A ROBUST METHOD FOR TV LOGO TRACKING IN VIDEO STREAMS

Jinqiao Wang[1,2], *Lingyu Duan*[2,3], *Zhenglong Li*[1], *Jing Liu*[1], *Hanqing Lu*[1], *Jesse S. Jin*[3]

[1] National Lab of Pattern Recognition
Institute of Automation, Chinese Academy of Sciences, Beijing 100080, China
{jqwang, zlli, jliu, luhq}@nlpr.ia.ac.cn
[2] Institute for Infocomm Research, 21 Heng Mui Keng Terrace, Singapore 119613
{lingyu}@i2r.a-star.edu.sg
[3] The School of Design, Communication and Information Technology,
University of Newcastle, NSW 2308, Australia
{Jesse.Jin}@newcastle.edu.au

ABSTRACT

Most broadcast stations rely on TV logos to claim video content ownership or visually distinguish the broadcast from the interrupting commercial block. Detecting and tracking a TV logo is of interest to TV commercial skipping applications and logo-based broadcasting surveillance (abnormal signal is accompanied by logo absence). Pixel-wise difference computing within predetermined logo regions cannot address semi-transparent TV logos well for the blending effects of a logo itself and inconstant background images. Edge-based template matching is weak for semi-transparent ones when incomplete edges appear. In this paper we present a more robust approach to detect and track TV logos in video streams on the basis of multispectral images gradient. Instead of single frame based detection, our approach makes use of the temporal correlation of multiple consecutive frames. Since it is difficult to manually delineate logos of irregular shape, an adaptive threshold is applied to the gradient image in subpixel space to extract the logo mask. TV logo tracking is finally carried out by matching the masked region with a known template. An extensive comparison experiment has shown our proposed algorithm outperforms traditional methods such as frame difference, single frame-based edge matching. Our experimental dataset comes from part of TRECVID2005 news corpus and several Chinese TV channels with challenging TV logos.

1. INTRODUCTION

TV broadcast stations rely on logos to claim video content ownership or visually distinguish the broadcast clearly from the interrupting commercial block. A robust method for detecting and tracking a TV logo in video streams is useful for commercial skipping type applications [1, 2] and broadcast surveillance. Firstly, the existence and absence of a TV logo is a reliable indicator of program segments and commercial segments in many television programs such as CNN, NBC, MSNBC. Secondly, since cable TV via the satellite network has developed as a programming vehicle, the tracking of a TV logo can be employed to monitor the signal status of a particular TV channel at the local transmitter side to secure safe broadcasting. Herein it is assumed that abnormal TV signals usually cause the absence or distortion of a TV logo. Thirdly, exactly detecting a TV logo is required for automatic TV logo removal with image inpainting techniques [3, 4] in order to improve the viewing experience of rebroadcast programs. TV logo has three types: opaque, semi-transparent and animated. Examples are given in Fig. 1.

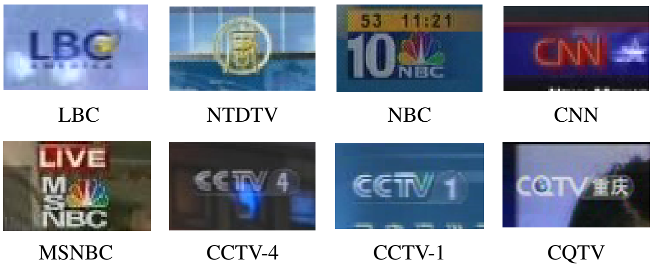

Fig. 1. Examples of the magnified TV logo images(LBC, NTDTV, NBC, CNN and MSNBC are opaque logos; CCTV-4 and CCTV-1 are semi-transparent logos; CQTV is an animated logo.)

For the convenience of reviewing previous work about TV logos, we introduce two terms "logo detection" and "logo tracking". The first is meant to locate and extract the mask of a logo within a whole frame or a predetermined region from a single image or a sequence of images. The latter is meant to track the existence or absence of a known logo over time in video streams. Logo detection can be manually or automatically accomplished. An exact mask is of interest to image inpainting related applications. Logo tracking is required to be insensitive to image noises, cluttered background, and partial occlusion derived from the blending effect of a semi-transparent logo. A key issue of logo tracking is to robustly model a logo with low-level visual features (e.g. color, edge, texture, etc.).

Below we review related work. Meisinger *et al.* [4] used the difference image between consecutive frames to extract the logo mask through assuming that the video content changes over time except for the logo. This assumption implied the limitation of an opaque logo only. Yan *et al.* [3] utilized a learning approach (i.e., neural network) to classify candidate logo regions as True or False by using local color and texture features. Like [4] difference images were used to determine the candidate logo regions. In order to achieve a reliable solution, learning-based approaches rely on large amounts of manually labeled samples. The real performance tightly depends on training samples. Albiol *et al.* [1] used the time-averaged gradients of a series of successive images plus morphological operations to extract the coarse mask of a TV logo wherein stable contours are assumed to characterize logo regions. In [1], the TV logo is utilized to determine the commercial segments at the shot-level. Only one frame (the last one) is selected from a shot to perform the gradients-based matching within the coarse logo mask. As discussed above, [3] and [4] are closely related to logo detection. In term of logo mask, some post-processing techniques have been proposed to refine the coarse mask such as Markov Random Fields (MRF) based contour relaxation [4] and heuristic edge selection [3]. Essentially, [1] is related to logo tracking. But only one frame per shot was applied. As the visual content changes within a shot may encounter the blending effects induced "occlusion", one frame is insufficient for determining the logo's existence or absence at the shot level. A granularity issue (i.e. frame level, shot level, or uniform temporal window level) actually exists in the context of logo tracking. Once the granularity is more than one frame, the temporal constraints from adjacent frames could be taken into account at the phase of logo matching.

As discussed above, logo detection and logo tracking are dealt with as two different tasks. Previous work mainly concern logo detection. In this paper we focus on robust logo tracking in extensive TV broadcast video stream. A gradient detection technique for multispectral images [5] is extended to a sequence of images for capturing persistent contour over time. Hereafter we call such a temporally extended gradient as a generalized gradient. Particularly in the case of a semi-transparent TV logo, the generalized gradient is able to alleviate the noisy edges from the cluttered background and enhance the incomplete contour (i.e. remove partial occlusion from blending) by temporal accumulation. By choosing a suitable granularity, missed detection and false alarm can be reduced. Comparison experiments have shown prominent advantages of the generalized gradient in terms of the contour representation of a semi-transparent logo. Also we propose the use of Ostu's binarization method [6] to accomplish logo detection. OSTU method is a locally adaptive binarization method for document images. This method is useful for extracting logo masks from images with low contrast, variable background intensity and noise. It complements the pixel-wise difference based approaches, which may fail when the logo is semi-transparent, irregular, or hollow with fine contours such as CCTV logos in Fig. 1.

The rest of this paper is organized as below. Section 2 briefly introduces TV logo detection. Section 3 presents the generalized gradient algorithm and logo matching. Section 4 gives an evaluation method of logo tracking and discusses experimental results. Section 5 finally draws a conclusion.

2. LOGO DETECTION

TV logos are often fixed and comprise very few image pixels, say several hundred pixels for CIF image size in MPEG-1 video streams. In order to make full use of the mask information, we conduct logo detection in sub-pixel space. Our implementation enlarges the image by triple using bilinear interpolation. Given an image, OSTU algorithm [6] is employed to automatically determine an optimal threshold to binarize the image. This algorithm is executed by maximizing discriminant measure variable of an image in gray levels. Given an image represented by L gray levels $[1, 2, 3, ...L]$. The number of pixels at level i is n_i and the total number of pixels is N. the gray-level histogram is normalized and regard as a probability distribution $p_i = \frac{n_i}{N}$. By utilizing the zeroth cumulative moment $\omega(k) = \sum_{i=1}^{k} p_i$ and the first cumulative moment $\mu(k) = \sum_{i=1}^{k} ip_i$ of the histogram, the optimal threshold k^* can be obtained by discriminative criterion as below.

$$\sigma^2(k^*) = \max_{1\leq k\leq L}\left(\frac{(\mu(L)\omega(k) - \mu(k))^2}{\omega(k)(1-\omega(k))}\right) \tag{1}$$

OSTU is originally used in the binarization of document images. Like document images, TV logo regions show similar visual characteristics and pose similar challenges such as low contrast, variable background and noise. Since OSTU method is nonparametric and unsupervised, it is simple and effective for extracting the logo mask. Fig. 2 illustrates the detection of a semi-transparent CCTV logo.

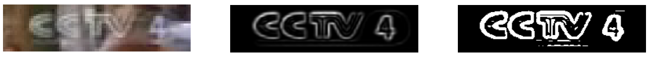

Fig. 2. TV logo detection (Left to right: the original logo image, the gray-level image, and the logo's binary mask by OSTU)

3. LOGO TRACKING

One advantage of our proposed approach lies in the incorporation of temporal context to enhance logo template modeling and matching. Gradient information is used as low-level visual features. The so-called generalized gradient seeks persistent gradients by obtaining support from neighbouring frames.

3.1. Generalized gradient

A generalized gradient is referred to the temporal extension of traditional gradient detection from a single image to a sequence of images. Different from simply averaging the gradients of multiple frames over time [1], we employ the technique of tensor gradient of a multi-image [5]. Explicit formulas for the direction along which the rate of change is maximum, as well as for the maximum rate of change itself, over multiple images in a video, can be derived.

A video segment is a sequence of neighbouring images, which can be packed and treated as a multi-valued image by modeling as an array of ordinary color images. The method of gradient calculation in a multi-valued image is then applied. Let $\Theta(x_1, x_2) : R^2 \rightarrow R^m$ be a multi-valued image with components $\Theta_i(x_1, x_2) : R^2 \rightarrow R, i = 1, 2, ..., m$. For a color image, there are R,G,B components, namely $(m = 3)$. Let us consider a video segment consisting of n frames, each frame having 3 color components. Through integrating temporal information, a video segment is expressed as: $\Theta(x_1, x_2) : R^2 \rightarrow R^{3n}$. The image value at a given spatial point (x_1^1, x_2^1) is a vector in R^{3n}. The difference between two image values at the points $M = (x_1^1, x_2^1)$ and $N = (x_1^2, x_2^2)$ is denoted by $\triangle\Theta = \Theta(M) - \theta(N)$. By dealing with $M - N$ as an infinitesimal displacement, the image value difference becomes the differential

$$d\Theta = \sum_{i=1}^{2} \frac{\partial\Theta}{\partial \mathbf{x}_i} dx_i \tag{2}$$

The squared norm $d\Theta^2$ is called the first fundamental form and is given by:

$$d\Theta^2 = \sum_{i=1}^{2}\sum_{j=1}^{2} \frac{\partial\Theta}{\partial x_i}\frac{\partial\Theta}{\partial x_j} dx_i dx_j \tag{3}$$

$$d\Theta^2 = \sum_{i=1}^{2}\sum_{j=1}^{2} g_{i,j} dx_i dx_j = \begin{bmatrix} dx_1 \\ dx_2 \end{bmatrix}^T \begin{bmatrix} g_{11} & g_{12} \\ g_{21} & g_{22} \end{bmatrix} \begin{bmatrix} dx_1 \\ dx_2 \end{bmatrix} \tag{4}$$

where $g_{i,j} := \frac{\partial\Theta}{\partial x_i} \cdot \frac{\partial\Theta}{\partial x_j}$,$g_{11}, g_{12}, g_{21}, g_{22}$ are the components of a symmetric tensor field. $d\Theta^2$ is a measure of the rate of change in the multi-value image, the corresponding eigenvalues are

$$\lambda_{\pm} = \frac{g_{11} + g_{22} \pm \sqrt{(g_{11} - g_{22})^2 + 4g_{12}^2}}{2} \tag{5}$$

and the corresponding eigenvectors are given by $(cos\phi_{\pm}, sin\phi_{\pm})$, where the angles $\phi_{\pm}$ (modulo π) are

$$\phi_+ = \frac{1}{2}\arctan\frac{2g_{12}}{g_{11}-g_{22}} \quad \phi_- = \frac{\pi}{2} + \frac{1}{2}\arctan\frac{2g_{12}}{g_{11}-g_{22}} \quad (6)$$

For a given spatial point $x = \{x_1, x_2\}$, the eigenvectors provide the direction of maximal and minimal change in the spatio-temporal domain; the eigenvalues are the corresponding rates of change in the temporal and spatial domain using an Euclidean metric. From a mathematical point of view, $\Theta(x_1, x_2)$ is a vector valued function defined over a manifold, hence the generalized gradient must be a tensor [5]. Let ϕ_+ and ϕ_- denote the direction of maximal change and minimal change, respectively, λ_+ and λ_- the maximal rate of change and the minimal rate of change correspondingly.

Below we discuss two particular cases: single gray image ($m = 1$) and single color image ($m = 3$). For $m = 1, \lambda_+ \equiv \|\nabla\Theta\|^2, \lambda_- \equiv 0$, the generalized gradient is always perpendicular to the level-sets; for $m = 3, \Theta(x, y) = (R(x, y), G(x, y), B(x, y))$, Let r, g, b denote the unitary vectors associated with the R, G, B axis, $u = \frac{\partial\Theta}{\partial x} = \frac{\partial R}{\partial x}r + \frac{\partial G}{\partial x}g + \frac{\partial B}{\partial x}b$, and $v = \frac{\partial\Theta}{\partial y} = \frac{\partial R}{\partial y}r + \frac{\partial G}{\partial y}g + \frac{\partial B}{\partial y}b$, then g_{11}, g_{12}, g_{21} and g_{22} can be written as

$$g_{11} = u \cdot u = \left|\frac{\partial R}{\partial x}\right|^2 + \left|\frac{\partial G}{\partial x}\right|^2 + \left|\frac{\partial B}{\partial x}\right|^2 \quad (7)$$

$$g_{22} = u \cdot v = \left|\frac{\partial R}{\partial y}\right|^2 + \left|\frac{\partial G}{\partial y}\right|^2 + \left|\frac{\partial B}{\partial y}\right|^2 \quad (8)$$

$$g_{12} = g_{21} = v \cdot v = \frac{\partial R}{\partial x}\frac{\partial R}{\partial y} + \frac{\partial G}{\partial x}\frac{\partial G}{\partial y} + \frac{\partial B}{\partial x}\frac{\partial B}{\partial y} \quad (9)$$

For a sequence of images, the resulting edge is not simply given by the rate of maximal change λ_+, but by comparing λ_+ to λ_-. If $\lambda_+ = \lambda_-$, the image sequence changes at equal rates in all directions. Image discontinuities can be detected by defining a function $f = f(\lambda_+, \lambda_-)$ to measure the dissimilarity between λ_+ and λ_-. A suitable choice of the function is in the form $f = f(\lambda_+ - \lambda_-)$ [7]. Since $f(\lambda_+ - \lambda_-)$ is the analog mutispectral extension of $f = f(\|\nabla\theta\|^2)$ for single gray images($i.e., m = 1$), it reduces to the gradient-based edge detector.

By jointly considering R, G, B color components and employing temporal accumulation, the generalized gradients enhance the persistent edges belonging to a TV logo. It helps remove or weaken the noisy edges from changing background video content, as those edges are instable over time. As illustrated in Fig. 3, the energy distribution of enhanced edges at the CCTV-4 channel logo is stronger than that at the background area in the generalized gradient image.

3.2. Matching

In consideration of the constant position of a TV logo, we calculate the generalized gradients over a sequence of images to build the matching template. Such temporal accumulation helps reduce background noises and produce a clear contour of the TV logo. The final template is obtained by binarizing the generalized gradient image by using OSTU method at the phase of logo detection.

The number of neighbouring frames is closely related to the template modeling and matching. Insufficient frames cannot eliminate noises from background contents that could increase false alarms. Too many frames tend to blur a logo's inherent features that could increase missed segments. Moreover, the number of adjacent frames affects the temporal resolution of tracking.

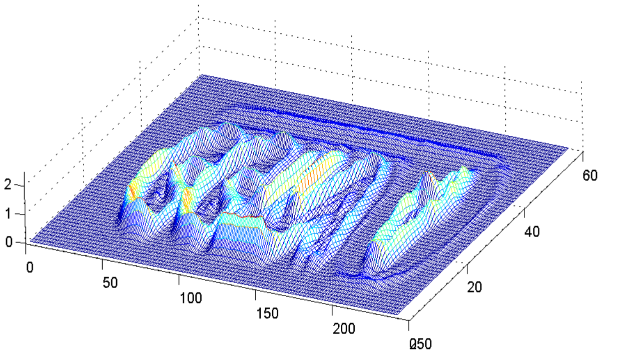

Fig. 3. Energy distribution of edges based on generalized gradients computation at the CCTV-4 logo area

Accordingly, a two-level logo matching scheme is proposed. Overlapped sliding windows are applied. At both levels, the existence or absence of the TV logo is decided by the matching criteria:

$$C(I, T) = \sum_{T(i,j)=1} (I(i, j)) \quad (10)$$

where $I(i, j)$ is the binary image derived from the gradients of consequent frames, $T(i, j)$ is the matching template. If $C(I, T) \geq Th$, the TV logo is existent; otherwise, the TV logo is absent. At the first level, a coarse resolution (i.e., more adjacent frames, say 90 frames) is used to roughly determine the boundaries of segments in the absence or the existence of a logo. At the second level, a fine resolution (say 1 frame) is used to precisely locate the transition points by shifting windows backwards and forwards around current time stamp. Twin thresholds are consequently applied. A tight threshold, say 0.8, is used at the first level while a loose threshold, say 0.5, at the second level. A fast running speed (more than real-time) also benefits from this coarse-to-fine scheme.

4. EXPERIMENT

Our experimental video data (around 4 hours) is extensively collected from TRECVID'05 news video corpus and several challenging chinese TV channels. The video is in MPEG-1 format with the frame rate of 29.97 fps and the frame size of 352×240. TRECVID'05 corpus includes 6 channels: CNN, NBC, MSNBC, LBC, CCTV4, NTDBC; Chinese TV channels include CCTV1, CCTV8, HNTV, GZTV and CQTV. CQTV channel has an animated logo.

Referring to Fig. 4, our approach is compared with two previous algorithms: edge-based matching and pixel-wise difference computing. For each approach, the results by using different number of neighbouring frames are also compared. For edge-based matching, Canny edge detector is applied and the resulting edge images are time-averaged. In order to reduce false alarms, OSTU method is employed to derive the final edge mask instead of morphology operators [1]. For pixel-wise difference computing, the gray-level difference images are temporally accumulated. OSTU method is also applied to get the final mask. As illustrated in Fig. 4, our approach produces a solider and clearer contour with 40 frames to calculate generalized gradient than the other approaches. When the neighbouring frames are 150, all the three approaches get a clear contour.

The logo tracking performance varies with different channels as indicated in Table 1. Since CCTV-4 logo in TRECVID'05 corpus is

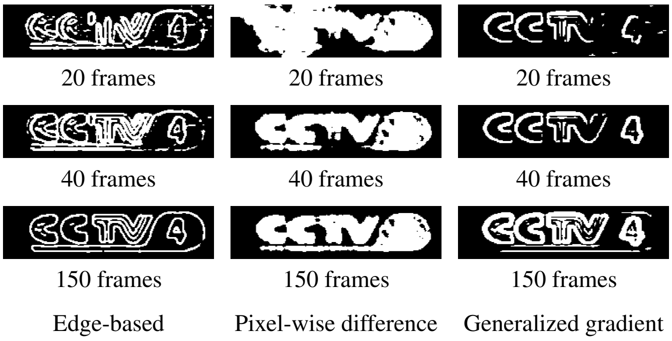

Fig. 4. Comparison of three TV logo detection algorithms

challenging, we use it to quantitively compare three approaches. As CCTV-4 has no negative samples (the logo always remain there), we use videos from other channels as negative samples (the logo region is replaced by changing video content). F1 ($F1 = \frac{2*Recall*Precision}{Recall+Precision}$) is used to evaluate the result. Fig. 5 shows the results of three approaches over different number of neighbouring frames. For relatively fewer neighbouring frames, our approach generally delivers better results. The number of neighbouring frames is decided by an application as it affects the temporal resolution. The tracking error comes from false alarm and false reject as shown in Fig. 6. With the increase of neighbouring frames incorporated to TV logo tracking, the contour of the binarized TV logo image is clearer and the false alarm rate(FAR) and the false reject rate(FRR) are both becoming lower for our method and edge-based method. When the number of neighbouring frames is higher than 300, both the FAR and FRR are reach zero for the three approaches. For pixel-wise difference method, the FAR is very high, initially about 95%. With the increase of neighbouring frames, the FRR increases while the FAR decreases.

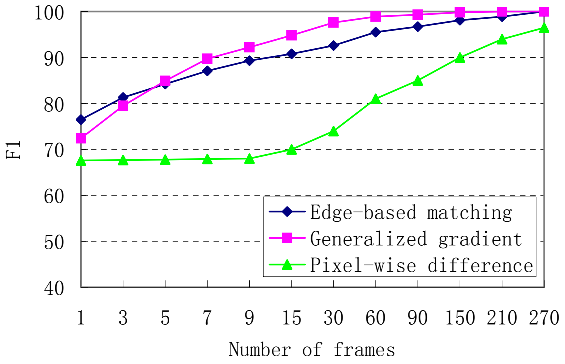

Fig. 5. Logo tracking results with different number of neighbouring frames and comparison with edge-based and pixel-wise approaches.

Table 1 lists the tracking results of different channels including opaque, semi-transparent, and animated logos as shown in Fig. 1, where the number of neighbouring frames is set to 90 (around 3 sec in length). The video duration of each channel is around 30 mins.

5. CONCLUSION

OSTU based TV logo detection and generalized gradients based TV logo tracking approach is suggested. It is shown to outperform traditional techniques. Satisfactory results have been achieved on opaque, semi-transparent, and simple animated TV logos. Future work includes the improvement of temporal resolution of logo tracking.

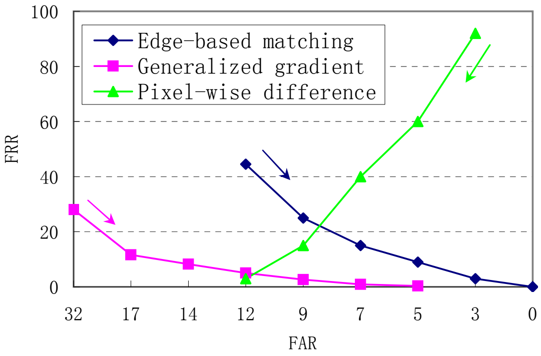

Fig. 6. Curve of FAR and FRR. The arrows indicate the curve trend along the increasing number of neighbouring frames.

Table 1. Experiment results of logo tracking

Name of TV channels	FAR	FRR	F1
MSNBC	0.4%	0%	99.80%
NTDTV	2.2%	0.61%	98.59%
LBC	1.74%	0.24%	99%
CNN	1.03%	0.24%	99.36%
NBC	2.42%	0.2%	98.68%
CCTV-4	3.9%	0.52%	97.76%
CQTV	2.4%	0.32%	98.63%

6. ACKNOWLEDGEMENT

This work is supported by National Natural Science Foundation of China (Grant No. 60475010 and 60121302).

7. REFERENCES

[1] A. Albial, M. J. C. Fullà, A. Albial, and L. Torres, “Detection of tv commercials,” *Proc. ICASSP’04*, May 2004.

[2] Nevenka Dimitrova, Thomas Mc Gee, and Jan Hermanus Elenbaas, “Apparatus and method for locating a commercial disposed within a video data stream,” *United States Patent 6100941*, August 2000.

[3] Wei-Qi Yan, Jun Wang, and Mohan S. Kankanhalli, “Automatic video logo detection and removal,” *ACM Trans. on Multimedia System*, July 2005.

[4] Katrin Meisinger, Tobias Troeger, Marcus Zeller, and André Kaup, “Automatic tv logo removal using statistical based logo detection and frequency selective inpainting,” *Proc. European Signal Processing Conference’05*, September 2005.

[5] S. Di Zenzo, “A note on the gradient of a multi-image,” *Comput. Vision Graphics Image Processing*, vol. 33, pp. 116–125, 1986.

[6] N. Otsu, “A threshold selection method from gray-level histograms,” *IEEE Trans. Systems, Man and Cybernetics*, vol. 9, pp. 62–66, 1979.

[7] G. Sapiro, “Vector (self) snakes: a geometric framework for color, texture, and multiscale image segmentation,” *Proc. ICIP’96*, vol. 1, pp. 817–820, Sept 1996.