Received 16 September 2014; revised 8 December 2014; accepted 14 December 2014.
Date of publication 7 January, 2015; date of current version 10 June, 2015.

Digital Object Identifier 10.1109/TETC.2014.2386140

Enhancing Product Detection With Multicue Optimization for TV Shopping Applications

FAUSTO C. FLEITES[1], (Student Member, IEEE), HAOHONG WANG[2], (Senior Member, IEEE), AND SHU-CHING CHEN[1], (Senior Member, IEEE)

[1]School of Computing and Information Sciences, Florida International University, Miami, FL 33174 USA
[2]General Manager of TCL Research America, Santa Clara, CA 95134 USA
CORRESPONDING AUTHOR: F. C. FLEITES (fflei001@cs.fiu.edu)

ABSTRACT Smart TVs allow consumers to watch TV, interact with applications, and access the Internet, thus enhancing the consumer experience. However, the consumers are still unable to seamlessly interact with the contents being streamed, as it is highlighted by TV-enabled shopping. For example, if a consumer is watching a TV show and is interested in purchasing a product being displayed, the consumer can only go to a store or access the Web to make the purchase. It would be more convenient if the consumer could interact with the TV to purchase interesting items. To realize this use case, products in the content stream must be detected so that the TV system notifies consumers of possibly interesting ones. A practical solution must address the detection of complex products, i.e., those that do not have a rigid form and can appear in various poses, which poses a significant challenge. To this end, a multicue product detection framework is proposed for TV shopping. The framework is generic as it is not tied to specific object detection approaches. Instead, it utilizes appearance, topological, and spatio-temporal cues that make use of a related, easier to detect object class to improve the detection results of the target, more difficult product class. The three cues are jointly considered to select the best path that occurrences of the target product class can follow in the video and thus eliminate false positive occurrences. The empirical results demonstrate the advantages of the proposed approach in improving the precision of the results.

INDEX TERMS Smart TV, TV shopping, spatio-temporal information, multimedia content analysis, dynamic programming.

I. INTRODUCTION

Recently, smart TVs have raised the TV experience to a new level by combining the TV, Internet, and PC technologies. Consumers are able to browse the web, interact with a variety of applications, and watch TV channels. Nevertheless, smart TVs still do not allow consumers to seamlessly interact with the contents being streamed. One example of such a drawback is TV shopping. In this use case, the consumer interacts with the TV to purchase an interesting product that is displayed in the current show. For instance, consider the consumer is watching a fashion show. The TV system detects hand bags and apparel in the content stream and notifies the consumer via a non-intrusive notification. When the consumer sees the notification, he or she can activate it and is then presented with the list of products detected in the show over a time window. If the consumer is interested in an item, he or she selects the item and proceeds to purchase it. The purchase can be realized, for example, by providing the detected items as search input in an online store. Clearly, such a content-enabled application is of commercial value and would significantly enrich the interaction of consumers with their TV systems.

The fundamental challenge to realize TV shopping is detecting objects in the content stream to be able to signal consumers of interesting products. Particularly, the system must be able to detect complex objects, i.e., those that do not have a rigid form or can appear in a variety of poses. In the fashion show use case, detecting hand bags is very difficult as hand bags do not have a definitive shape, can present deformations, be occluded by hands or arms, and appear in many poses. The object detection task can be addressed by considering the video as an unrelated sequence of frames and perform static object detection [1]–[5]. On the other hand, it can be tackled by utilizing the additional information

offered by the progression of the video sequence [6]–[12]. Nevertheless, most of these approaches fail to detect complex objects and perform well only on ideal conditions. In the case of video based approaches, these mostly concentrate on using motion information to detect moving objects, which may not work well for difficult objects; in the case of hand bags, the motion of the bag would be masked by the motion of the person carrying the bag.

The significant challenge posed by complex objects gives rise to using additional information to improve the detection results. For still images, most approaches employ additional information from co-occurrence and/or spatial relationships between object labels [13]–[20], but these do not incorporate the temporal information embedded in video sequences. Very few approaches do address the temporal information as an additional cue [21]–[23]. However, most of these are meant for surveillance applications, which have different requirements and scene characteristics than those of fashion shows. The latter are characterized by many shots with varying background motion, making very difficult the differentiation of foreground motion.

Building upon the idea of utilizing additional information to improve detection results, this article proposes a generic, multi-cue product detection framework for TV shopping. The approach is generic as is not tied to a specific detection approach. It makes use of multi-cue information to enhance the detection of complex objects in unconstrained video sequences, i.e., no assumptions are made about foreground or background motion in the video. Three cues are considered to detect objects of a target product class. The first one is the appearance cue, which dictates that the visual appearance of an object must represent the target product class. The second and third are topological and spatio-temporal cues, which consider the relationship between the target product class and a related, easier-to-detect object class. Within a video frame, the topological cue enforces a spatial relationship between detections of both classes. Across consecutive frames, the spatio-temporal cue assumes there is a correlation between the spatial positions of detections in both classes. The three cues are jointly considered to formulate an optimization problem that selects the best path objects of the target product class can follow in the video. Then, detections that do not belong to the selected path are regarded are false positive detections. To the be best of our knowledge, the proposed approach is the first attempt to combine appearance, spatio-temporal, and topological information into a path-optimization problem to enhance object detection in unconstrained video sequences.

The rest of the paper is organized as follows. Section II discusses the related work. Section III presents the proposed multi-cue product detection framework. Section IV describes the experiments and results. Finally, Section V concludes this article.

II. RELATED WORK

The utilization of additional information has been approached in recent years to try to overcome the challenges posed by object detection. Plenty of approaches tackle this task in static images utilizing co-occurrence and/or spatial relationships [13]–[20], [24], [25]. However, very few approaches address this problem for videos. These additionally include temporal relationships to exploit the inherent spatio-temporal information [21]–[23].

A. CO-OCCURRENCE AND SPATIAL RELATIONSHIPS

Even though they utilize additional information to aid object detection, the following approaches are intended for image data and thus do not take into consideration temporal information.

Galleguillos et al. [15] present an object detection framework that utilizes co-occurrence and spatial relationships as an additional source of information. Their framework is called CoLA for co-occurrence, location, and appearance. It first segments the objects based on their appearance, and the contextual information is integrated via a conditional random field (CRF) that aims at maximizing the agreement between object labels. The spatial information consists of four relationships: above, below, inside, and around, where each relationship is represented via a context matrix that encodes corresponding co-occurrence information.

Heitz et al. [14] propose a probabilistic framework that models contextual information to enhance the detection results of off-the-shelf detectors. It does so by rescoring the detections scores with the intent of lowering the scores of false-positive detections. The framework captures contextual relationships between ''stuff'' (i.e., regions with homogeneous or repetitive patterns) and ''things'' (i.e., monolithic objects) and does not require manual labeling of the ''stuff'' regions, only limited ground-truth labeling of object detections. In this sense, this approach combines co-occurrence and spatial relationships. Image regions (''stuff'') are provided as input to this approach, and the framework clusters the regions based on their ability to serve as context for object detection. A probabilistic model is then learned to link detections with ''stuff'' clusters.

Zheng et al. [20] propose a context-modeling framework that extends the idea of Heitz et al. [14]. The authors categorize types of context as ''Scene-Thing'', ''Stuff-Stuff'', ''Thing-Thing'', and ''Thing-Stuff'', and their framework models ''Thing-Thing'' and ''Thing-Stuff'' contexts by learning co-occurrence and spatial contextual relationships. Contextual information is represented via a polar geometric context descriptor. The framework then utilizes a maximum margin context (MMC) model to evaluate the usefulness of contextual information and fuse context information with appearance information. It does so by discriminatively learning a context risk function that measures the rank information between true positive and false positive detections. The empirical results on several PASCAL VOC datasets show that the framework outperforms that of Heitz et al. [14] for some concepts and achieves similar performance for others.

B. ADDITIONAL TEMPORAL RELATIONSHIP

The following approaches target video data and include temporal information.

However, most of them target surveillance applications, which have different requirements and assumptions (e.g., fixed camera) that are invalid in the TV shopping use case.

Sheikh et al. [21] introduce an object detection framework for surveillance videos that is able to model dynamic backgrounds. Different from previous approaches, the framework does not model pixel intensities as independent random variables; instead, the framework models the background as a single probability density using non-parametric kernel density estimation over joint location-color representation of image pixels. Object detection is approached by also maintaining a foreground model that is modeled similarly to that of the background and using both models competitively in a decision framework. The foreground model is enhanced with a temporal criterion, under which foreground objects are assumed to maintain small frame-to-frame color transformations and spatial changes. The decision framework is based on a maximum a posteriori Markov Random Field (MAP-MRF), which transforms the problem of object detection into a pixel-level binary classification problem that combines the foreground and background probability models in the likelihood function for each pixel. Even though this approach models dynamic backgrounds, it does so specifically for surveillance videos and thus assumes a fixed camera. Hence, it cannot be applied in completely unconstrained videos.

Yan et al. [22] propose to use pairwise constraints to aid video object classification with insufficient labeled data in surveillance videos. Indicating whether two examples are of the same class or not, pairwise constraints exploit the spatio-temporal continuity of video streams. As an example, the authors illustrate their method with the use case of classifying people's identities. In this task, two overlapping objects from consecutive frames can be considered to have the same identity, but two objects that appear in the same video frame cannot. Moreover, identities can be differentiated using a face comparison mechanism, which represents another source of pairwise constraints, instead of building statistical models for every possible subject. The authors present three discriminative learning methods that minimize the regularized empirical risk and incorporate pairwise constraints by penalizing their violation. This approach focuses on a use case that is different from TV shopping, where the definition of pairwise constraints is not directly applicable. In addition, it is meant for surveillance videos.

Yang et al. [23] introduce an object tracking framework that utilizes additional information to diminish the possibility of drifting. Their idea is to automatically mine auxiliary regions that have high co-occurrence and motion correlation, at least for a short period of time, with the target object and use their collaborative tracking to prevent the target tracker from drifting. Such auxiliary regions consist of ''significant'' color regions that are obtained using the classical split-merge quad-tree color segmentation and are represented via color histograms. Using simple histogram matching, coherent color regions are matched as the frame sequence progresses, a transaction set is constructed, and regions with high co-occurrence with the target object are chosen as candidate auxiliary regions. Such candidate regions are then tracked using a mean-shift tracker. The framework determines the candidate regions that have motion correlation with the target object via a subspace analysis on an assumed affine model between the candidate auxiliary regions and the target object. Once co-occurrent, motion-correlated auxiliary regions are determined, collaborative tracking is achieved by modeling a random field among the auxiliary regions and the target object. The random field is formulated under a Markov network with a star topology, and a two-step belief propagation algorithm is used to estimate the posterior probabilities of the network. Lastly, the framework includes a mechanism to detect inconsistent tracking estimates, which are regarded as outliers. If the outlier is an auxiliary object, then it is removed from the collaborative tracking; however, if the outlier is the target object, then it is considered to be experiencing occlusion or drift, and it is suspended from the tracking temporarily. This approach uses auxiliary regions to improve object tracking, which can be considered to have the same purpose as the occurrences of the related object class in the proposed framework. However, the method is different than the proposed one and assumes that auxiliary regions can be both obtained via color segmentation and tracked by a simple mean-shift tracker, which are not practical assumptions in the TV shopping use case.

III. MULTI-CUE PRODUCT DETECTION

As previously introduced, effectively addressing the TV shopping use case requires the detection of possibly complex objects of a target product class $\mathcal{C}$ in an unconstrained video sequence $\mathcal{V} = \{\mathcal{F}_i\}$, where $\mathcal{F}_i$ is the ith frame in $\mathcal{V}$. Hence, the problem at hand consists of obtaining the product occurrences of class $\mathcal{C}$ in $\mathcal{V}$. This article proposes to solve the stated problem by dividing $\mathcal{D}$ into shots $\{S_k\}$, followed by detecting all product occurrences of class $\mathcal{C}$ in each shot $\mathcal{S}_i$ using additional cues. More specifically, the approach consists of two steps that are applied on $\mathcal{S}_i$. Firstly, an object detector is executed on each frame. The detection threshold of the particular object detector is lowered to increase the changes of detecting complex objects, at the expense of increasing the number of false positive detections but increasing the changes of detecting complex product occurrences. Secondly, additional cues are utilized to obtain the optimal path product occurrences should follow across $\mathcal{S}_i$. The optimal path identifies the true positive occurrences and serves to weed out false positive detections. Since the video is divided into shots, it is assumed the shot-boundary method employed for this task will not fragment a continuous scene into many separate shots.

The additional cues consist of appearance, topological, and spatio-temporal relationships. The appearance cue refers to the visual appearance of the target product class, i.e., how

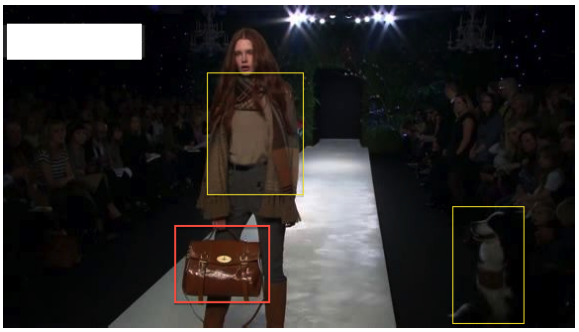


FIGURE 1. Example of how the appearance cue helps discern between possible object detections.

much influence has the visual appearance of the object in determining its class. An example of this cue is depicted in figure 1. It represents the task of detecting bags in fashion shows, where the target product class is ''bags''. An object detector for bags could generate the three bounding boxes shown in the figure. However, the visual features of the red bounding box should indicate that indeed this is the correct detection. The detection score of each bounding box can be used to quantify the visual information. Additionally, besides analyzing a frame in a vacuum, the appearance cue applies to nearby frames. Product occurrences in a neighborhood of frames must have similar visual appearance as the same object should have small changes in appearance from frame to frame.

In contrast, the topological and spatio-temporal cues refer to relations the target product class has with a related object class $\mathcal{C}_R$. An implicit requirement is the related object class must be easier to detect in the sense there is a mature technology that robustly detects objects of $\mathcal{C}_R$. For example, for the task of detecting bags in fashion shows, the related object class is ''faces''. The technology for face detection is quite robust, and thus it is possible to use face detection results to enhance the detection of objects of the class ''bags''. Nonetheless, false positive detections of the related object class can still occur. The topological relationship constricts the possible locations for occurrences of the target product class with respect to locations of occurrences of the related object class. Resuming the fashion show example, in the case the model is carrying a bag as depicted in figure 2, there is a clear positional relationship between the model's face and the bag. Based on this topological relationship, it is clearly possible to use the position of the model's face to restrict the possible locations for the bag. Lastly, the progression of video frames creates a spatio-temporal correlation between consecutive positions of the target product class and consecutive positions of the related object class, as depicted in figure 3a. Another example suitable for the related object class in fashion shows is ''persons''. Figure 4 depicts the positional relationship between the bag and the model.

Based the multiple cues, the proposed approach analyzes the best path occurrences of the target product class can

FIGURE 2. Example of topological relationship between "bags" and "faces" in a fashion show. Yellow bounding boxes depict object detections.

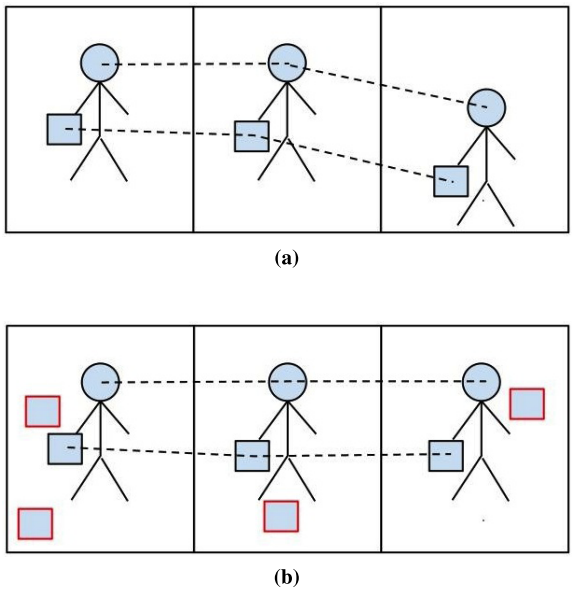


FIGURE 3. Illustration of the optimal path. (a) Spatio-temporal correlation between paths of the target product class and the related object class. (b) The optimal path according to the multiple cues is selected, pruning false positive detections of the target product class.

follow in a video shot. Figure 3b depicts how the optimal path weeds out false positive detections. In a succession of video frames $\{\mathcal{F}_i\}_{i=1}^{M}$, let a possible path be $\mathcal{P} = \{\mathcal{O}_i\}_{i=1}^{M}$, where $\mathcal{O}_i$ is an occurrence of the target product class in $\mathcal{F}_i$; $Q(\mathcal{P})$ be the quality of $\mathcal{P}$; and $\mathcal{R}_i$ is an occurrence of the related object

FIGURE 4. Example of topological relationship between "bags" and "persons" in a fashion show.

class in $\mathcal{F}_i$. The definition of $Q(\mathcal{P})$ is based on the following criteria imposed by the aforementioned cues.

A. APPEARANCE

The appearance cue is modeled as the probability $P(\mathcal{O}_i|\mathcal{C})$, which can be obtained by training an object detector to detect occurrences of the target product class. Any object detection can be utilized, but it is recommended a robust one that can detect occlusions and pose variations. The trained object detector must provide $P(\mathcal{O}_i|\mathcal{C})$ as well as the corresponding bounding box, i.e., the location and size of $\mathcal{O}_i$. Moreover, consecutive product occurrences in the path must have a high appearance similarity, which is defined by

$$\Omega(\mathcal{O}_i, \mathcal{O}_j) = \begin{cases} 0 & \text{if } i \leq 0 \\ s(\tau(\mathcal{O}_i), \tau(\mathcal{O}_j)) & \text{otherwise} \end{cases} \quad (1)$$

where $i < j$, $\tau(.)$ is the feature vector representation of a product occurrence's bounding box, and $s(.)$ is a function that measures the similarity between the feature vectors of two occurrences, where the image of $s(.)$ is [0, 1].

B. TOPOLOGICAL

Based on the location of the related object class, the topological cue specifies the set of locations from which a product occurrence should not deviate in a frame. With respect to the example of detecting bags in a fashion show, occurrences of "bags" should not be located too far from the location of the model's face. This requirement is modeled via the following function:

$$\Psi(\mathcal{O}_i) = \frac{d_l(l(\mathcal{O}_i), l(\mathcal{R}_i))}{b_m} \quad (2)$$

where $l(.)$ provides the location of a detection in the video frame, $d_l(.)$ is the Euclidean distance between two positions in the frame, and b_m is a constant that measures the diagonal length of the video frames. Hence, $\Psi(\mathcal{O}_i)$ assigns larger values the farther the product occurrence is from the related object, and its image is in [0, 1]. It is important to highlight that $\Psi(.)$ can be defined differently depending on the use case.

C. SPATIO-TEMPORAL

With respect to the path of the related object class, the spatio-temporal cue imposes a similar within-path deviation in the trajectory of the target product class. This constraint is modeled via the function:

$$\Gamma(\mathcal{O}_i, \mathcal{O}_j) = \begin{cases} 0 & \text{if } i \leq 0 \\ \dfrac{\min(d_l(\mathcal{O}_i, \mathcal{O}_j), d_l(\mathcal{R}_i, \mathcal{R}_j))}{\max(d_l(\mathcal{O}_i, \mathcal{O}_j), d_l(\mathcal{R}_i, \mathcal{R}_j)) + \epsilon} & \text{otherwise} \end{cases} \quad (3)$$

where $i < j$ and ϵ is a small constant ≥ 0 to avoid dividing by zero. The function $\Gamma(.)$ is proportional to the translational difference between the target product class and the related object class, and its image is in [0, 1].

The best path $\mathcal{P}^*$ that occurrences of the target product class can follow in $\{\mathcal{F}_i\}$ must have the highest $\sum_{i=1}^{M} P(\mathcal{O}_i|\mathcal{C})$, the highest $\sum_{i=1}^{M} \Omega(\mathcal{O}_{i-1}, \mathcal{O}_i)$, the lowest $\sum_{i=1}^{M} \Psi(\mathcal{O}_i)$, and the highest $\sum_{i=1}^{M} \Gamma(\mathcal{O}_{i-1}, \mathcal{O}_i)$. The optimal path can then be obtained by solving the following optimization problem:

$$\begin{aligned} \text{maximize } Q(\mathcal{P}) = \sum_{i=1}^{M} \Big\{ & \alpha P(\mathcal{O}_i|\mathcal{C}) + \beta \Omega(\mathcal{O}_{i-1}, \mathcal{O}_i) \gamma \big[1 - \Psi(\mathcal{O}_i)\big] \\ & + (1 - \alpha - \beta - \gamma) \Gamma(\mathcal{O}_{i-1}, \mathcal{O}_i) \Big\} \end{aligned} \quad (4)$$

where α, β, and γ are weight parameters in [0, 1].

For practical considerations, it is worth discussing the case where there are no occurrences of the related object class in a video shot. This scenario can be handled by equation (4) by setting $\gamma = 0$ and $\alpha + \beta = 1$, such that the influence of the topological and spatio-temporal cues becomes nil. The optimization problem then equates to

$$\text{maximize } Q(\mathcal{P}) = \sum_{i=1}^{M} \Big\{ \alpha P(\mathcal{O}_i|\mathcal{C}) + \beta \Omega(\mathcal{O}_{i-1}, \mathcal{O}_i) \Big\} \quad (5)$$

Equation (5) is a specific case of equation (4) that only considers the appearance cue, and thus the accuracy of the results largely depends on the performance provided by the detector of the target product class.

The optimal solution to equation (4) can be efficiently obtained using dynamic programming (DP). Let $\mathcal{P}_k$ denote the path formed by the first k elements in $\mathcal{P}$. Firstly, the cost function $\mathcal{G}_k(\mathcal{P}_k)$ is created to represent the maximum cost solution for the first k elements of $\mathcal{P}$ subject to the kth element is $\mathcal{O}_k$:

$$\mathcal{G}_k(\mathcal{P}_k) = \text{maximize}_{\mathcal{P}_{k-1}} Q(\mathcal{P}_k) \quad (6)$$

It is clear that maximizing $\mathcal{G}_M(\mathcal{P}_M)$ implies maximizing $Q(\mathcal{P})$. In addition, $\mathcal{G}_{k+1}(\mathcal{P}_{k+1})$ can be written as:

$$\begin{aligned} \mathcal{G}_{k+1}(\mathcal{P}_{k+1}) = \mathcal{G}_k(\mathcal{P}_k) + \Big\{ & \alpha P(\mathcal{O}_{k+1}|\mathcal{C}) + \beta \Omega(\mathcal{O}_k, \mathcal{O}_{k+1}) \gamma \big[1 - \Psi(\mathcal{O}_{k+1})\big] \\ & + (1 - \alpha - \beta - \gamma) \Gamma(\mathcal{O}_k, \mathcal{O}_{k+1}) \Big\} \end{aligned} \quad (7)$$

which shows that the selection of the $k+1$ occurrence in the path does not depend on the previously selected occurrences. This recursive representation makes the next step of the optimization process independent of the previous step, which is the foundation of DP.

Therefore, the problem can be interpreted as finding the longest path in a weighted, directed acyclic graph (DAG) $G = (V, E, w)$, where V is the set of vertices consisting of all the product occurrences found in $\{\mathcal{F}_i\}_{i=1}^{M}$, E is the set of edges $\{(\mathcal{O}_i, \mathcal{O}_{i+1})\}$ where $\mathcal{O}_i$ is a product occurrence in $\mathcal{F}_i$ and $\mathcal{O}_{i+1}$ in $\mathcal{F}_{i+1}$, and $w : E \rightarrow (R)$ is an edge-weight function that assigns a weight to each edge as follows:

$$w(\mathcal{O}_i, \mathcal{O}_{i+1}) = \Big\{\alpha P(\mathcal{O}_{i+1}|\mathcal{C}) + \beta\Omega(\mathcal{O}_k, \mathcal{O}_{i+1})\gamma\big[1 - \Psi(\mathcal{O}_{k+1})\big] + (1-\alpha-\beta-\gamma)\Gamma(\mathcal{O}_i, \mathcal{O}_{i+1})\Big\} \quad (8)$$

Obtaining the longest path in the DAG via DP takes $O(Mt_{max}^2)$, where t_{max} where t_{max} is the maximum number of object appearances in a frame of $\{\mathcal{F}_i\}$.

IV. EXPERIMENTS AND RESULTS

This section presents and analyzes experiments that demonstrate the advantages of the proposed multi-cue product detection approach. The experiments were conducted in a MacBook Pro with 4GB of RAM, 200GB of HDD, a dual core Intel(R) Core(TM) i7 CPU, and 512MB of video memory. The proposed approach was implemented using Matlab 2012a [26] and OpenCV 2.4.5 [27].

The evaluation was performed on fashion shows as they provide commercial value to a prototype TV shopping system. Three fashion show videos were obtained from YouTube [28] with high resolution. Figure 5 shows a few sample frames. Given the required manual labeling and evaluation efforts, a 2,074-frame clip was extracted from each video, for a total of 6,222 frames with a resolution of 576 × 324. The extracted clips from the videos are referred to as VC1, VC2, and VC3. The target product class consists of hand bags, which represent a significant detection challenge as described in the introduction, and the models' faces make up the related object class. Figure 2 shows bounding boxes for these two classes.

Color histograms were utilized as the feature representation $\tau(.)$ required in equation (1), with histogram intersection used as the similarity function $s(.)$. In addition, to detect occurrences of the target and related classes, the following object detectors were utilized:

- The widely utilized Viola-Jones object detector [29] was used to detect the models' faces. The implementation and trained models that are provided by OpenCV 2.4.5 were incorporated into the Matlab implementation of the proposed approach via the mexopencv [30] library, which provides Matlab mex functions for the OpenCV library.

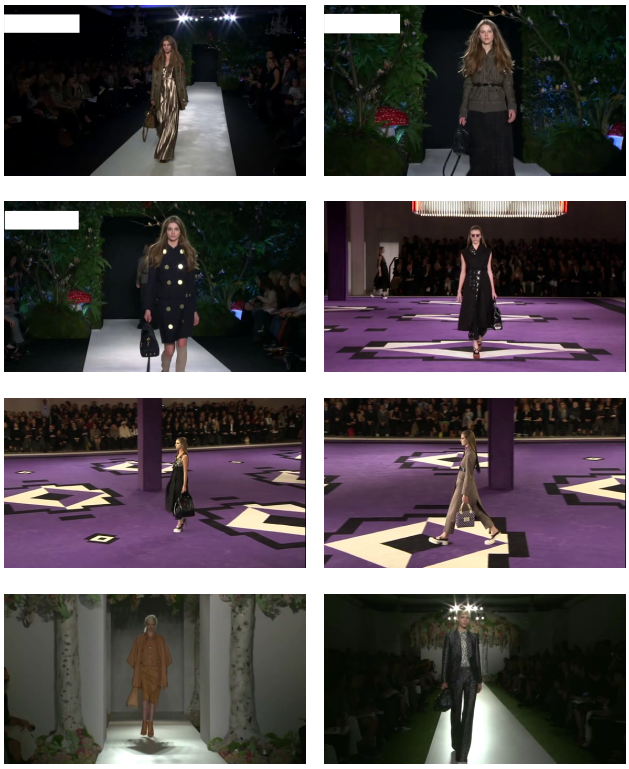


FIGURE 5. **Sample frames from test videos.**

- The discriminatively trained object detector based on deformable part models of Felzenszwalb et. al [1], [31] was utilized to detect hand bags. It represents variable object classes by using mixtures of deformable part models at different scales. This detector has achieved state-of-the-art results in the PASCAL object detection challenges [32], and its ability to detect non-rigid deformations and partial occlusions in the objects makes it a suitable approach for detecting hand bags. Moreover, a Matlab implementation of this approach is available online [31], which was directly incorporated into the proposed approach. To train this detector, 500 frames were extracted from the three fashion show videos (excluding the extracted clips). Out of the 500 frames, 250 were positive frames (i.e., contained hand bags), while the other 250 were negative frames (i.e., no hand bags). The bounding boxes for the hand bags in the positive set were manually labeled, thus creating the ground-truth set of bounding boxes.

The experiments consisted in comparing the detection performance of the proposed approach with three other approaches.

The first comparison is with the approach of Sheikh et al. [21], referred to as the surveillance-application approach. It is representative of the majority methods that use temporal information, which are meant for surveillance videos. Since they assume a fixed camera, these approaches are not suitable for fashion shows that are characterized by

many shots and varying background motion. Nevertheless, the comparison with the proposed approach is made to validate this claim. To make a proper comparison, the clips VC1, VC2, and VC3 were divided into shots, and the approach of Sheikh et al. was executed for each shot. The code was obtained from the project's web site.[1] It generates a detection mask for each frame, which was post-processed using morphological operators. The resulting regions were then represented by bounding boxes. For this and following comparisons, the proposed approach was executed with the weight parameters α, β, and γ of equation (1) set to $\frac{1}{4}$; i.e., it equally utilized all the equations derived from the three cues.

TABLE 1. Detection results for the surveillance-application approach.

	VC1	VC2	VC3
Precision	0.14	0.01	0.09
Recall	0.21	0.06	0.18

TABLE 2. Detection results for the proposed approach.

	VC1	VC2	VC3
Precision	0.94	0.92	0.97
Recall	0.38	0.54	0.31

Tables 1 and 2 show the detection results in terms of precision and recall values. Precision is defined as the fraction of predicted bounding boxes that are true positives, and recall is defined as the fraction of ground-truth bounding boxes that are predicted. A bounding box is considered true positive if it overlaps more than 50% of the ground-truth bounding box; otherwise, it is considered a false positive. Moreover, if multiple bounding boxes are predicted that overlap with the same ground-truth bounding box, only one is considered correct, and the other ones are labeled as false positives. Precision-recall curves were not generated as the approach of Sheikh et al. does not generate detection scores.

As shown in the tables, the proposed approach significantly outperforms the surveillance-application approach. For all three clips, the proposed approach achieved higher recall values as well as precision values over 0.90. On the other hand, the maximum values achieved by the surveillance-application approach were 0.14 precision and 0.21 recall. Such a difference in performance highlights the claim that approaches meant for surveillance videos are not appropriate for unconstrained video sequences. Additionally, some complex object classes such as hand bags are very difficult to detect using mainly motion estimation as these are likely to be immersed in the motion of the person carrying them.

The other comparisons are with the following three approaches, which assume nothing about background motion in the video sequences. The first one is a ''plain'' product detection approach, which does not use any additional cues and consists of performing object detection on each frame of the video sequence. The second approach is that of Heitz et al. [14] that uses a things-and-stuff (TAS) context model as described in section II. This approach is representative of methods that do not use temporal cues. They can be applied to the TV shopping use case by processing each frame of the video sequence individually. The TAS approach was chosen as its goal is in line with the proposed approach. That is, it is not tied to a specific detection approach but rather utilizes contextual information to enhance the detection results of object detectors. The code was obtained online,[2] the CEDD features [33] were used to represent the image regions this approach uses for context, and the code was trained using the same ground-truth bag data used to train object detector of Felzenszwalb et. al. Finally, the third approach consists of one sub-optimal version of the proposed approach. This version sets the weight parameters α and β to 0.5, thus effectively disabling both the topological and spatio-temporal cues to be able to analyze their combined effect. For each frame, each approach predicts bounding boxes of the hand bags along with corresponding detection scores. These scores were thresholded to then obtain precision-recall curves for each clip.

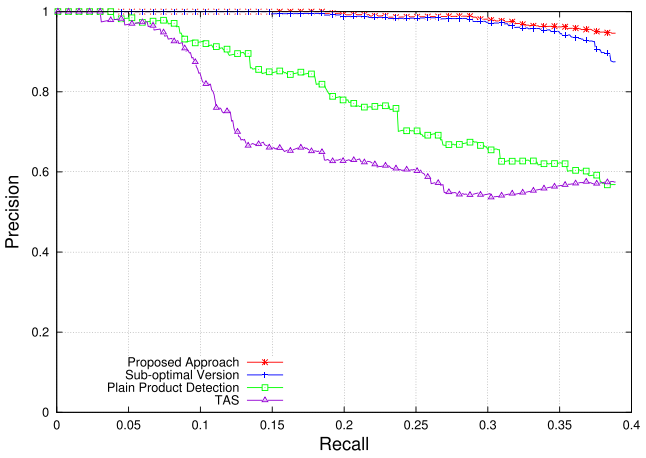


FIGURE 6. Precision-recall curve for VC1.

Figures 6, 7, and 8 depict the precision-recall curves for VC1, VC2, and VC3, respectively. As shown, the proposed approach significantly outperforms both the plain product detection and TAS detection approaches in terms of precision. The advantage is achieved by removing a significant number of false positive detections and thus improving the precision for the same recall value. For VC1, the maximum recall achieved is 0.38, at which the proposed approach obtains a precision of 0.94 vs 0.56 for the plain object detector (67.86% improvement) and 0.57 for the TAS approach (64.91% improvement). For VC2, the maximum recall achieved is 0.54, at which the proposed approach obtains a precision of 0.92 vs 0.43 for the plain object detection

[1] http://crcv.ucf.edu/projects/backgroundsub/

[2] http://ai.stanford.edu/~gaheitz/Research/TAS/

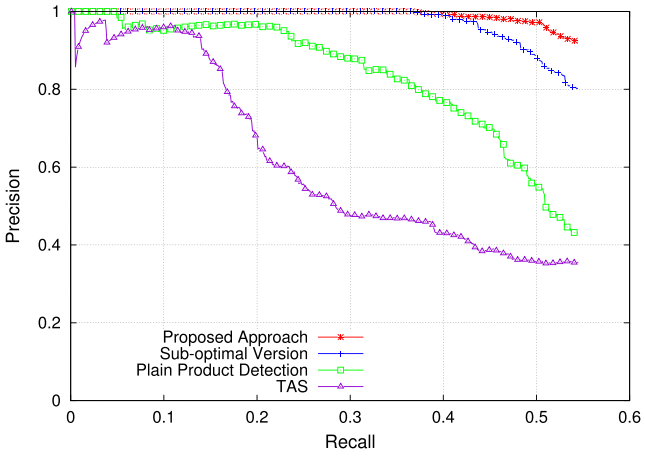


FIGURE 7. Precision-recall curve for VC2.

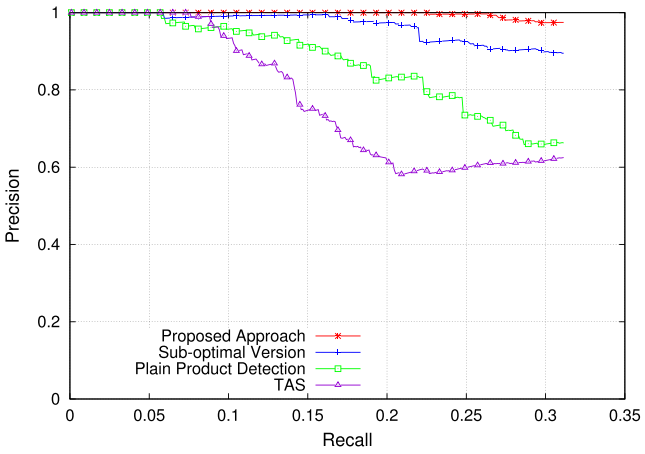


FIGURE 8. Precision-recall curve for VC3.

(113.56% improvement) and 0.36 for the TAS approach (155.56% improvement). For VC3, the maximum recall achieved is 0.31, at which the proposed approach obtains a precision of 0.97 vs 0.66, (47.07% improvement) and 0.63 for the TAS approach (53.97% improvement). Sample bounding boxes generated by the proposed and sub-optimal approaches in VC1 are depicted in figure 9. The sub-optimal detector generates both true positive (red bounding boxes) and false positive (yellow bounding boxes) detections due to the difficulty of detecting hand bags. On the other hand, the proposed approach is able to eliminate the false positive boxes. This result empirically proves the claim that the proposed approach can be used with a low detection threshold to increase the recall of complex objects but still achieve high precision values.

Figure 10 highlights the improvement in precision of the proposed approach compared to that of the sub-optimal version. As shown, the improvement is zero for low recall values but rises as the recall increases. A maximum improvement of 8.15% is achieved for VC1, 14.80% for VC2, and 8.92% for VC3. The reason for a larger improvement in VC2 is that more false positive detections were corrected

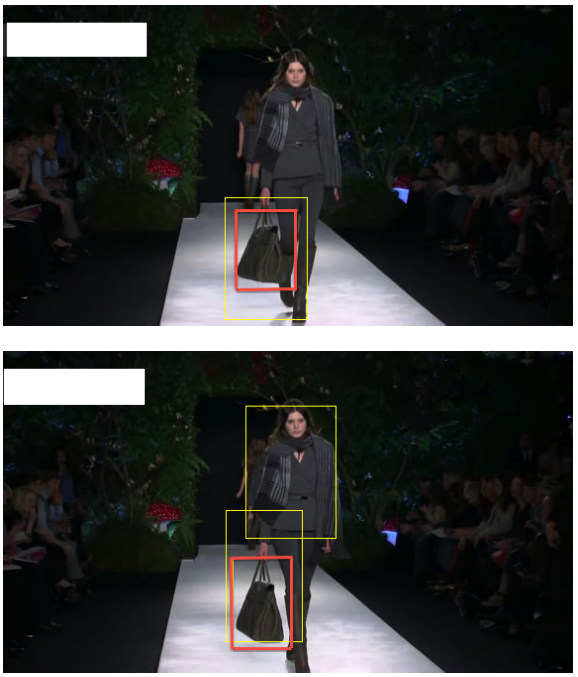


FIGURE 9. Examples of predicted bounding boxes for hand bags in clip VC1. The plain product detection approach predicts all bounding boxes (both yellow and red ones), whereas the proposed multi-cue approach only predicts the red boxes and eliminates the yellow boxes as false positives.

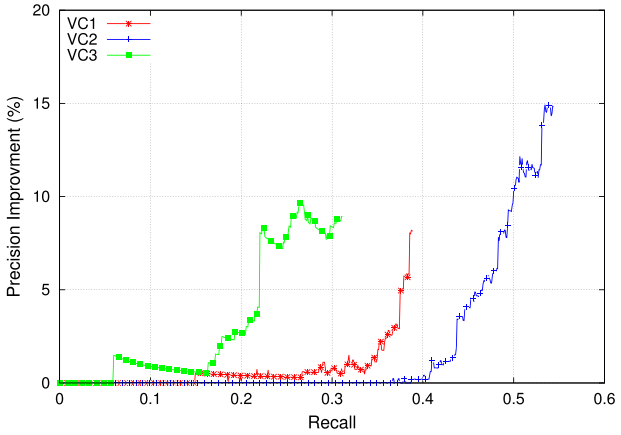


FIGURE 10. Precision improvement (%) of the proposed approach vs. the sub-optimal version.

by using the topological and spatio-temporal cues. Moreover, this result underlines the importance of utilizing the related object class in achieving higher precision at larger recall values.

Another aspect worth discussing is the improved performance of the plain detection approach over that of the TAS approach, in all three video clips. The TAS approach failed to correctly rescore a significant portion of the detections. An example of which is depicted in figure 11, where the red bounding box is the one that was scored the highest by the TAS approach in that frame. The explanation is that the image regions used by the TAS approach as context around the hand

FIGURE 11. Example of incorrect rescoring by TAS approach.

bags are very similar to those around other parts of the model, which makes it difficult for the TAS approach to correctly rescore the detections.

TABLE 3. Maximum recall achieved.

	VC1	VC2	VC3
Per Occurrence	0.38	0.58	0.34
Per Product	1.00	1.00	0.90

Moreover, it is important to highlight that the relatively low recall values reported in the precision-recall curves do not detriment the applicability of the proposed approach on fashion shows. These values can be considered to be on a ''per occurrence'' basis, i.e., computed for each true positive bounding box. However, for the purpose of TV shopping, the important criterion is to detect each particular bag at least once, and not necessarily on all the frames the bag appears on. For example, if a bag appears consecutively from frames 1 through 100 as the model walks through the stage, it can be considered a successful product detection if the bag is detected on a subset of the 100 frames, even if it is detected only once or twice. The rationale is that the consumer will still be notified of the existence of a possibly interesting bag. Hence, it can be stated that recall on a ''per product'' basis is more important. Table 3 compares the per-occurrence and per-product recall values for the three clips. It shows that on a per-product basis, the recall is very high, achieving 1.0 for both VC1 and VC2, and 0.9 for VC3. Precision, on the contrary, still should be considered on a per-occurrence basis as having to many false-positive occurrences has a negative impact on the usability of the system.

V. CONCLUSION AND FUTURE WORK

The lack of content understanding does not allow smart TVs to provide consumers with a seamless TV shopping experience. To purchase interesting items displayed in the current TV show, consumers must inconveniently resort to a store or the Web. Object detection is one of the tasks that is required for realizing the TV shopping use case, but the detection of complex objects poses a significant challenge. To this end, this article proposes a multi-cue product detection framework for TV shopping. Three main characteristics define the proposed approach. Firstly, it is generic in the sense that it is not tied to a specific object detection approach. Secondly, it does not make any assumption about motion in the video. Thirdly, it utilizes three cues as additional information to improve the detection results of a target product class. The appearance cue is related to the probability of a product occurrence of corresponding to the target class. The other two consists of topological and spatio-temporal relationships between the target product class and a related, easier-to-detect object class. These enforce spatial relationships within a video frame and across consecutive frames, respectively. The proposed approach jointly considers the three cues as a path-optimization problem that aims at selecting the correct product occurrences and weed out false positive detections. The empirical results demonstrate the advantages of the proposed framework in improving the detection results.

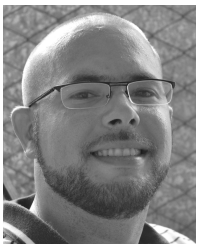
Future work comprises three aspects. The first one is extending the experimental results with three scenarios: (a) using another related object class, e.g., ''persons'' for which robust detectors exist in the literature; (b) handling other target product classes such as apparel, shoes, and/or watches, and (c) employing different detectors for the target product class. Since the proposed detection framework is independent of the detection mechanism utilized for the target product class, handling other product classes entails training detectors and plugging them into the framework. Accuracy results for these may vary according to the chosen detector, but the proposed framework is likely to enhance the performance via the additional information provided by the related object class. Moreover, analyzing the performance results produced by different detectors will yield important conclusions on the benefits provided by the related object class with respect to different accuracy levels obtained for the target product class. The second aspect of future work is the development of a distributed system that allows the real-time application of the proposed product detection framework. The computational performance of the proposed framework mainly depends on the computational cost of detecting the related objects and target products in each frame. Nevertheless, within a shot, frames can be processed in parallel, and the proposed optimization problem can be efficiently solved once all objects in the shot have been obtained. Hence, the detection phase can be distributed and parallelized to achieve the desired performance. The third one is the applicability of the framework in other TV shopping use cases such as on-demand movies. Such a use case involves the same functional requirements as that of fashion shows, except that on-demand movies can be batched-processed offline and the results saved for future retrieval.

REFERENCES

[1] P. F. Felzenszwalb, R. B. Girshick, D. McAllester, and D. Ramanan, ''Object detection with discriminatively trained part-based models,'' *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 32, no. 9, pp. 1627–1645, Sep. 2010.

[2] B. Leibe, A. Leonardis, and B. Schiele, ''Robust object detection with interleaved categorization and segmentation,'' *Int. J. Comput. Vis.*, vol. 77, nos. 1–3, pp. 259–289, May 2008. [Online]. Available: http://dx.doi.org/10.1007/s11263-007-0095-3

[3] J. Zhang, M. Marszalek, S. Lazebnik, and C. Schmid, "Local features and kernels for classification of texture and object categories: A comprehensive study," in *Proc. Conf. Comput. Vis. Pattern Recognit. Workshop (CVPRW)*, Jun. 2006, p. 13.
[4] P. Viola and M. Jones, "Rapid object detection using a boosted cascade of simple features," in *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit. (CVPR)*, vol. 1. Dec. 2001, pp. 511–518.
[5] M. Weber, M. Welling, and P. Perona, "Towards automatic discovery of object categories," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, vol. 2. Jun. 2000, pp. 101–108.
[6] A. Ayvaci and S. Soatto, "Detachable object detection: Segmentation and depth ordering from short-baseline video," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 34, no. 10, pp. 1942–1951, Oct. 2012. [Online]. Available: http://dx.doi.org/10.1109/TPAMI.2011.271
[7] Y. Gurwicz, R. Yehezkel, and B. Lachover, "Multiclass object classification for real-time video surveillance systems," *Pattern Recognit. Lett.*, vol. 32, no. 6, pp. 805–815, Apr. 2011. [Online]. Available: http://dx.doi.org/10.1016/j.patrec.2011.01.005
[8] D. Liu, M.-L. Shyu, Q. Zhu, and S.-C. Chen, "Moving object detection under object occlusion situations in video sequences," in *Proc. IEEE Int. Symp. Multimedia (ISM)*, Dec. 2011, pp. 271–278. [Online]. Available: http://dx.doi.org/10.1109/ISM.2011.50
[9] A. Vazquez-Reina, S. Avidan, H. Pfister, and E. Miller, "Multiple hypothesis video segmentation from superpixel flows," in *Proc. 11th Eur. Conf. Comput. Vis. (ECCV)*, 2010, pp. 268–281. [Online]. Available: http://dl.acm.org/citation.cfm?id=1888150.1888172
[10] J. Kim, G. Ye, and D. Kim, "Moving object detection under free-moving camera," in *Proc. 17th IEEE Int. Conf. Image Process. (ICIP)*, Sep. 2010, pp. 4669–4672.
[11] G. van Essen, S. Marsland, and J. Lewis, "Hierarchical block-based image registration for computing multiple image motions," in *Proc. 24th Int. Conf. Image Vision Comput. New Zealand (IVCNZ)*, Nov. 2009, pp. 425–430.
[12] B. Qi, M. Ghazal, and A. Amer, "Robust global motion estimation oriented to video object segmentation," *IEEE Trans. Image Process.*, vol. 17, no. 6, pp. 958–967, Jun. 2008.
[13] S. Kumar and M. Hebert, "A hierarchical field framework for unified context-based classification," in *Proc. 10th IEEE Int. Conf. Comput. Vis. (ICCV)*, vol. 2. Oct. 2005, pp. 1284–1291. [Online]. Available: http://dx.doi.org/10.1109/ICCV.2005.9
[14] G. Heitz and D. Koller, "Learning spatial context: Using stuff to find things," in *Proc. 10th Eur. Conf. Comput. Vis. (ECCV)*, 2008, pp. 30–43. [Online]. Available: http://dx.doi.org/10.1007/978-3-540-88682-2_4
[15] C. Galleguillos, A. Rabinovich, and S. Belongie, "Object categorization using co-occurrence, location and appearance," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2008, pp. 1–8.
[16] C. Galleguillos and S. Belongie, "Context based object categorization: A critical survey," *Comput. Vis. Image Understand.*, vol. 114, no. 6, pp. 712–722, Jun. 2010. [Online]. Available: http://dx.doi.org/10.1016/j.cviu.2010.02.004
[17] A. Torralba, K. P. Murphy, and W. T. Freeman, "Using the forest to see the trees: exploiting context for visual object detection and localization," *Commun. ACM*, vol. 53, no. 3, pp. 107–114, Mar. 2010. [Online]. Available: http://doi.acm.org/10.1145/1666420.1666446
[18] P. F. Felzenszwalb, R. B. Girshick, D. McAllester, and D. Ramanan, "Object detection with discriminatively trained part-based models," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 32, no. 9, pp. 1627–1645, Sep. 2010.
[19] L. Wang, Y. Wu, T. Lu, and K. Chen, "Multiclass object detection by combining local appearances and context," in *Proc. 19th ACM Int. Conf. Multimedia (MM)*, 2011, pp. 1161–1164. [Online]. Available: http://doi.acm.org/10.1145/2072298.2071964
[20] W.-S. Zheng, S. Gong, and T. Xiang, "Quantifying and transferring contextual information in object detection," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 34, no. 4, pp. 762–777, Apr. 2012.
[21] Y. Sheikh and M. Shah, "Bayesian modeling of dynamic scenes for object detection," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 27, no. 11, pp. 1778–1792, Nov. 2005.
[22] R. Yan, J. Zhang, J. Yang, and A. G. Hauptmann, "A discriminative learning framework with pairwise constraints for video object classification," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 28, no. 4, pp. 578–593, Apr. 2006.
[23] M. Yang, Y. Wu, and G. Hua, "Context-aware visual tracking," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 31, no. 7, pp. 1195–1209, Jul. 2009.
[24] A. Torralba, "Contextual priming for object detection," *Int. J. Comput. Vis.*, vol. 53, no. 2, pp. 169–191, Jul. 2003. [Online]. Available: http://dx.doi.org/10.1023/A:1023052124951
[25] P. Carbonetto, N. de Freitas, and K. Barnard, "A statistical model for general contextual object recognition," in *Proc. 8th Eur. Conf. Comput. Vis.*, 2004, pp. 350–362.
[26] MathWorks. (Feb. 2014). *Mathworks Announces Release 2012a of the MATLAB and Simulink Product Families*. [Online]. Available: http://www.mathworks.com / company / newsroom / MathWorks-Announces-Release-2012a-of-the-MATLAB-and-Simulink-Product-Families.html
[27] OpenCV. (Feb. 2014). [Online]. Available: http://opencv.org,
[28] Youtube. (Mar. 2014). [Online]. Available: http://www.youtube.com/
[29] P. Viola and M. Jones, "Rapid object detection using a boosted cascade of simple features," in *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit. (CVPR)*, vol. 1. Dec. 2001, pp. I-511–I-518.
[30] K. Yamaguchi. *Mexopencv*. [Online]. Available: http://www.cs.stonybrook.edu/~kyamagu/mexopencv/, accessed Jun. 11, 2014.
[31] P. F. Felzenszwalb, R. B. Girshick, and D. McAllester. *Discriminatively Trained Deformable Part Models, Release 4*. [Online]. Available: http://people.cs.uchicago.edu/~pff/latent-release4/, accessed Jun. 11, 2014.
[32] M. Everingham, L. V. Gool, C. K. I. Williams, J. Winn, and A. Zisserman. *The PASCAL Visual Object Classes Challenge 2009 (VOC2009) Results*. [Online]. Available: http://www.pascal-network.org/challenges/VOC/voc2009/workshop/index.html, accessed Jun. 11, 2014.
[33] S. A. Chatzichristofis and Y. S. Boutalis, "CEDD: Color and edge directivity descriptor: A compact descriptor for image indexing and retrieval," in *Proc. 6th Int. Conf. Comput. Vis. Syst. (ICVS)*, 2008, pp. 312–322.



FAUSTO C. FLEITES received the B.S., M.S., and Ph.D. degrees in computer science from Florida International University, Miami, FL, USA, in 2009, 2012, and 2014, respectively. His research interests are in multimedia mining and indexing and big data. He was an Intern with TCL Research America, Santa Clara, CA, USA, where he was involved in research on large-scale object detection and mining.

HAOHONG WANG is currently the General Manager of TCL Research America, Santa Clara, CA, USA, the North America's research arm of TCL Corporation. Prior to joining TCL Research America, he held various technical and management positions with AT&T, Dallas, TX, USA, Catapult Communications, Doylestown, PA, USA, Qualcomm, San Diego, CA, USA, Marvell, Hamilton, Bermuda, and Cisco, San Jose, CA, USA. His research involves in the areas of multimedia computing and communications, data mining, and recommender systems. He has authored over 50 articles in peer-reviewed journals and international conferences. He is the inventor of more than 50 patents and pending applications. He has co-authored the books entitled *3-D Visual Communications* (John Wiley & Sons, 2013), *4G Wireless Video Communications* (John Wiley & Sons, 2009), and *Computer Graphics* (1997). He received the Ph.D. degree from Northwestern University, Evanston, IL, USA.

Dr. Wang is the Editor-in-Chief of the *Journal of Communications*, the Vice President of the Asia-Pacific Signal and Information Processing Association, the Co-Chair of the IEEE Systems, Man, and Cybernetics Society Technical Committee on Human Perception and Multimedia Computing, and the Chair of the Steering Committee of the International Conference on Computing, Networking and Communications. He served as the General Chair of the IEEE ICME 2011 (Barcelona), the IEEE ICCCN 2011 (Maui), and the IEEE ICCCN 2008 (US Virgin Island), and the Technical Program Committee Chair of the IEEE GLOBECOME 2010 (Miami). He chaired the IEEE Multimedia Communications Technical Committee (2010–2012). He was a member of the Steering Committee of the IEEE TRANSACTION ON MULTIMEDIA (2010–2013) and the IEEE ICME conference (2010–2012), and an Editor of many journals. He has received the Distinguished Service Award by the IEEE ComSoc MMTC in 2013.

SHU-CHING CHEN has been a Full Professor with the School of Computing and Information Sciences, Florida International University (FIU), Miami, FL, USA, since 2009, where he has been an Assistant/Associate Professor since 1999. He received the Ph.D. degree in electrical and computer engineering and the master's degrees in computer science, electrical engineering, and civil engineering from Purdue University, West Lafayette, IN, USA, in 1998, 1992, 1995, and 1996, respectively.

He is currently the Director of the Distributed Multimedia Information Systems Laboratory with the School of Computing and Information Sciences. His main research interests include content-based image/video retrieval, distributed multimedia database management systems, multimedia data mining, multimedia systems, and disaster information management. He has authored or co-authored over 280 research papers in journals, refereed conference/symposium/workshop proceedings, book chapters, and four books.

Dr. Chen was a recipient of the ACM Distinguished Scientist Award in 2011 and the best paper award from the IEEE International Symposium on Multimedia in 2006. He was a recipient of the IEEE Systems, Man, and Cybernetics (SMC) Society's Outstanding Contribution Award in 2005 and a co-recipient of the IEEE Most Active SMC Technical Committee Award in 2006. He was also a recipient of the Inaugural Excellence in Graduate Mentorship Award from FIU in 2006, the University Outstanding Faculty Research Award from FIU in 2004, the Excellence in Mentorship Award from the School of Computing and Information Sciences, FIU, in 2010, the Outstanding Faculty Service Award from the School of Computing and Information Sciences, FIU, in 2004, and the Outstanding Faculty Research Award from the School of Computing and Information Sciences, FIU, in 2002 and 2012. He is a fellow of the Society of Information Reuse and Integration.

He has been the General Chair and Program Chair for more than 40 conferences, symposiums, and workshops. He is the Founding Editor-in-Chief of the *International Journal of Multimedia Data Engineering and Management* and an Associate Editor/Editorial Board Member for other 20 journals. He is the Chair of the IEEE Computer Society Technical Committee on Multimedia Computing and the Co-Chair of the IEEE SMC Society's Technical Committee on Knowledge Acquisition in Intelligent Systems. He has been a Guest Editor for more than 10 journal special issues. He was a member of three steering committees (including, the IEEE TRANSACTIONS ON MULTIMEDIA) and several panels for conferences and NSF. He served as a member of the Technical Program Committee for more than 320 professional meetings.