HOLLOW TV LOGO DETECTION

Liang Zhang[1,2], *Tian Xia*[1], *Yongdong Zhang*[1], *Jintao Li*[1]

[1]Advanced Computing Research Laboratory, Institute of Computing Technology,
Chinese Academy of Sciences, Beijing 100190, China
[2]Graduate University of Chinese Academy of Sciences, Beijing 100049, China

ABSTRACT

Most existing TV logo detection methods regard the logo's region as a whole and depend on the global features derived from it, these methods fit to solid TV logos. However, we find one special kind of TV logo which has large hollow areas within the logo's region and we name it "hollow TV logo." In this case, the global features are dramatically varying due to the changing content in the hollow areas, and traditional global feature based methods fail to detect such TV logos. In this paper, we propose a local feature based approach for hollow TV logo detection. It successfully suppresses the noises from hollow areas, and can complete detection in single-frame scenario. Compared to the traditional edge-based method, our method achieves 40.5% and 29.3% improvements on false alarm rate and false reject rate separately for hollow TV logos.

Index Terms— TV logo detection, single frame, local features, global features

1. INTRODUCTION

TV broadcast stations often use a unique TV logo to claim video content ownership, and TV logo detection is desirable in detecting commercial video, monitoring TV signal status, and re-broadcasting programs with a new logo.

Generally, traditional detection is usually performed in multi-frame scenario, and the difference between consecutive frames can be utilized to obtain a stable TV logo region. Based on the region, the global features, e.g., texture, shape, contour and color, are extracted to generate a template for unsupervised detection through template matching [1][2] or train a model for supervised detection through classification [3][4]. In practice, the learning-based approaches are not very feasible since they rely heavily on large amounts of manually labeled samples, and the negative samples (without a TV logo) are hard to be completely collected due to their rich diversity. Therefore, template based methods are more popular in real applications, and we focus on them in this paper. Moreover, the results of template matching can tell us not only whether a TV logo appears in a video, but also which TV station it is. It means the detection and recognition are both completed in this process. However, in this paper, we still utilize the term of "detection" to denote this process as most of the previous papers [2][3] did.

Although there is plenty of previous work on TV logo detection, we find one special kind of TV logo is ignored unfortunately. These TV logos are named as "hollow TV logos." As shown in Fig.1, we can see that a hollow TV logo mainly consists of irregular streak-shaped strokes, and there are plenty of gaps between the strokes. However, as shown in Fig.2, a traditional solid TV logo mainly consists of a solid pattern (as shown in a-d), or a solid pattern with small gaps in it (as shown in e-f).

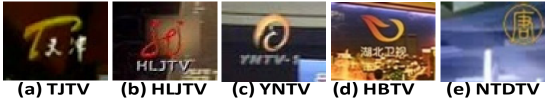


Fig.1. Some samples of hollow TV logos which have large hollow areas within the logo's region.

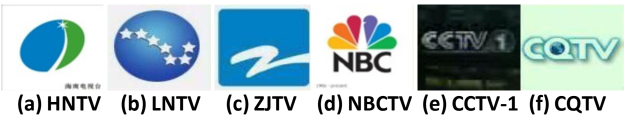


Fig.2. Some samples of traditional solid TV logos which mainly consist of a solid pattern.

In order to illustrate the difference formally, we present several concepts first. As shown in Fig.3, "cover region" (CR) is the area within a circumscribed circle/rectangle covering the whole logo; "logo region" (LR) is the area occupied by the logo in CR; and "hollow region" (HR) is the left area in CR. Obviously, we can get: CR=LR+HR.

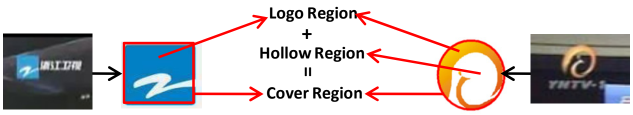


Fig.3. The composition of a TV logo: Cover Region consists of Logo Region and Hollow Region; the left one is the composition of a solid TV logo, and the right one is of a hollow TV logo.

From Fig.3, we can see that the main difference between hollow TV logos and solid ones lies on the ratio of HR-to-CR. A hollow TV logo has a large ratio of HR-to-CR, usually more than 40% (all the five hollow TV logos in Fig.1 have HR-to-CR ratios ranging from 40% to 60%); while a solid one has a small ratio of HR-to-CR, usually less than 20% (all the five TV logos in Fig.2 have HR-to-CR

978-1-4577-1303-3/11/$26.00 ©2011 IEEE

ratios below 20%), and we name it as a “solid TV logo.” In practical detection, a high HR-to-CR implies that only a small part of content (i.e., LR) retains stable in CR and a large part of content (i.e., HR) is the dramatically changing background. Traditional TV logo detection methods ignore such difference and mainly target solid logos. Therefore, existing methods fail to detect hollow TV logos, and the failure is mainly caused by two issues:

(1) It is hard to locate the hollow TV logo well through differentiating consecutive frames, since LR of hollow TV logo is too small.

(2) It is hard to get exact match through global features derived from the whole CR, since much of the information carried by the global feature is noisy due to the large HR area.

Therefore, in this paper, we investigate on how to detect hollow TV logos, and propose a local feature based method to handle the above two issues. Based on the assumption that the interference from a local region is much less than the one from the whole CR, we utilize a SIFT-like local feature to generate the logo template through visual words, and perform the template matching to detect hollow TV logos. Moreover, the local feature based template matching does not need exact logo localization first as traditional methods did, thus can perform detection in a single frame scenario, which is more preferable in real-time applications.

2. ANALYSIS FOR A HOLLOW TV LOGO

In this section, we investigate two important issues: one is how to locate a hollow TV logo, and the other is how to represent a hollow TV logo.

2.1 Rough Location of a Hollow TV Logo

A TV logo has a stable position for a long duration within a video to draw the attention of viewers. The difference image between consecutive frames is often used to locate LR in a multi-frame scenario. The difference image for a solid logo is shown in Fig.4 (a), a stable LR is obtained here. However, it is not feasible for a hollow logo as shown in Fig.4 (b), since LR is too small and the large HR areas bring with many interferences.

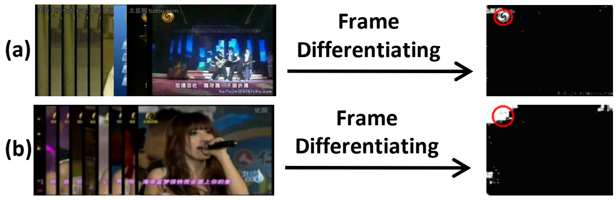


Fig.4 Frame difference for solid and hollow TV logos: (a) Logo region can be localized correctly through frame differentiating for a solid TV logo; (b) Logo region cannot be localized correctly through frame differentiating for a hollow TV logo.

Different from general logo detection in images, there is prior knowledge about a TV logo’s location in frames. As shown in Fig.5, some TV logos of Chinese channels usually appear in the upper left corner of the frame. Therefore, we can utilize this prior knowledge by performing detection only within the upper-left corner of the frame, i.e., the rectangle with 1/4 of the frame’s width and height as shown in Fig.5 (the rough region can be modified with different prior knowledge). Moreover, our method depends on local feature based matching, which does not need exact localization, and this rough localization is sufficient.

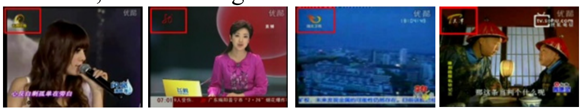


Fig.5. Prior knowledge about TV logo’s rough location in video frames, some Chinese channels have logos appearing in the upper left corner of the frame.

2.2 Representation of a Hollow TV Logo

The representation of a hollow TV logo is an important issue in detection. In this section, the advantages of the local representation compared to the global one is analyzed, and a SIFT-like representation of a hollow TV logo is presented.

2.2.1 Global Representation v.s. Local Representation

Traditional TV logo detection relies on the global feature derived from CR, since it targets on a solid logo which has stable CR content. However it is not reasonable for a hollow TV logo, since it has large HR areas within CR as shown in Fig.3, and the HR areas will bring noises into detection.

Compared with the global representation, local representation can suppress the noises from HR. Take the following test as an example. We compare the standard deviations of the pixel-level gray values for two regions: one is the “global region” which is the rough logo region as shown in Fig.5, the four images in Fig. 6 illustrate the “global regions” in four frame images; and the other is the “local region” which is a small round area (4*4 pixels) centered at the corner points as shown in Fig. 6, in which the red circle regions are the “local regions.”

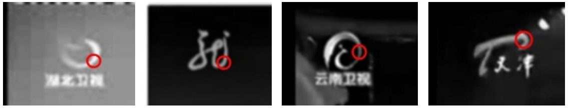


Fig.6. Four testing samples of global region and local region: the whole sub-image is a global region; the red marked region is a local region, which is a small round area centered at the corner point.

	Sample 1	Sample 2	Sample 3	Sample 4
Global std.	40.62	32.58	38.96	33.15
Local std.	3.26	3.23	4.57	4.38

Table 1 The standard deviations of gray values for global region and local region over the four samples in Fig.6.

Table 1 summarizes the statistics for four testing samples, we can find that the “local region” has much smaller variation than the “global region,” which means that the "local region" is smoother and has the ability to suppress the interference from the background. Therefore, local representation is beneficial to hollow TV logo detection.

2.2.2 Description of a Local Region

The local region mentioned above should be described in mathematical form. Fig.7 shows the gradient orientation histograms of gray pixels for the local region (red marked region in Fig.6), and the three histograms represent the gradient orientation distributions under different color backgrounds. By observing these histograms, despite the ever-changing of backgrounds, gradient orientations for the local region remain stable. This illustrates that the gradient orientation histogram is a robust descriptor.

SIFT [5] describes the local image gradients at the selected scale in the region around each key point, so it seems a good choice for our task. Moreover, considering the characteristics of TV logos, the utilized descriptor in our method is a simplified version of SIFT. SIFT's invariance to the geometry changes including scaling, rotation and affine transformations is eliminated; however, the invariance to the photometry changes is retained. We name this simplified local descriptor as "SIFT-like descriptor." In our method, the SIFT-like descriptor is 128 dimensional (4*4 pixels * 8 bins).

3. DETECTION PROCESS

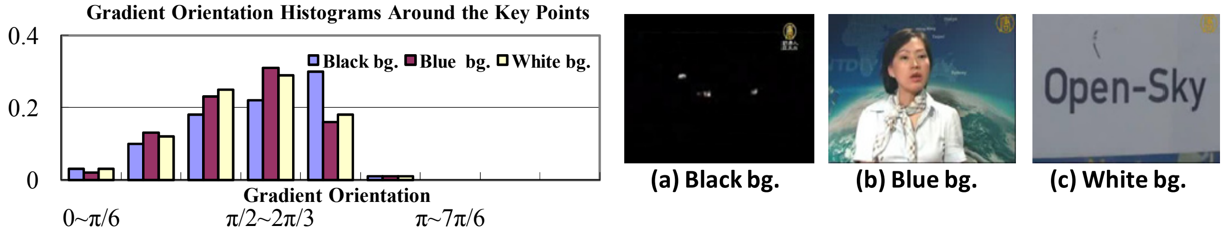
Based on the above analysis, our detection process is divided into two basic steps: template generation and template matching. The flowchart is shown in Fig.8.

3.1 Template Generation

To further enhance the robustness of the template, smooth background frames selected manually are used in template generation. Harris corner detection [6] is performed in the rough logo region as shown in Fig.5. 128 dimensional SIFT-like descriptors are figured out for the key points. To obtain a robust template, SIFT-like descriptors are quantized into visual words using K-means clustering algorithm. The generation of the template is actually the building of a visual dictionary on the selected frames, and the template can be regarded as the learned visual vocabulary.

3.2 Template Matching

The key idea of template matching is that local regions of the TV logo are described with the visual words from the template. However, local feature is always variable due to the noises, illumination changes, and the instability in the feature detection. In order to suppress the interferences from the variable local features, the distances between each SIFT-like vector and the cluster centers are compared, and each descriptor is soft-assigned to the top five nearest clusters. If the sum of the weights from each cluster exceeds a predefined threshold, one soft-assigning match is established between the current frame and a template. Soft-assignment provides a robust match since it gives large weights to the close clusters and small ones to the distant clusters. To avoid false detection as much as possible, the number of the soft-assigning match should be greater than 1/2 of the size of the vocabulary.



(a) Black bg. (b) Blue bg. (c) White bg.

Fig.7. The gradient orientation histograms around the key point under different backgrounds, they seem to have similar distributions illustrating the robustness of the descriptor (bg. stands for background).

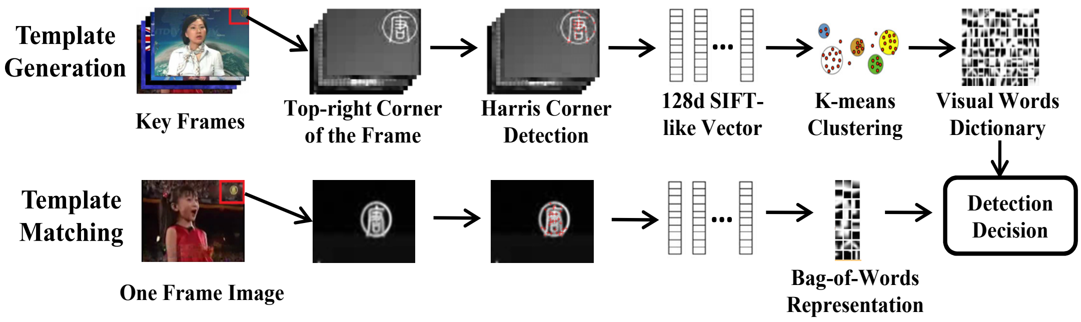


Fig.8. The flowchart of our detection process.

4. EXPERIMENT RESULT

In this section, we conduct two experiments: one is the performance comparison with the global feature based method [7]. The other is a verification on a large scale web video dataset to evaluate the precision of our method, since the precision attracts more attention in real practical scenarios, e.g., video filtering and broadcasting surveillance. All the experiments are carried out on a PC with Core2 Duo 3.3 GHz CPU and 2GB memory.

With the consideration of time efficiency, 30 is chosen as the number of key points on each frame and 20 is chosen as the number of visual words for the template generation.

4.1 Performance Comparison

In this section, we choose five hollow TV logos as shown in Fig.1, i.e., TJTV, HLJTV, YNTV, HBTV, and NTDTV. For each logo, 11 videos are collected with 10 videos for testing and one video for template generation. As for testing, 200 frames are uniformly extracted from each testing video. Therefore, the testing set consists of 100,000 frames, and 2,000 frames for each logo. The training set for generating a template of each logo is a set of manually selected frames with clear logo appearance from the training video, and 20 frames are selected to form the set in our experiments.

We compare our method to the edge-based method in [7], in which Canny edge detection and matching is performed on the rough logo region as shown in Fig.5. False alarm rate (FAR) and false reject rate (FRR) are used as the evaluation metric. From Table 2, we can see that our method outperforms edge-based method on all the five hollow logos, and achieves 40.5% improvement on average FAR, 29.3% improvement on average FRR, and 16.8% improvement on efficiency.

Moreover, we further conduct evaluation on three solid TV logos as shown in Fig.2, i.e., ZJTV, HNTV, and LNTV. From Table 2, we can see that our method achieves 16.7% improvement on average FAR, but with higher FRR. The reason is that solid logos tend to be short of local information which degrades the quality of the template.

4.2 Verification on the Web Video Dataset

In order to evaluate the precision, the verification on a large scale web video dataset [8] including 256,661key frames from 9,170 videos of YouTube is conducted. None of the five hollow TV logos concerned in Section 4.1appears in the dataset, and only 1/100,000 of them are incorrectly detected as having one of the five logos by our method. Therefore, our method achieves a very high precision, and is applicable in broadcasting surveillance.

5. CONCLUSION

In this paper, we focus on the detection of hollow TV logos which feature large hollow areas embedded in logos, and propose a local feature based method to perform the detection. It successfully suppresses the noises from hollow areas, and can be applied in single-frame scenarios.

We also find that our method achieves limited success on some solid TV logos due to little local information embedded in them. In the future, we will focus on the measurement of hollow degree of TV logos, and investigate the relationship between hollow degree and the performance. Furthermore, we will try to facilitate template generation through automatically selecting training frames from a given video.

TV channel	Our method			Edge-based method		
	FAR (%)	FRR (%)	Cost time (ms)	FAR (%)	FRR (%)	Cost time (ms)
TJTV	1.6	30.8	26	2.2	50.1	32
HLJTV	2.2	28.1	28	3.6	42.1	33
YNTV	1.8	21.6	26	2.9	31.6	35
HBTV	1.6	23.6	29	2.6	30.5	31
NTDTV	0	32.0	25	0.8	38.1	30
ZJTV	2.5	36.1	29	3.9	23.3	32
HNTV	2.6	50.5	25	2.6	22.8	33
LNTV	2.9	32.2	26	3.1	22.6	31

Table2 Performance comparison to the edge-based approach, the above five are hollow TV logos, the last three are solid TV logos.

6. ACKNOWLEDGEMENT

This work was supported by the Beijing Natural Science Foundation under Grant 4112055; by the National Basic Research Program of China (973 Program) under Grant 2007CB311100; by the National High Technology and Research Development Program of China (863 Program) under Grant 2009AA01A403; by the Co-building Program of Beijing Municipal Education Commission.

7. REFERENCES

[1] J. Wang, et al., “A Robust Method for TV Logo Tracking in Video Streams,” in Proc. ICME, pp.1041-1044, 2006.

[2] J. Wang, et al., “Automatic TV Logo Detection, Tracking and Removal in Broadcast Video,” in Proc. MMM, pp. 63-72, 2007.

[3] W. Yan, et al., “Automatic Video Logo Detection and Removal,” in ACM Trans. on Multimedia System, pp. 379-391, July 2005.

[4] P. Nieto, et al., “A TV-logo Classification and Learning System,” in Proc. ICIP, pp.2548-2551, 2008.

[5] D. Lowe, “Distinctive Image Features from Scale-Invariant Keypoints,” in IJCV, vol. 20, pp.91–110, 2003.

[6] C. Harris, et al., “A Combined Corner and Edge Detector,” in Proc. BMVC, pp. 147-151, 1988.

[7] A.R. Santos, et al., “Real-Time Opaque and Semi-Transparent TV Logos Detection,” in Proc. 5th International Information and Telecommunication Technologies Symposium, 2006.

[8] J. Cao, et al., “MCG-WEBV: A Benchmark Dataset for Web Video Analysis,” Technical Report, ICT-MCG-09-001, Institute of Computing Technology, CAS, 2009.