Mutual Enhancement for Detection of Multiple Logos in Sports Videos

Yuan Liao[1], Xiaoqing Lu[1], Chengcui Zhang[2], Yongtao Wang[1], Zhi Tang[1]
[1]Institute of Computer Science and Technology, Peking University, Beijing, China
[2]Department of Computer Science, University of Alabama at Birmingham, USA
{liao_yuan, lvxiaoqing, wangyongtao, tangzhi}@pku.edu.cn, czhang02@uab.edu

Abstract

Detecting logo frequency and duration in sports videos provides sponsors an effective way to evaluate their advertising efforts. However, general-purposed object detection methods cannot address all the challenges in sports videos. In this paper, we propose a mutual-enhanced approach that can improve the detection of a logo through the information obtained from other simultaneously occurred logos. In a Fast-RCNN-based framework, we first introduce a homogeneity-enhanced re-ranking method by analyzing the characteristics of homogeneous logos in each frame, including type repetition, color consistency, and mutual exclusion. Different from conventional enhance mechanism that improves the weak proposals with the dominant proposals, our mutual method can also enhance the relatively significant proposals with weak proposals. Mutual enhancement is also included in our frame propagation mechanism that improves logo detection by utilizing the continuity of logos across frames. We use a tennis video dataset and an associated logo collection for detection evaluation. Experiments show that the proposed method outperforms existing methods with a higher accuracy.

1. Introduction

Logo detection in images and videos has been gaining considerable attention in the last decade, with many applications such as traffic-control systems and measurement of brand exposure. Detecting and classifying logos in videos are valuable for business analysis, especially for television advertisers. The frequency and duration of logos allow sponsors to evaluate the effect of their advertising efforts.

Detection requires localizing objects within an image. Existing researches on logo detection and recognition have made great achievements in still image [1, 2, 20, 23, 28]. Most logo databases adopted by retrieval systems consist of still images only [13, 22, 23, 26]. The logos in still images are generally clear and likely to be captured from a front facing view. However in videos, due to zooming, panning and changing in camera exposures, the chrominance level changes frequently, and the logos are often deformed and blurred. Consequently, the logo detection for videos faces many more challenges.

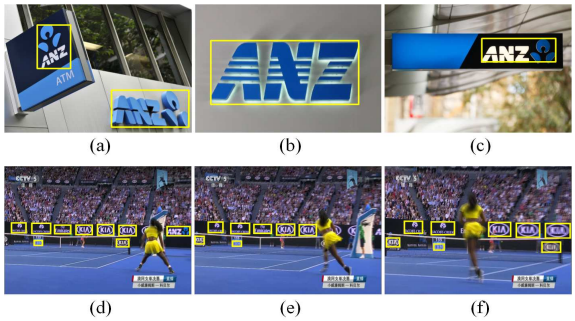

Figure 1. Challenges in logo detection. The logos from the same sponsor, (a), (b), and (c) have different layouts and deformation in tennis videos. The logos in neighboring frames (d), (e), and (f), suffer from motion blur and partial occlusion.

Logo detection could be considered as a sub problem of object detection. The existing approaches [3, 7, 15] generally consist of two important modules: detecting candidate object regions and classifying those regions. Besides the common challenges in object detection in videos, the detection of sponsor logos faces at least three specific obstacles. First, as shown in Figures 1(a), (b) and (c), according to the requirement of sponsors, the same sponsor's logos may exhibit different layouts and thus cause higher inner-cluster differences than ordinary objects; second, multiple instances of the same logo sometimes appear simultaneously but with various visual qualities, as shown in Figures 1(d), (e), and (f). Traditional methods can hardly detect all of them simultaneously. Furthermore, as a result of the panning and zooming of the camera, consecutive frames can produce different degrees of fuzziness, and the detection in fuzzy frames can be eased by utilizing relevant information from adjacent frames. To overcome the above-mentioned obstacles, we propose a logo detection framework based on a mutual enhancement mechanism aiming

to improve the detection of a logo leveraging the information obtained from other simultaneously occurring logos. In a Fast-RCNN-based framework, we introduce a homogeneity-enhanced re-ranking method by analyzing the characteristics of logos in videos to improve the region proposal accuracy in frames, including the type repetition of appearance, color consistency and mutual exclusion. As our most import contribution, this technique can be applied to other kinds of videos that contain multiple homogenous objects appearing simultaneously but with various visual qualities. Different from the conventional enhance mechanism that improves the weak proposals with the dominant proposals, our method enables all the proposals in one frame to mutually enhance each other, including improving those relatively significant or non-obvious proposals with weak proposals by their common characteristics and the potential alignment information. Moreover, a frame propagation enhancement method is also presented to assist the detection in contiguous frames.

The remainder of this paper is structured as follows. Section 2 introduces the related works of logo recognition and detection. In Section 3, the framework of the proposed method is presented. The homogeneity-enhanced re-ranking is detailed in Section 4. The propagating-enhancement and control mechanism is introduced in Section 5. Section 6 introduces the dataset used in this work, and presents the experimental results. Section 7 concludes this paper.

2. Related Works

Previous works generally establish detection models with key-point representations commendably capturing specific patterns present in graphic logos, such as key-point-based detectors and bag-of-words models [14, 22, 23, 24, 30, 31, 33, 34]. These methods extract local features such as SIFT [17] or HoG [4] from images, cluster and quantize these features into visual words and finally measure the similarity between a test image and a logo image according to these visual words. For instance, Romberg *et al*. [24] proposed a shape representation built with found key-points and their respective SIFT representation for scalable logo recognition in images. In [23], bundles of SIFT features are built from local regions around each key-point to index specific graphic logo patterns. In [28], an improved topological constraint, which considers the relative position between a key-point and its neighbor points, is proposed to reduce the number of mismatched key-points.

In recent years, deep learning has shown its good performance in object detection. State-of-the-art deep learning-based object detection networks such as R-CNN [10] depend on region proposal algorithms to hypothesize object locations. Further improvements Fast R-CNN [9] have reduced the running time of these detection networks, exposing region proposal computation as a bottleneck. Region Proposal Network (RPN) [21] shares full-image convolutional features with the detection network, thus enabling nearly cost-free region proposals. There are also a lot of attempts to apply deep learning in logo detection, and obtain excellent results [1, 12, 20]. Bianco *et al*. [1] involves the selection of candidate subwindows using an unsupervised segmentation algorithm, and the SVM-based classification of such candidate regions uses features computed by a CNN. Oliveira *et al*. [20] adopts the transfer learning to leverage powerful Convolutional Neural Network models trained with large-scale datasets and repurposes them in the context of graphic logo detection.

Video object tracking is another task close to video logo detection. Object tracking (especially single object tracking) refers to tasks that estimate the object state in subsequent frames with a given initial object state in the first frame. Tracking-by-detection methods gradually become the main-stream in the field of object tracking because of its outstanding performance [5, 6, 11, 18, 32]. Wang and Yeung [32] propose a framework of offline training and online fine-tuning, which largely solved the problem of insufficient training samples. While the majority of tracking targets are foreground, logos in sports videos generally belong to background. Therefore, it is difficult to distinguish logos from other background by movement analysis. Besides, the logos drift in and out of the camera frequently, which makes the detection of logos in sport videos more challenging than typical tracking tasks.

Some previous efforts attempt to recognize logos in videos [3, 7, 15]. Richard *et al*. [7] propose string based template matching to recognize logos in video stills. Chattopadhyay and Sinha [3] propose a system to automatically recognize the logos from sports videos for channel hyperlinking in the client end.

Despite the past efforts in video logo detection, the overall accuracy is still far from satisfactory. Different types of sports videos have their own unique characteristics. For example, in soccer videos, logos may look small and blurry in high-angle bleacher shots, while in tennis videos, logos may appear large in size but more likely to be occluded. (Figure 1(f)). Rich interframe information in videos has also been underutilized. In this paper, we focus on the characteristics in sports videos, utilize the mutual information in tennis videos and the enhanced deep learning to detect trademarks in videos.

3. The Proposed Framework

Deep Convolution Neural Networks (CNNs) are increasingly used in objection detection. A typical CNN-based framework for logo detection [20] generally consists of video preprocessing, candidate logo proposition and CNN-based classification modules. In this paper, the proposed

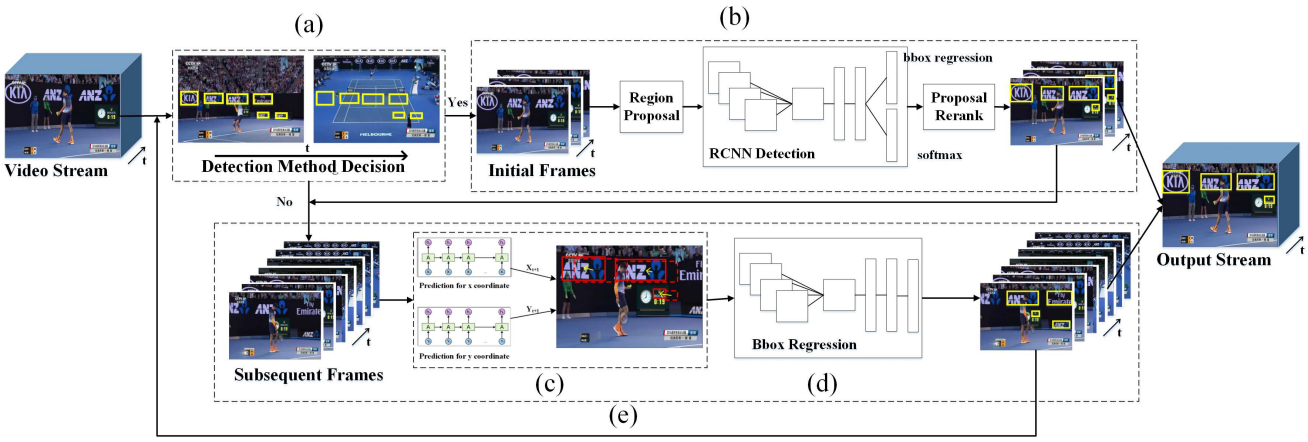

Figure 2. Video logo detection framework. The proposed video logo detection framework contains two major modules: proposal re-ranking and frame propagation. (a) Detection method decision. If the similarity between two frames is less than the threshold, we regard the coming frame as a shot boundary, then (b) is performed, otherwise (e) is performed. (b) When a shot boundary is detected, a few frames at the beginning of the shot are processed by region proposal and RCNN. The proposal rerank is operated to adjust the ranking of detection results. Detection results of these frames are utilized to propagate the subsequent frames. (c) Motion estimation module which predicts the next location of a logo using LSTM. The dotted red boxes are the logo location in the previous frame, and the solid red boxes are the location predicted. The yellow arrows show the translation of centers. (d) Bounding box regression is utilized to adjust that of each predicted logo. (e) The process of propagation.

framework for logo detection in sports videos is different from the conventional approaches. As shown in Figure 2, two phases, mutual-enhancement-based logo proposal re-ranking in frames and motion-estimation-based propagation in video clips, are introduced in our framework. The re-ranking method analyzes the characteristic of homogeneous logos within the frame to improve the accuracy of logo proposals, and the motion-estimation-based propagation enhances the detection in contiguous frames. The workflow of the method is described briefly as follows:

At the beginning of a shot, we detect logos in the first several frames with Fast-RCNN and mutual-enhancement-based re-ranking. Then we collect the detection results and utilize them to initialize the relay potency for each detected potential logo region. With the locations and the information of potential logos in previous frames obtained, we perform motion-estimation-based propagation to predict the new positions of potential logos in the current frame. A similarity verification step is then performed to select those qualified prediction results as the detection results of the current frame, and the relay potencies are updated for these results. When the relay potency of a potential logo decreases to 0, the propagation of this logo will be discontinued. During this process, we keep comparing the difference between the previous frame and the current one. If the magnitude of change exceeds a relatively conservative threshold, the current propagation will be canceled, and the detection with Fast-RCNN will be restarted from the current frame, just as that for the start of a shot.

In this study, we adopt the Fast-RCNN-based object detection model introduced by Ross Girshick [9] to generate logo candidates. First, a region proposal algorithm [29] is used to select category-independent region proposals for further classification. Then feature extraction is performed using a CNN for each proposed region. These regions are then classified using a softmax classifier with fully connected layers. After classification, each region is assigned a confidence score. However, a fixed confidence threshold cannot accommodate the diversity of visual quality in a video. To mutually enhance all possible proposals, we screen the proposed logos with a re-ranking method by analyzing the homogeneous logos occurring simultaneously, including the type repetition of appearance, color consistency and mutual exclusion (more details are provided in Section 4).

When the location of a logo is obtained in one frame, the corresponding location of its repetition in the next frame can be predicted utilizing the continuity between adjacent frames [16, 19, 25, 35]. In this paper, the motion-estimation-based propagation enhances the detection in contiguous frames, which is described in Section 5.

4. Mutual Enhancement for Proposals

Fast-RCNN has shown excellent performance in general-purposed object detection. However, it suffers from low recall in logo detection [1, 12, 20]. One of the primary reasons is the camera panning, which blurs logo regions and leads to low confidences. In this section, we present a filtering phase that re-ranks the proposals according to their

local context, including the type repetition of appearance, color consistency and mutual exclusion.

In a frame set K generated from a video clip, for the $i\text{-}th$ frame image $F_i \in K$, we utilize the Fast-RCNN-based object detection model to generate logo proposals in it, and then bag proposals as a candidate set P_i,

$$P_i = \{P_i^1, P_i^2, \ldots, P_i^j, \ldots, P_i^{n_i}\}$$

where P_i^j is the $j\text{-}th$ logo proposal of frame F_i, and n_i is the total number of proposals in frame F_i. The corresponding confidence of proposal P_i^j is defined as Φ_i^j, and the category of proposal P_i^j is $C_i^j \in \{1, 2, \ldots, m\}$, where m is the number of categories.

To improve the accuracy of proposal, we notice that the confidence Φ_i^j could be adjusted on the basis of context analysis. First, it is common in sport videos that several instances of the same logo appear simultaneously but with different visual quality, and the low confidence of a proposal can be promoted when more proposals of the same type have been detected. We denote this factor as ρ_i^j, which will be explained in Subsection 4.1. Second, different from randomly-appearing objects, such as pedestrians and vehicles, the logos in sports videos seldom overlap each other. We adopt an aggressive strategy to punish the overlapped proposals in Subsection 4.2, and define this factor as τ_i^j. As a consequence, the adjustment of Φ_i^j is estimated based on the homogeneous promotion and inhomogeneous exclusion, as follows:

$$\widehat{\Phi}_i^j = \Phi_i^j + \boldsymbol{W} \cdot [\rho_i^j, \tau_i^j]^{\mathrm{T}} \tag{1}$$

where $\boldsymbol{W} = [w_1, w_2]$ are the weight parameters, and $\widehat{\Phi}_i^j$ is the updated value of confidence.

4.1. Proposal Promotion with Homogeneous Siblings

As shown in Figure 1, it is common in sport videos that several instances of the same logo appear simultaneously, but their visual qualities may vary due to their individual perspective or distance to the camera. In a general Fast-RCNN framework, the decision about whether a region contains an object depends on one or more threshold. In our framework, a sibling-similarity-based strategy is proposed to improve the proposal selection. The proposals with high confidence are firstly chosen as delegates. And then more proposals in the same category of the delegate can obtain eligibility based on their co-appearance with a delegate. Such promotion also takes into account other characteristic similarities among the candidate regions in addition to the class information. Last but not the least, the layout information of the qualified proposals can reversely enhance all the proposals including the delegates according to the regularity degree of their spatial arrangement.

We denote the delegates as the most salient logo proposals among a group of logo proposals that appear simultaneously in a frame. They are selected with a relatively high confidence. This step is implemented by introducing a sign indicator S_i^j for each proposal P_i^j with the following formulas:

$$S_i^j = \begin{cases} 1, & \Phi_i^j \geq \delta_h \\ 0, & otherwise \end{cases} \qquad \delta_h = \theta + \varepsilon \tag{2}$$

where, θ is the experimental threshold in conventional network, and the proposals with $S_i^j = 1$ are chosen as delegates. ε is a two-way correction for the threshold adjustment, which is discussed in Section 6.2.

For each category of proposals in the current frame, we choose the maximum confidence among the delegates in the category as a recommendation factor $\widetilde{\Phi}$:

$$\widetilde{\Phi}_i^j = \max\{\Phi_i^k | (S_i^k > 0) \wedge (C_i^k = C_i^j)\} \qquad k \in \{1, 2, \ldots, n_i\} \tag{3}$$

We utilize the delegates and their category information to promote the detection of weak proposals and to remove unrelated proposals. First of all, a dynamic threshold is defined as:

$$\delta_{l,C_i^j} = \theta - \varepsilon \times exp(\widetilde{\Phi}_i^j - 1) \tag{4}$$

The relatively low threshold δ_{l,C_i^j} of the category C_i^j enables the maximum inclusion of the logo proposals of the same category. The adjustment $\varepsilon \times exp(\widetilde{\Phi}_i^j - 1)$ of the threshold depends on the confidence of the most significant delegate.

In many situations, the logos that could not be detected, even when other homogeneous logos in the same frame are obtained, are occluded by players or become blurred due to camera movement. To enhance the identification of these weak candidates, other characteristic similarities between the available regions of those and the delegates should be taken into account. In this paper, we adopt the color consistency to implement further enhancement, because it remains relatively stable in the cases of a slightly occluded logo and a blurred logo. Moreover, many instances in the same logo category can have different layouts, but they still have a relatively high consistency of color. For each class of logos, we smooth each training sample by Rolling Guidance Filter [36] to extract dominant color information, calculate the color histogram of the smoothed training sample, and compute the average histogram as the class template $ColorHistTemplate_{Class}$. For each logo proposal, we compare its color histogram with its corresponding class template. We denote the color consistency λ_i^j as

$$\lambda_i^j = \frac{\sum_I (H_1 - \bar{H_1})(H_2 - \bar{H_2})}{\sqrt{\sum_I (H_1 - \bar{H_1})^2 \sum_I (H_2 - \bar{H_2})^2}} \tag{5}$$

where H_1 is the color histogram of P_i^j, H_2 is the color histogram template of C_i^j, and

$$\bar{H_k} = \frac{1}{N} \sum\nolimits_j H_k(j)$$

N is the number of bins in color histogram, and λ_i^j is the correlation similarity of color histograms between a logo proposal P_i^j and the corresponding class template $ColorHistTemplate_{C_i^j}$. Combining the two factors, sibling co-occurrence and color consistency, we can update the criteria for promotion of logo proposals as:

$$\widetilde{\rho}_i^j = S_i^j \times \exp(\frac{\Phi_i^j - \delta_{l,C_i^j}}{\widetilde{\Phi}_i^j}) - \alpha \times \lambda_i^j \tag{6}$$

At last, a largely ignored characteristic when multiple logos appear in the same frame, i.e., the regularity of their spatial arrangement, can be utilized to verify or enhance all the proposals. In a typical sports video, the relative positions among the logos occurring simultaneously are not random, such as aligned in one direction. In this paper, we extract the local alignment information between two adjacent proposals to adjust their confidence. Suppose P_i^k is the nearest neighbor of the current proposal P_i^j within a predefined radius of neighborhood, the alinement σ between them can be estimated with the following formula.

$$\sigma = \frac{min(|\triangle x|, |\triangle y|)}{max(d1, d2)} \tag{7}$$

where, $\triangle x$ and $\triangle y$ are the coordinate differences between the two proposal centers. $d1$ and $d2$ are the diagonals of P_i^j and P_i^k, respectively. If σ is small enough, the two proposals can be considered aligned in one direction. The adjustment for the current proposal can be expressed as:

$$\rho_i^j = \begin{cases} \widetilde{\rho}_i^j \times exp(1-\sigma), & \sigma < \xi \\ \widetilde{\rho}_i^j, & otherwise \end{cases} \tag{8}$$

where ξ is a threshold for judging whether an alignment exists. The spatial-relationship-based enhancement is mutual and can apply to all the proposals, including improving the delegates with the weak proposals.

4.2. Exclusion with Inhomogeneous Siblings

Unlike typical objects in videos, such as pedestrians and vehicles, logo instances in sports videos rarely overlap each other. The observation motivates us to punish inhomogeneous overlapped proposals. In our approach, when two or more proposals are overlapped, the proposal with the highest confidence is selected as the dominate proposal. Simultaneously, the other overlapped proposals in the same region are severely punished. The penalty of an individual proposal is commensurate with the degree of overlapping according to the following formulas.

$$\tau_i^j = \begin{cases} 0 & \textit{if } C_i^j = C_i^k \\ 1 - e^{OReg(P_i^j, P_i^k)} & otherwise \end{cases} \tag{9}$$

$$\textit{where } OReg(P_i^j, P_i^k) = \frac{P_i^j \cap P_i^k}{P_i^j \cup P_i^k}$$

$OReg(P_i^j, P_i^k)$ indicates the overlap degree between the logo proposal P_i^j and the dominant proposal P_i^k. As overlap increases, the penalty increases rapidly.

5. Mutual Enhancement via Frame Propagation

It is more efficient to predict the logo positions on the basis of temporal correlation than to detect them on every frame individually, since most logos in the same shot may have small variation between adjacent frames. Meanwhile, the detection in fuzzy frames could be assisted by their adjacent frames. Several works have been proposed to dig and utilize the relationships between frames [5, 6, 11, 18]. In our framework, we first simply use long short-term memory (LSTM) network to predict the location in coming frames, then take advantages of the time continuity to enhance the logo detection with a propagation mechanism.

5.1. Motion-Estimation-based Prediction

Motion estimation is one of the widely used methods for object detection and tracking in the field of video retrieval and surveillance [27]. When we detect and collect the logo proposals in previous frames, we also record the center $\{x_i^j, y_i^j\}$ of each proposal P_i^j. Let the center sequence of P^j proposal be:

$$X^j = \{x_1^j, x_2^j, \ldots, x_i^j\}, \quad Y^j = \{y_1^j, y_2^j, \ldots, y_i^j\}$$

We train a LSTM on the set of center sequences for predicting the center $\{x_{i+1}^j, y_{i+1}^j\}$ of P_{i+1}^j. Then, a region which centers at $\{x_{i+1}^j, y_{i+1}^j\}$ and has the same size as P_i^j is regarded as P_{i+1}^j.

To verify the reliability of the propagated P_{i+1}^j, we calculate the similarity between P_i^j and P_{i+1}^j based on the mean absolute deviation (MAD) cost. Suppose B is one of the proposal blocks and B' is the corresponding block in the next frame. The MAD value between them is calculated as:

$$MAD(B, B') = \frac{1}{w \times h} \sum_{x=1}^{w} \sum_{y=1}^{h} |B(x,y) - B'(x,y)| \tag{10}$$

where w and h are the width and height of the block region, respectively, $B(x,y)$ and $B'(x,y)$ are the pixels in the current block and the referenced block, respectively.

The distance between P_i^j and estimated P_{i+1}^j can also be calculated by MAD.

$$D_i^j = MAD(P_i^j, P_{i+1}^j) \tag{11}$$

and the similarity of two adjacent frames is defined as:

$$\Psi(F_i, F_{i+1}) = \frac{1}{n_i} \sum_{j=1}^{n_i} (1 - D_i^j) \tag{12}$$

where n_i is the total number of proposals in frame F_i.

At the end of each frame propagation, a bounding box regression is adopted to adjust P_{i+1}^j.

5.2. Control Mechanism for Propagation

Before a frame is processed, a control mechanism is adopted to make sure a right decision is made to adopt either the propagation or the detection with Fast-RCNN.

First, different than traditional methods that firstly divide a video into shots, our method automatically decides whether to perform detection by Fast-RCNN or by propagation based on the frame continuity. Once the propagation starts, the proposed method keeps on comparing the difference between the current frame and the next one. The frame similarity $\Psi(F_i, F_{i+1})$ is compared with a relatively high threshold θ_s. When $\Psi(F_i, F_{i+1}) > \theta_s$, the two adjacent frames are considered to be in the same shot and the propagation can continue. If $\Psi(F_i, F_{i+1}) \leq \theta_s$, the propagation will be discontinued, and the detection with Fast-RCNN will be restarted from the next frame. The significant camera changes, such as shot switch, fast panning and zooming, may lead to a low $\Psi(F_i, F_{i+1})$ and a Fast-RCNN will be adopted to avoid problematic propagation.

Moreover, a group of relay potency, R_i, is introduced to describe the propagation ability of proposals in P_i.

$$R_i = \{R_i^1, R_i^2, \ldots, R_i^j, \ldots, R_i^{n_i}\}$$

where, i denotes the i-th frame, j denotes the j-th proposal in F_i. R_i^j is initialized according to Φ_i^j when the propagation procedure starts or re-starts. During the propagation, the relay potency decreases gradually. On each iteration of the propagation, R_i^j is transferred to the new proposal in the next frame, but part of its energy is lost according to the dissimilarity between the corresponding proposals in adjacent frames.

$$R_{i+1}^j = R_i^j - D_i^j \tag{13}$$

When the relay potency drops to zero, the corresponding proposal also has to withdraw from propagation. When all the proposals run out of potency, the propagation will terminate and the detection with Fast-RCNN will restart. Obviously, significant camera changes lead to fast attenuation of relay potency, which also avoids problematic propagation.

6. Experiments

In our experiments, we firstly evaluate the proposed method in commonly used logo recognition image datasets, and then evaluate the logo detection method in a sports video dataset.

6.1. Dataset

We evaluate the proposed re-ranking method in two challenging and commonly used logo recognition datasets, FlickrLogos-32[24] and FlickrLogos-27[14], both of which

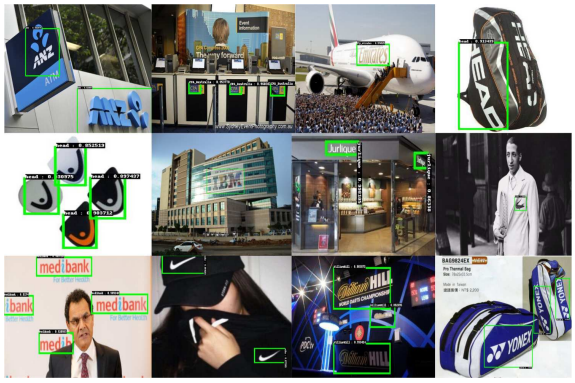

Figure 3. Some examples of detection results in the proposed logo image dataset. Logos in the green boxes are the detected results, and the detected categories are tagged around the green box.

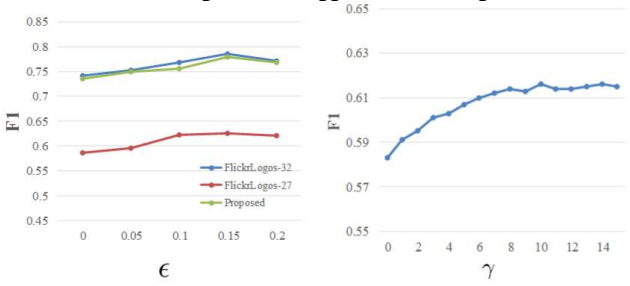

Figure 4. (a): The comparison of different ε in three image datasets. (b): The comparison of different γ in the video dataset.

are specifically designed for real-world logo recognition. FlickrLogos-32 is a collection of photos showing 32 different logo brands, which contains 8,240 images. FlickrLogos-27 contains more than four thousand images. They are established for the evaluation of logo retrieval and multi-class logo detection/recognition systems.

Since there is still no public video logo dataset, we collected a set of tennis videos, annotating each frame in the training set, and established a dataset for the detection evaluation. The dataset contains 20 different tennis video clips (more than twenty five thousand frames) including various camera motions such as switch of camera views, panning, zooming, and rotating, and the resulted object blurring and occlusion. The logos appear both in the background and on players' and staff's clothes. We annotated the positions of the top 20 high frequency categories of logos in videos. Meanwhile, we collect a corresponding logo recognition image dataset. Figure 3 shows some examples of logos in this image dataset. The new dataset contains the same 20 logo categories and about 100 images for each category. We also labeled the position of each logo in the image dataset. To evaluate the proposed approach, we randomly select 70% of data for training, and 30% for testing, respectively.

Logo	DPM	DTC	FRCN	Faster-RCNN	Proposed
adidas	9.75	43.6	61.6	50.2	**63.87**
aldi	25.9	**93.9**	67.2	71.6	66.2
apple	50.0	2.29	84.9	82.9	**86.22**
becks	N/A	**83.5**	72.5	64.7	62.86
bmw	N/A	33.3	70.0	81.3	**85.01**
carlsberg	5.4	58.1	49.6	63.5	**69.45**
chimay	2.9	**90.0**	71.9	79.1	82.53
cocacola	12.1	74.3	33.0	73.3	**81.84**
corona	38.5	**100.0**	92.9	98.0	96.97
dhl	0.02	**87.9**	53.5	69.2	62.55
erdinger	38.3	74.0	80.1	**84.1**	78.54
esso	62.4	86.7	88.8	90.1	**91.81**
fedex	8.75	**74.9**	61.3	73.0	65.98
ferrari	14.8	71.7	**90.0**	85.5	84.38
ford	33.4	74.0	84.2	84.1	**91.27**
fosters	27.4	60.0	79.7	84.5	**93.27**
google	58.2	86.3	85.2	95.0	**96.88**
guiness	46.1	**91.9**	89.4	90.3	91.82
heineken	3.13	**88.9**	57.8	77.3	81.53
HP	N/A	50.5	N/A	59.3	**64.40**
milka	0.18	49.1	34.6	56.2	**56.27**
nvidia	2.42	**57.5**	50.3	54.0	49.94
paulaner	31.0	96.4	98.6	98.0	**100.00**
pepsi	6.85	N/A	34.2	**43.0**	41.97
rittersport	3.2	74.3	63.0	77.0	**85.68**
shell	42.1	27.2	**57.4**	52.0	53.78
singha	80.0	85.1	**94.2**	85.9	91.98
starbucks	22.5	96.7	95.9	98.0	**99.50**
stellaartois	44.8	**96.7**	82.2	80.5	90.00
texaco	52.1	66.5	87.4	86.2	**87.95**
tsingtao	16.5	84.2	84.3	80.5	**89.44**
ups	54.7	77.3	**81.5**	80.5	72.50
MAP	27.4	72.2	74.4	76.4	**78.64**

Table 1. Comparison of MAPs(%) in each class and average MAPs(%).

6.2. Parameter Setting

The most commonly used experimental threshold θ in conventional network is 0.8 [9, 10, 21]. After testing the effect of different ε in three training data, we select 0.15 for ε as an empirical value. The parameters w_1 and w_2 in Eq. 1 are assigned 0.7 and 0.3, respectively. The parameter ξ in Eq. 8 is assigned 0.25. The comparison results are shown in Figure 4(a). For the initial frame set size γ for Fast-RCNN, we test the F1 values with different γ in video dataset. Experiments show that the detection performance converges when γ reaches 10, as shown in Figure 4(b).

6.3. Evaluation in Still Frames

We pre-train a basic VGG16 network with the ImageNet dataset, fine-tune and test the model in the three datasets separately. Training data is augmented by rotating, flipping, blurring and sharpening. Because of the wide use of FlickrLogos-32 in logo detection, we compare the re-ranking method on this dataset with several other logo detection methods, and then show the comparison of the proposed method with the baseline[9] in all the three datasets.

Method	Precision	Recall	F1
FRCN[9]	0.8	0.69	0.74
FRCN[9]+ρ (without λ)	0.73	0.76	0.74
FRCN[9]+ρ (with λ)	0.77	0.76	0.76
Proposed	0.81	0.76	0.79

Table 2. Comparison of results in FlickrLogos-32.

DataSet	Method	Precision	Recall	F1
FlickrLogos-32	FRCN[9]	0.801	0.691	0.742
	Proposed	0.812	0.760	0.785
FlickrLogos-27	FRCN[9]	0.582	0.591	0.586
	Proposed	0.673	0.585	0.626
Our own logo Dataset	FRCN[9]	0.731	0.741	0.736
	Proposed	0.818	0.745	0.780

Table 3. Comparison of results in the three logo dataset.

For FlickrLogos-32, we compare the MAP values with DPM [8] which detects logos based on mixtures of multiscale deformable part models, DTC [28] which uses improved topological constraint for SIFT features, FRCN [9] and Faster-RCNN [21] which use conventional RCNN to detect logos. Table 1 shows the comparison results with different detection algorithms. The proposed method obviously has a higher average MAP than the other methods, and gains the highest MAPs in 16 categories (out of 32).

Table 2 shows the effectiveness of each re-ranking factor. The factor ρ is effective in reducing the false negatives caused by blurred logos but introduces more false positives. The use of color consistency and inhomogeneous exclusion on top of ρ obviously boosts up the overall precision without undermining the recall.

The FlickrLogos-27 dataset is labeled mainly for classification without location information. So we annotate the data of FlickrLogos-27 manually. Since there are very few repetitions in FlickrLogos-27, the benefit of mutual-enhancement is less significant. Results in Table 3 show that the proposed method outperforms conventional Fast-RCNN in FlickrLogos-27 and our own logo dataset. Figure 3 illustrates some examples of detection results in our own dataset.

6.4. Evaluation in Videos

Figure 5 shows some comparison in the video dataset. When the similarity between adjacent frames obtained by Eq. 12 is less than the threshold, the logo candidates will be propagated. Otherwise, Fast-RCNN detection will restart to generate new logo candidates.

(a) Detection results in a frame sequence with blur and camera zooming

(b) Detection results in a frame sequence with occlusion and camera's translation

(c) Detection results in a frame sequence with camera's rotation

Figure 5. Detection results in video sequences with various camera changes.

Method	Precision	Recall	F1
FRCN[9]	0.738	0.482	0.583
FRCN[9]+re-ranking	0.725	0.516	0.603
FRCN[9]+propogation	0.741	0.478	0.581
Proposed	0.745	0.525	0.616

Table 4. Comparison of results in the video dataset.

We compare the propagation method with Fast-RCNN [9]. The values of precision, recall, and F1 are listed in Table 4. The re-ranking method gains a higher recall, and propagation further improves the precision. Overall, the proposed method has a higher F1. It shows that the characteristics of homogenous logos are indeed powerful features of sports video.

Our experiments on football and basketball videos show that the proposed method also outperforms the existing methods. However, none of the detection results on these videos are as good as the result on tennis videos. This is because that there are more challenging problems, such as small logos in long shots and the incompleteness due to frequent occlusions.

To measure the time efficiency of our method, we evaluate various stages of the proposed method in the proposed video dataset which contains 25,193 frames, and achieve a performance about 0.59s/frame. We obtain the most potential regions by propagation rather than Fast-RCNN detection in similar contiguous frames, which greatly reduces the region proposing and detection time. Meanwhile, frame propagation finds more true positive regions and discards more false positive regions for classification. For the fuzzy frames, frame propagation utilizes the location of the object in contiguous frames to predict the location in the current frame, which helps obtain more true positive regions compared with those methods that consider only still frames. Besides, the relay potency of false positive regions decreases rapidly, which leads to the corresponding regions to be removed immediately and further reduces the processing time.

7. Conclusion

In this study, we propose an integrative, effective, and efficient framework for logo detection in sports videos. The framework efficiently combines Fast-RCNN-based object detection with frame propagation. Based on the analysis of the characteristics of logos, proposal re-ranking, prediction, and verification schemes are extensively investigated and evaluated. The proposed approach shows consistent performance improvement over existing logo detections methods for sports videos. However, several challenges still remain to be addressed in the future. For instance, the extraction and utilization of texts in logos, rich variations in logo layout, and more tracking information could be considered.

8. Acknowledgments

This work is supported by National Natural Science Foundation of China under Grant 61673029.

References

[1] S. Bianco, M. Buzzelli, D. Mazzini, and R. Schettini. Logo recognition using cnn features. In *International Conference on Image Analysis and Processing*, pages 438–448. Springer, 2015.

[2] R. Boia and C. Florea. Homographic class template for logo localization and recognition. In *Iberian Conference on Pattern Recognition and Image Analysis*, pages 487–495. Springer, 2015.

[3] T. Chattopadhyay and A. Sinha. Recognition of trademarks from sports videos for channel hyperlinking in consumer end. In *2009 IEEE 13th International Symposium on Consumer Electronics*, pages 943–947. IEEE, 2009.

[4] N. Dalal and B. Triggs. Histograms of oriented gradients for human detection. In *2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05)*, volume 1, pages 886–893. IEEE, 2005.

[5] M. Danelljan, G. Hager, F. Shahbaz Khan, and M. Felsberg. Convolutional features for correlation filter based visual tracking. In *Proceedings of the IEEE International Conference on Computer Vision Workshops*, pages 58–66, 2015.

[6] M. Danelljan, A. Robinson, F. S. Khan, and M. Felsberg. Beyond correlation filters: Learning continuous convolution operators for visual tracking. In *European Conference on Computer Vision*, pages 472–488. Springer, 2016.

[7] R. J. den Hollander and A. Hanjalic. Logo recognition in video stills by string matching. In *Image Processing, 2003. ICIP 2003. Proceedings. 2003 International Conference on*, volume 3, pages III–517. IEEE, 2003.

[8] P. F. Felzenszwalb, R. B. Girshick, D. McAllester, and D. Ramanan. Object detection with discriminatively trained part-based models. *IEEE transactions on pattern analysis and machine intelligence*, 32(9):1627–1645, 2010.

[9] R. Girshick. Fast r-cnn. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 1440–1448, 2015.

[10] R. Girshick, J. Donahue, T. Darrell, and J. Malik. Rich feature hierarchies for accurate object detection and semantic segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 580–587, 2014.

[11] J. F. Henriques, R. Caseiro, P. Martins, and J. Batista. High-speed tracking with kernelized correlation filters. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 37(3):583–596, 2015.

[12] S. C. Hoi, X. Wu, H. Liu, Y. Wu, H. Wang, H. Xue, and Q. Wu. Logo-net: Large-scale deep logo detection and brand recognition with deep region-based convolutional networks. *arXiv preprint arXiv:1511.02462*, 2015.

[13] R. Jain and D. S. Doermann. Logo retrieval in document images. In *Document Analysis Systems*, pages 135–139, 2012.

[14] Y. Kalantidis, L. G. Pueyo, M. Trevisiol, R. van Zwol, and Y. Avrithis. Scalable triangulation-based logo recognition. In *Proceedings of the 1st ACM International Conference on Multimedia Retrieval*, page 20. ACM, 2011.

[15] B. Kovar and A. Hanjalic. Logo detection and classification in a sport video: video indexing for sponsorship revenue control. In *Electronic Imaging 2002*, pages 183–193. International Society for Optics and Photonics, 2001.

[16] L. Li, X. Wang, W. Zhang, G. Yang, and G. Hu. Detecting removed object from video with stationary background. In *International Conference on Digital Forensics and Watermaking*, 2012.

[17] D. G. Lowe. Distinctive image features from scale-invariant keypoints. *International journal of computer vision*, 60(2):91–110, 2004.

[18] H. Nam and B. Han. Learning multi-domain convolutional neural networks for visual tracking. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4293–4302, 2016.

[19] H. Nisar, A. S. Malik, and T. S. Choi. Content adaptive fast motion estimation based on spatio-temporal homogeneity analysis and motion classification. *Pattern Recognition Letters*, 33(1):52–61, 2012.

[20] G. Oliveira, X. Frazão, A. Pimentel, and B. Ribeiro. Automatic graphic logo detection via fast region-based convolutional networks. *arXiv preprint arXiv:1604.06083*, 2016.

[21] S. Ren, K. He, R. Girshick, and J. Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. In *Advances in neural information processing systems*, pages 91–99, 2015.

[22] J. Revaud, M. Douze, and C. Schmid. Correlation-based burstiness for logo retrieval. In *Proceedings of the 20th ACM international conference on Multimedia*, pages 965–968. ACM, 2012.

[23] S. Romberg and R. Lienhart. Bundle min-hashing for logo recognition. In *Proceedings of the 3rd ACM conference on International conference on multimedia retrieval*, pages 113–120. ACM, 2013.

[24] S. Romberg, L. G. Pueyo, R. Lienhart, and R. Van Zwol. Scalable logo recognition in real-world images. In *Proceedings of the 1st ACM International Conference on Multimedia Retrieval*, page 25. ACM, 2011.

[25] A. Saha, J. Mukherjee, and S. Sural. A neighborhood elimination approach for block matching in motion estimation. *Signal Processing Image Communication*, 26(8C9):438–454, 2011.

[26] H. Sahbi, L. Ballan, G. Serra, and A. Del Bimbo. Context-dependent logo matching and recognition. *IEEE Transactions on Image Processing*, 22(3):1018–1031, 2013.

[27] P. Shenolikar and S. Narote. Different approaches for motion estimation. In *Control, Automation, Communication and Energy Conservation, 2009. INCACEC 2009. 2009 International Conference on*, pages 1–4. IEEE, 2009.

[28] P. Tang and Y. Peng. Logo recognition via improved topological constraint. In *International Conference on Multimedia Modeling*, pages 150–161. Springer, 2016.

[29] J. R. R. Uijlings, K. E. A. V. D. Sande, T. Gevers, and A. W. M. Smeulders. Selective search for object recognition. *International Journal of Computer Vision*, 104(2):154–171, 2013.

[30] C. Wan, Z. Zhao, X. Guo, and A. Cai. Tree-based shape descriptor for scalable logo detection. In *Visual Communications and Image Processing (VCIP), 2013*, pages 1–6. IEEE, 2013.

[31] J. Wang, Q. Liu, L. Duan, H. Lu, and C. Xu. Automatic tv logo detection, tracking and removal in broadcast video. In *Advances in Multimedia Modeling, International Multimedia Modeling Conference, MMM 2007, Singapore, January 9-12, 2007. Proceedings*, pages 63–72, 2007.

[32] N. Wang and D.-Y. Yeung. Learning a deep compact image representation for visual tracking. In *Advances in neural information processing systems*, pages 809–817, 2013.

[33] X. Wu and K. Kashino. Image retrieval based on spatial context with relaxed gabriel graph pyramid. In *2014 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 6879–6883. IEEE, 2014.

[34] W. Q. Yan, J. Wang, and M. S. Kankanhalli. Automatic video logo detection and removal. *Multimedia Systems*, 10(5):379–391, 2005.

[35] B. Zhang, L. Wang, Z. Wang, Y. Qiao, and H. Wang. Real-time action recognition with enhanced motion vector cnns. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2718–2726, 2016.

[36] Q. Zhang, X. Shen, L. Xu, and J. Jia. Rolling guidance filter. In *European Conference on Computer Vision*, pages 815–830. Springer, 2014.