevance for targeted advertising, traditional techniques for relevance feedback

in multimedia information retrieval can be used to deliver user relevant advertisements on the fly [92], [124]. A recent work has attempted to recognize human behavior and understand a user's mobility from sensor data like GPS logs [123]. Based on the mobility information, the advertisements can be context-aware in terms of location. Behavior targeting is one of the important research topics in the general advertising. By studying the demographic and user behavior in an advertising ecosystem, the advertisements could be tailored to the targeted users [37], [113].

- 5) To improve user involvement of the advertisement, the emotional or affective models could be used to compute the emotional involvement of video program [32], [33], [121]. This kind of involvement information has been proved to be a key effect on selecting suitable advertisements within a contextual program [20], [28].

How can we achieve the best trade-off advertising strategy which can satisfy different users at the same time? As we have discussed in Section IV that different combinations of relevance and different computation of contextual intrusiveness fit different roles (i.e., publishers, advertisers, and viewers) in an advertising ecosystem. Although we have adopted to make the inserted advertisements look similar to the source media, which strategy is the best still remains an open issue. The studies from the perspective of psychology and user behavior analysis are desired. Another possible solution is to make advertising interactive and let viewers be engaged and interact with publishers and advertisers.

How can we design suitable advertising approaches for different media genres? As we have various advertising approaches in the literature, which one is the best for a specific kind of medium? For example, if the source medium is user-generated content like an amateur video from Youtube [118], the intelligent overlay video advertising proposed in [31], [76] might be more suitable due to the typical short duration of the source video. Otherwise, if the source medium is a premier feature movie or professional video from Hulu [43], AOL [5], or MSN Video [82], the in-video advertising [79], [80] which is more aggressive than overlay advertising might be more suitable. Another example is that if the source medium is a high resolution image with huge file size and considerable transmission latency, then the inside image advertising proposed in [65] might be the best.

How can we make advertising more impressive so that users are more willing to interact with the attractive advertisements? This problem can be partially tackled from the following perspectives.

- 1) Designing advertising in a game form to make users participate the game and get some incentives simultaneously. For example, GameSense has proved that game-like advertising is a powerful

complement to existing multimedia advertising that can attract viewer's attention [59], [60].

- 2) Leveraging techniques in computer graphics to render the advertisements in a more visually appealing way. Rendering effect is one of the key problems that will influence user experience. It is not true that more aggressive rendering of advertisement is more effective [57], [74]. Sometimes, a moderate rendering strategy can make viewers feel more comfortable about the advertisement. For example, one can show the remaining duration of the now playing advertisement at certain position on the video or player. Another example is to enable users to click or minimize/maximize the overlay advertisements according to their interests.
- 3) Rather than directly push the product or service information around the media, we can automatically provide rich and related valuable information along with the advertisement embedded in-media. In this way, the user experience can be enriched by connecting more comprehensive and useful information with the contents of advertisements (e.g. weather, discount, traffic, education, traditional TV commercials, and so on) while users are consuming the advertisements [26], [105]. This would be a new advertising scenario for enriching advertising service. The techniques in human computer interaction such as user modeling and computer vision such as local feature based image matching may help this issue.

How can we adapt existing multimedia advertising to mobile domain? As the media capture and GPS functionalities have been widely equipped in mobile phones, PDAs, and other digital devices, mobile multimedia advertising has become a promising direction. The distinctiveness of mobile devices, including limited screen size and bandwidth, media capture (e.g., photo, video, and audio), GPS traces, and so on, bring quite a few challenges and new applications to multimedia advertising. One scenario is querying relevant advertising information about a product via photo capturing through mobile phone. The mobile search techniques in [48], [93], [110] can be used for searching advertisements with different modalities.

Although we have summarized the multimedia advertising as two generations, we believe that we are now facing the third generation, i.e., impressionative advertising. Using successful applications in the so-called Web 2.0 era as references, we believe that impressionative advertising is the next big bet. The primary characteristic of impressionative advertising is that more and more user interactions and mash-up of multimedia applications are getting more and more involved. Advertising will eventually become impressive and "game-ized." Through impressionative advertising, the users will find that the ads are more impressive, interactive, as well as game-like compared with the first and second advertising generations. ■

REFERENCES

- [1] AdSense. [Online]. Available: <http://www.google.com/adsense/>
- [2] AdWords. [Online]. Available: <http://adwords.google.com/>
- [3] L. Ahn and L. Dabbish, "Labeling images with a computer game," in *Proc. SIGCHI Conf. Human Factors Comput. Syst.*, 2004, pp. 319–326.
- [4] A. Anagnostopoulos, A. Z. Border, E. Gabrilovich, V. Josifovski, and L. Riedel, "Just-in-time contextual advertising," in *Proc. ACM Conf. Inform. Knowl. Management*, 2007, pp. 331–340.
- [5] AOL. [Online]. Available: <http://www.aol.com/>
- [6] N. Archak, A. Ghose, and P. Ipeirotis, "Show me the money: Deriving the pricing power of product features by mining consumer reviews," in *Proc. ACM Conf. Knowl. Discovery and Data Mining*, 2007.
- [7] R. Baeza-Yates and B. Ribeiro-Neto, *Modern Information Retrieval*. Reading, MA: Addison Wesley, 1999.
- [8] S. Boll, "Multitube—Where multimedia and web 2.0 could meet," *IEEE Multimedia*, vol. 14, no. 1, pp. 9–13, Jan.–Mar. 2007.
- [9] S. Boyd and L. Vandenberghe, *Convex Optimization*. Cambridge, U.K.: Cambridge University Press, 2004.
- [10] BritePic. [Online]. Available: <http://www.britepic.com/>
- [11] A. Broder, M. Fontoura, V. Josifovski, and L. Riedel, "A semantic approach to contextual advertising," in *Proc. ACM SIGIR Conf. Res. and Dev. Inform. Retrieval*, 2007.
- [12] D. Cai, S. Yu, J.-R. Wen, and W.-Y. Ma, "Extracting content structure for web pages based on visual representation," in *Fifth Asia Pacific Web Conf.*, 2003.
- [13] D. Cai, S. Yu, J.-R. Wen, and W.-Y. Ma, "VIPS: A vision-based page segmentation algorithm," Microsoft Technical Report (MSR-TR-2003-79), 2003.
- [14] D. Chakrabarti, D. Agarwal, and V. Josifovski, "Contextual advertising by combining relevance with click feedback," in *Proc. Int. WWW Conf.*, 2008, pp. 417–426.
- [15] C.-H. Chang, K.-Y. Hsieh, M.-C. Chung, and J.-L. Wu, "ViSA: Virtual spotlighted advertising," in *Proc. ACM Multimedia*, 2008, pp. 837–840.
- [16] P. Chatterjee, D. L. Hoffman, and T. P. Novak, "Modeling the clickstream: Implications for web-based advertising efforts," *Marketing Science*, vol. 22, no. 4, pp. 520–541, 2003.
- [17] X. Chen and H.-J. Zhang, "Text area detection from video frames," in *Proc. IEEE Pacific Rim Conf. Multimedia*, 2001, pp. 222–228.
- [18] Y.-Y. Chuang, "A collaborative benchmark for region of interest detection algorithms," in *Proc. IEEE Conf. Comput. Vision Pattern Recognit.*, 2009.
- [19] C. Colombo, A. D. Bimbo, and P. Pala, "Retrieval of commercials by semantic content: The semiotic perspective," *Multimedia Tools and Applications*, vol. 13, pp. 93–118, 2001.
- [20] K. S. Coulter, "The effects of affective response to media context on advertising evaluations," *J. Adv.*, vol. XXVII, no. 4, pp. 41–51, 1998.
- [21] H. Cui, J.-R. Wen, J.-Y. Nie, and W.-Y. Ma, "Probabilistic query expansion using query logs," in *Proc. Int. Conf. WWW*, 2002, pp. 325–332.
- [22] H. Dai, L. Zhao, Z. Nie, J.-R. Wen, L. Wang, and Y. Li, "Detecting online commercial intention," in *Proc. Int. WWW Conf.*, 2006.
- [23] R. Datta, D. Joshi, J. Li, and J. Z. Wang, "Image retrieval: Ideas, influences, and trends of the new age," *ACM Comput. Surveys*, vol. 40, no. 65, 2008.
- [24] DoubleClick. [Online]. Available: <http://www.doubleclick.com/>
- [25] 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 *Proc. ACM Multimedia*, 2006, pp. 201–210.
- [26] L.-Y. Duan, J. Wang, Y. Zheng, H. Lu, and J. S. Jin, "Digesting commercial clips from TV streams," *IEEE MultiMedia*, vol. 15, no. 1, pp. 28–41, 2008.
- [27] L. Fei-Fei, R. Fergus, and P. Perona, "One-shot learning of object categories," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 128, no. 4, pp. 594–611, 2006.
- [28] T. S. Feltham and S. J. Arnold, "Program involvement and ad/program consistency as moderators of program context effects," *J. Consumer Psychology*, vol. 3, no. 1, pp. 51–77, 1994.
- [29] R. J. Gerrig and P. G. Zimbardo, *Psychology and Life (16 Edition)*. Boston, MA: Allyn & Bacon, 2001.
- [30] Google Image Labeler. [Online]. Available: <http://images.google.com/imagelabeler/>; <http://images.google.com/imagelabeler/>
- [31] J. Guo, T. Mei, F. Liu, and X.-S. Hua, "AdOn: An intelligent overlay video advertising system," in *Proc. ACM SIGIR Conf. Res. Dev. Inform. Retrieval*, 2009, pp. 628–629.
- [32] A. Hanjalic, "Extracting moods from pictures and sounds: Towards truly personalized TV," *IEEE Signal Process. Mag.*, vol. 23, no. 2, pp. 90–100, Mar. 2006.
- [33] A. Hanjalic and L.-Q. Xu, "Affective video content representation and modeling," *IEEE Trans. Multimedia*, vol. 7, no. 1, pp. 143–154, 2005.
- [34] A. G. Hauptmann, M. Christel, and R. Yan, "Video retrieval based on semantic concepts," *Proc. IEEE*, pp. 602–622, 2008.
- [35] A. G. Hauptmann, W. H. Lin, R. Yan, J. Yang, and M. Y. Chen, "Extreme video retrieval: Joint maximization of human and computer performance," in *Proc. ACM Int. Conf. Multimedia*, Santa Barbara, 2006.
- [36] A. G. Hauptmann, R. Yan, W.-H. Lin, M. Christel, and H. Wactlar, "Can high-level concepts fill the semantic gap in video retrieval? A case study with broadcast news," *IEEE Trans. Multimedia*, vol. 9, no. 5, pp. 958–966, 2007.
- [37] J. Hu, H.-J. Zeng, H. Li, C. Niu, and Z. Chen, "Demographic prediction based on user's browsing behavior," in *Proc. Int. WWW Conf.*, 2007.
- [38] X.-S. Hua, L. Lu, and H.-J. Zhang, "Optimization-based automated home video editing system," *IEEE Trans. Circuit Syst. for Video Technol.*, vol. 14, no. 5, pp. 572–583, 2004.
- [39] X.-S. Hua, L. Lu, and H.-J. Zhang, "Robust learning-based TV commercial detection," in *Proc. ICME*, Jul. 2005.
- [40] X.-S. Hua, T. Mei, and S. Li, "When multimedia advertising meets the new internet era," in *Proc. IEEE Int. Workshop Multimedia Signal Process.*, 2008, pp. 1–5.
- [41] X.-S. Hua, P. Yin, and H.-J. Zhang, "Efficient video text recognition using multiple frame integration," in *Proc. IEEE Int. Conf. Image Process.*, 2002, pp. 397–400.
- [42] T. S. Huang, C. K. Dagli, S. Rajaram, E. Y. Chang, M. I. Mandel, G. E. Poliner, and D. P. W. Ellis, "Active learning for interactive multimedia retrieval," *Proc. IEEE*, vol. 96, no. 4, pp. 648–667, Apr. 2008.
- [43] Hulu. [Online]. Available: <http://www.hulu.com/>
- [44] IT Facts. [Online]. Available: <http://www.itfacts.biz/50-blm-digital-photos-taken-in-2007-60-blm-by-2011/8985>
- [45] iTVx. [Online]. Available: <http://www.itvx.com/>
- [46] R. J. K. Jacob and K. S. Karn, "Eye tracking in human-computer interaction and usability research: Ready to deliver the promises," in *The Mind's Eye: Cognitive and Applied Aspects of Eye Movement Research*, R. Radach, J. Hyona, and H. Deubel, Eds. Boston: North-Holland/Elsevier, 2003.
- [47] K. Jarvelin and J. Kekalainen, "Cumulated gain-based evaluation of IR techniques," *ACM Trans. Inform. Syst.*, vol. 20, no. 4, pp. 442–446, 2002.
- [48] W. Jiang, D. Liu, X. Xie, M. R. Scott, J. Tien, and D. Xiang, "An online advertisement platform based on image content bidding," in *IEEE Int. Conf. Multimedia & Expo*, 2009.
- [49] A. Joshi and R. Motwani, "Keyword generation for search engine advertising," in *Proc. Workshops IEEE Int. Conf. Data Mining*, 2006.
- [50] Jupiter Research. [Online]. Available: <http://www.jupiterresearch.com/>
- [51] G. Kastidou and R. Cohen, "An approach for delivering personalized ads in interactive TV customized to both users and advertisers," in *Proc. Eur. Conf. Interactive Television*, 2006.
- [52] L. Kennedy, S.-F. Chang, and A. Natsev, "Query-adaptive fusion for multimodal search," *Proc. IEEE*, vol. 96, no. 4, pp. 567–588, 2008.
- [53] L. Kennedy, M. Naaman, S. Ahern, R. Nair, and T. Rattenbury, "How Flickr helps us make sense of the world: Context and content in community-contributed media collections," in *Proc. ACM Multimedia*, Augsburg, Germany, 2007.
- [54] A. Lacerda, M. Cristo, M. A. Goncalves et al., "Learning to advertise," in *Proc. ACM SIGIR Conf. Res. Dev. Inform. Retrieval*, 2006.
- [55] G. Lekakos, D. Papakiriakopoulos, and K. Chorianopoulos, "An integrated approach to interactive and personalized TV advertising," in *Proc. Workshop on Personalization in Future TV*, 2001.
- [56] D. Li, B. Wang, Z. Li, N. Yu, and M. Li, "On detection of advertising images," in *IEEE Int. Conf. Multimedia & Expo*, 2007.
- [57] H. Li, S. M. Edwards, and J.-H. Lee, "Measuring the intrusiveness of advertisements: Scale development and validation," *J. Advertising*, vol. 31, no. 2, pp. 37–47, 2002.
- [58] H. Li, D. Zhang, J. Hu, H.-J. Zeng, and Z. Chen, "Finding keyword from online broadcasting content for targeted advertising," in *Int. Workshop on Data Mining and Audience Intelligence for Advertising*, 2007.
- [59] L. Li, T. Mei, and X.-S. Hua, "GameSense: Game-like in-image advertising," *Multimedia Tools and Applications*, 2009.

- [60] L. Li, T. Mei, C. Liu, and X.-S. Hua, "GameSense," in *Proc. Int. WWW Conf.*, Madrid, Spain, Apr. 2009, pp. 1097–1098.
- [61] S. Z. Li, L. Zhu, Z. Zhang, A. Blake, H.-J. Zhang, and H. Shum, "Statistical learning of multi-view face detection," in *Proc. Eur. Conf. Computer Vision*, Copenhagen, Denmark, May 2002, pp. 67–81.
- [62] T. Li, T. Mei, I.-S. Kweon, and X.-S. Hua, "Video^M: Multi-video synopsis," in *1st Int. Workshop on Video Mining, in Conjunction With the IEEE Int. Conf. Data Mining*, 2008, pp. 854–861.
- [63] X. Li, C. G. M. Snoek, and M. Worring, "Learning tag relevance by neighbor voting for social image retrieval," in *Proc. ACM Multimedia Inform. Retrieval*, 2008.
- [64] Y. Li, K. Wan, X. Yan, and C. Xu, "Advertisement insertion in baseball video based on advertisement effect," in *Proc. ACM Multimedia*, 2005, pp. 343–346.
- [65] Z. Li, L. Zhang, and W.-Y. Ma, "Delivering online advertisements inside images," in *Proc. ACM Multimedia*, 2008, pp. 1051–1060.
- [66] W.-S. Liao, K.-T. Chen, and W. H. Hsu, "Adimage: Video advertising by image matching and ad scheduling optimization," in *Proc. ACM SIGIR Conf. Res. Dev. Inform. Retrieval*, 2008, pp. 767–768.
- [67] R. Lienhart, C. Kuhmunch, and W. Effelsberg, "On the detection and recognition of television commercials," in *Proc. Int. Conf. Multimedia Comput. Syst.*, Ottawa, Canada, 1997, pp. 509–516.
- [68] D. Liu, X.-S. Hua, L. Yang, M. Wang, and H.-J. Zhang, "Tag ranking," in *Proc. Int. WWW Conf.*, 2009.
- [69] H. Liu, S. Jiang, Q. Huang, and C. Xu, "A generic virtual content insertion system based on visual attention analysis," in *Proc. ACM Int. Conf. Multimedia*, 2008, pp. 379–388.
- [70] T. Liu, J. Sun, N.-N. Zheng, X. Tang, and H.-Y. Shum, "Learning to detect a salient object," in *Proc. IEEE Conf. Comput. Vis. Pattern Recog.*, 2007, pp. 1–8.
- [71] D. G. Lowe, "Distinctive image features from scale-invariant keypoints," *Int. J. Comput. Vis.*, vol. 60, no. 2, pp. 91–110, 2004.
- [72] Y.-F. Ma, L. Lu, H.-J. Zhang, and M. Li, "A user attention model for video summarization," in *Proc. ACM Multimedia*, 2002, pp. 533–542.
- [73] Y.-F. Ma and H.-J. Zhang, "Contrast-based image attention analysis by using fuzzy growing," in *Proc. ACM Multimedia*, Nov. 2003, pp. 374–381.
- [74] S. Mccoy, A. Everard, P. Polak, and D. F. Galletta, "The effects of online advertising," *Commun. ACM*, vol. 50, no. 3, pp. 84–88, 2007.
- [75] A. Mehta, A. Saberi, U. Vazirani, and V. Vazirani, "Adwords and generalized on-line matching," *J. ACM*, vol. 54, no. 5, Oct. 2007.
- [76] T. Mei, J. Guo, X.-S. Hua, and F. Liu, "AdOn: Toward contextual overlay in-video advertising," *Multimedia Syst.*, 2010.
- [77] T. Mei, X.-S. Hua, W. Lai, L. Yang *et al.*, "MSRA-USTC-SJTU at TRECVID 2007: High-level feature extraction and search," in *TREC Video Retrieval Evaluation Online Proc.*, 2007.
- [78] T. Mei, X.-S. Hua, and S. Li, "Contextual in-image advertising," in *Proc. ACM Multimedia*, Vancouver, Canada, 2008, pp. 439–448.
- [79] T. Mei, X.-S. Hua, and S. Li, "VideoSense: A contextual in-video advertising system," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 19, no. 12, pp. 1866–1879, Dec. 2009.
- [80] T. Mei, X.-S. Hua, L. Yang, and S. Li, "VideoSense: Towards effective online video advertising," in *Proc. ACM Multimedia*, Augsburg, Germany, 2007, pp. 1075–1084.
- [81] T. Mei, X.-S. Hua, C.-Z. Zhu, H.-Q. Zhou, and S. Li, "Home video visual quality assessment with spatiotemporal factors," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 17, no. 6, pp. 699–706, Jun. 2007.
- [82] MSN Video. [Online]. Available: <http://video.msn.com/>
- [83] V. Murdock, M. Ciaramita, and V. Plachouras, "A noisy-channel approach to contextual advertising," in *1st Int. Workshop on Data Mining and Audience Intelligence for Advertising*, 2007, pp. 96–99.
- [84] M. Naphade, J. R. Smith, J. Tesic, S.-F. Chang, W. Hsu, L. Kennedy, A. Hauptmann, and J. Curtis, "Large-scale concept ontology for multimedia," *IEEE MultiMedia*, vol. 13, no. 3, pp. 86–91, Jul.–Sep. 2006.
- [85] Online Publishers. [Online]. Available: <http://www.online-publishers.org/>
- [86] A. Poole and L. J. Ball, "Eye tracking in human-computer interaction and usability research: Current status and future prospects," in *Encyclopedia of Human Computer Interaction*, Idea Group, 2004.
- [87] G.-J. Qi, X.-S. Hua, Y. Rui, J. Tang, T. Mei, and H.-J. Zhang, "Correlative multi-label video annotation," in *ACM Multimedia*, Augsburg, Germany, Sep. 2007, pp. 17–26.
- [88] Revver. [Online]. Available: <http://one.revver.com/revver>
- [89] B. Ribeiro-Neto, M. Cristo, P. B. Golgher, and E. S. Moura, "Impedance coupling in content-targeted advertising," in *Proc. ACM SIGIR Conf. Res. Devel. Inform. Retrieval*, 2005.
- [90] M. Richardson, E. Dominowska, and R. Ragno, "Predicting clicks: Estimating click-through rate for new ads," in *Proc. Int. WWW Conf.*, 2007.
- [91] C. Rohrer and J. Boyd, "The rise of intrusive online advertising and the response of user experience research at Yahoo!" in *Proc. SIGCHI Conf. Human Factors Comput. Syst.*, 2004.
- [92] Y. Rui, T. S. Huang, M. Ortega, and S. Mehrotra, "Relevance feedback: A power tool for interactive content-based image retrieval," *IEEE Trans. Circuits Video Technol.*, vol. 8, no. 5, pp. 644–655, Sep. 1998.
- [93] M. R. Scott, W. Jiang, X. Xie, J. Tien, G. Chen, and D. Xiang, "VisiAds: A vision-based advertising platform for camera phones," in *IEEE Int. Conf. Multimedia & Expo*, 2009.
- [94] D. Shen, J.-T. Sun, Q. Yang, and Z. Chen, "Building bridges for web query classification," in *Proc. ACM SIGIR Conf. Res. Devel. Inform. Retrieval*, 2006.
- [95] C. G. M. Snoek, M. Worring, J. C. van Gemert, J. M. Geusebroek, and A. W. M. Smeulders, "The challenge problem for automated detection of 101 semantic concepts in multimedia," in *Proc. ACM Int. Conf. Multimedia*, Santa Barbara, 2006.
- [96] S. H. Srinivasan, N. Sawant, and S. Wadhwa, "vAdeo: Video advertising system," in *Proc. ACM Multimedia*, 2007, pp. 455–456.
- [97] A. Thawani, S. Gopalan, and V. Sridhar, "Context aware personalized ad insertion in an interactive TV environment," in *Proc. Workshop on Personalization in Future TV*, 2004.
- [98] TRECVID. [Online]. Available: <http://www.nlp.ir.nist.gov/projects/trecvid/>
- [99] S. Velusamy, L. Gopal, S. Bhatnagar, and S. Varadarajan, "An efficient ad recommendation system for TV programs," *Multimedia Syst.*, vol. 14, pp. 73–87, Jul. 2008.
- [100] Vibrant Media. [Online]. Available: <http://www.vibrantmedia.com/>
- [101] Videoegg. [Online]. Available: <http://www.videoegg.com/>
- [102] E. M. Voorhees and D. K. Harman, *TREC: Experiment and Evaluation in Information Retrieval*. Cambridge, MA: MIT Press, 2005.
- [103] K. Wan, X. Yan, X. Yu, and C. Xu, "Robust goal-mouth detection for virtual content insertion," in *Proc. ACM Multimedia*, Nov. 2003, pp. 468–469.
- [104] G. Wang and D. Forsyth, "Object image retrieval by exploiting online knowledge resources," in *Proc. IEEE Conf. Comput. Vis. Pattern Recog.*, 2008.
- [105] J. Wang and L.-Y. Duan, "Linking video ads with product or service information by web search," in *IEEE Int. Conf. Multimedia & Expo*, 2009.
- [106] M. Weideman and T. Haig-Smith, "An investigation into search engines as a form of targeted advert delivery," in *South African Institute of Computer Scientists and Information Technologies on Enablement Through Technology*, 2002, pp. 258–258.
- [107] D. Whitley, "A genetic algorithm tutorial," *Stat. Comput.*, vol. 4, pp. 65–85, 1994.
- [108] Wikipedia. [Online]. Available: http://en.wikipedia.org/wiki/click-through_rate
- [109] L. Wu, X.-S. Hua, N. Yu, W.-Y. Ma, and S. Li, "Flickr distance," in *Proc. ACM Multimedia*, Vancouver, Canada, 2008, pp. 31–40.
- [110] X. Xie, L. Lu, M. Jia, H. Li, S. Frank, and W.-Y. Ma, "Mobile search with multimodal queries," *Proc. IEEE*, vol. 96, no. 4, pp. 589–601, Apr. 2008.
- [111] C. Xu, K. W. Wan, S. H. Bui, and Q. Tian, "Implanting virtual advertisement into broadcast soccer video," in *Proc. IEEE Pacific Rim Conf. Multimedia*, 2004, pp. 264–271.
- [112] Yahoo!. [Online]. Available: <http://www.yahoo.com/>
- [113] J. Yan, N. Liu, G. Wang, W. Zhang, Y. Jiang, and Z. Chen, "How much can behavioral targeting help online advertising?" in *Proc. WWW*, 2009, pp. 261–270.
- [114] B. Yang, T. Mei, X.-S. Hua, L. Yang, S.-Q. Yang, and M. Li, "Online video recommendation based on multimodal fusion and relevance feedback," in *Proc. ACM Int. Conf. Image Video Retrieval*, 2007.
- [115] K. Yang, M. Wang, and H.-J. Zhang, "Active tagging for image indexing," in *Proc. Int. Workshop Internet Multimedia Search Mining, in Conjunction With ICME*, 2009.
- [116] Y. Yang and X. Liu, "A re-examination of text categorization methods," in *Proc. ACM SIGIR Conf. Res. Devel. Inform. Retrieval*, 1999.
- [117] W.-T. Yih, J. Goodman, and V. R. Carvalho, "Finding advertising keywords on web pages," in *Proc. Int. WWW Conf.*, 2006.
- [118] YouTube. [Online]. Available: <http://www.youtube.com/>
- [119] Z.-J. Zha, T. Mei, Z. Wang, and X.-S. Hua, "Building a comprehensive ontology to refine video concept detection," in *Proc. ACM SIGMM Int. Conf. Workshop*

Mei and Hua: Contextual Internet Multimedia Advertising

Multimedia Inform. Retrieval, Augsburg, Germany, Sep. 2007, pp. 227–236.

- [120] H.-J. Zhang, A. Kankanhalli, and S. W. Smoliar, "Automatic partitioning of full-motion video," *Multimedia Syst.*, vol. 1, no. 1, pp. 10–28, Jun. 1993.
- [121] S. Zhang, Q. Tian, S. Jiang, Q. Huang, and W. Gao, "Affective mtv analysis based on arousal and valence features," in *Proc. ICME*, 2008, pp. 1369–1372.
- [122] L. Zhao, W. Qi, Y.-J. Wang, S.-Q. Yang, and H.-J. Zhang, "Video shot grouping using best first model merging," in *Proc. Storage and Retrieval for Media Database*, 2001, pp. 262–269.
- [123] Y. Zheng, Q. Li, Y. Chen, and X. Xie, "Understanding mobility based on GPS data," in *Proc. ACM Conf. Ubiquitous Comput.*, Seoul, Korea, 2008, pp. 312–321.
- [124] X. S. Zhou and T. S. Huang, "Relevance feedback for image retrieval: A comprehensive review," *Multimedia Syst.*, vol. 8, no. 6, pp. 536–544, Apr. 2003.

ABOUT THE AUTHORS

Tao Mei (Member, IEEE) received the B.E. degree in automation and the Ph.D. degree in pattern recognition and intelligent systems from the University of Science and Technology of China, Hefei, in 2001 and 2006, respectively. He joined Microsoft Research Asia, Beijing, China, as a Researcher Staff Member, in 2006. He was a visiting professor of the Xidian University, Xi'an, China, during 2009–2012. His current research interests include multimedia content analysis, computer vision, and internet multimedia applications such as search, advertising, management, social network, and mobile applications. He is the author of one book, five book chapters, and over 80 journal and conference papers in these areas, and holds more than 20 filed patents or pending applications.



Dr. Mei serves as an Editorial Board Member of *Journal of Multimedia*, a Guest Editor for *IEEE Multimedia Magazine*, *ACM/Springer Multimedia Systems*, and *Journal of Visual Communication and Image Representation*. He was the principle designer of the automatic video search system that achieved the best performance in the worldwide TRECVID evaluation in 2007. He received the Best Paper and Best Demonstration Awards in the ACM International Conference on Multimedia 2007, the Best Poster Paper Award in the IEEE International Workshop on Multimedia Signal Processing 2008, and the Best Paper Award in the ACM International Conference on Multimedia 2009.

Dr. Mei is a member of the Association for Computing Machinery.

Xian-Sheng Hua (Member, IEEE) received the B.S. and Ph.D. degrees from Peking University, Beijing, China, in 1996 and 2001, respectively, both in applied mathematics. When he was in Peking University, his major research interests were in the areas of image processing and multimedia watermarking. Since 2001, he has been with Microsoft Research Asia, Beijing, where he is currently a Lead Researcher with the media computing group. His current research interests are in the areas of video content analysis, multimedia search, management, authoring, sharing, mining, advertising and mobile multimedia computing. He has authored or co-authored more than 160 publications in these areas and has more than 40 filed patents or pending applications. He is now an adjunct professor of University of Science and Technology of China, and serves as an Associate Editor of *IEEE TRANSACTIONS ON MULTIMEDIA*, Associate Editor of *ACM Transactions on Intelligent Systems and Technology*, Editorial Board Member of *Advances in Multimedia* and *Multimedia Tools and Applications*, and editor of *Scholarpedia* (Multimedia Category). Dr. Hua won the Best Paper Award and Best Demonstration Award in ACM Multimedia 2007, Best Poster Paper Award in 2008 IEEE International Workshop on Multimedia Signal Processing, and Best Student Paper in ACM Conference on Information and Knowledge Management 2009. He also won 2008 MIT Technology Review TR35 Young Innovator Award, and named as one of the "Business Elites of People under 40 to Watch" by *Global Entrepreneur*.



Dr. Hua is a Senior Member of the Association for Computing Machinery.