A Phase-based Approach for Caption Detection in Videos

Shu Wen, Yonghong Song, Yuanlin Zhang, Yu Yu

Institute of Artificial Intelligence and Robotics, Xi'an Jiaotong University,
No.28, Xianning West Road, Xi'an, Shaanxi, P.R. China

Abstract. The captions in videos are closely related to the video contents, so the research of automatic caption detection contributes to video contents analysis and content-based retrieval. In this paper, a novel phase-based static caption detection approach is proposed. Our phase-based algorithm consists of two processes: candidate caption region detection and candidate caption region refinement. Firstly, the candidate caption regions are extracted from the caption saliency map, which is mainly generated by phase-only Fourier synthesis. Secondly, the candidate regions are refined by text region shape features. The comparison experimental results with existing methods show a better performance of our proposed approach.

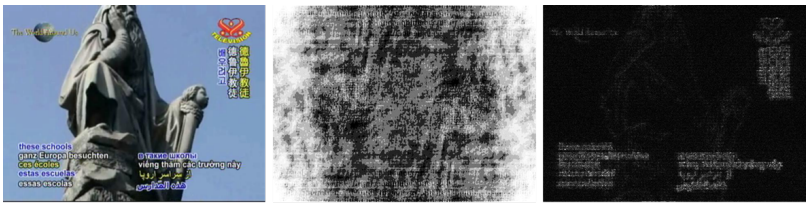
1 Introduction

With the rapid development of the Internet and video web sites (like YouTube), more and more digital videos are available for public to search and share. Most of the existing video databases are text-based, which are manually tagged with keywords associated with their contents for conveniently retrieval, and this manual work are laborious and subjective. However, the content-based video retrieval analyzes the video contents rather than the manually tags, and it is more efficient and objective. The caption is the textual version of video contents (like dialog and subtitle, etc.) to help viewers to follow, and because of its close relationship with the video contents, the research of automatic caption detection contributes to video contents analysis and content-based retrieval.

Automatic caption detection from videos is a challenging work [1]. Most of the recent related works can be classified into four kinds [2]: edge-based, texture-based, corner-based and stroke-based. The edge-based approach [3–6] extracts the candidate caption regions from the edge map, which is often straightforward and fast, but more false alarms in nontext regions with intensive edges. The texture-based method [7, 8] takes the text as a kind of special texture, and uses different texture descriptors and machine learning techniques to classify text and nontext. But it is time-consuming and training data-dependent. The corner-based approach [9] is inspired by the observation that the distribution of the corner points in text regions is dense and orderly. In literature [9], a fixed threshold is used to binarize the Harris response map to get the corner points of

the text regions. But for various kinds of videos, a fixed threshold is improper. The stroke-based method [2] is based on "two-side edges" feature of character stroke. A stroke-like edge filter and video temporal feature are combined to detect the caption in literature [2].

In this paper, a novel phase-based static caption detection approach is proposed. From Oppenheim's previous work [10], the phase-only image (normalize the magnitude without changing phase, which is called phase-only Fourier synthesis) and magnitude-only image (zero the phase without changing magnitude, which is called magnitude-only Fourier synthesis) represents the different information of an image. The phase-only image, which captures the edge information, is the foundation of image reconstruction. And it has been used in different contexts and applications [11, 12]. As shown in Fig. 1, the captions are much more salient than the background (nontext regions) in the phase-only image, so the phase-only Fourier synthesis can be applied in caption detection research.



(a) (b) (c)

Fig. 1. The Phase-only Image and Magnitude-only Image Represents Different Information of An Image. (a) a video frame with caption, (b) magnitude-only image (after histogram equalization for better visualization), (c) phase-only image.

Our proposed phase-based method mainly consists of two steps. In the first step, the caption saliency map is generated by phase-only Fourier synthesis based on the three features (edge-strong, salience and edge-intensive, detailed in Section 2.1) of caption region. Suppose that there are only two classes, text and nontext regions, in the caption saliency map, and the difference between them is significant. Then the candidate caption regions are extracted from the saliency map by simply clustering or binarization because of the obvious difference between the two regions. In the second step, the false positive regions (nontext regions) among the candidate caption regions can be refined by several text shape features.

The rest of this paper is organized as follows. In Section 2, the phase-based caption detection algorithm is presented, and the comparison experiments with two existing methods [6, 9] are shown and analyzed in Section 3. Finally, conclusions and future work are discussed in Section 4.

2 The Proposed Phase-based Caption Detection Algorithm

The proposed algorithm mainly consists of two processes: candidate caption region detection and candidate caption region refinement, and the flowchart of our algorithm is illustrated in Fig. 2.

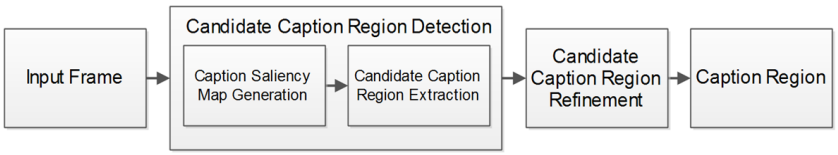


Fig. 2. Flowchart of the Proposed Caption Detection Algorithm.

2.1 Candidate Caption Region Detection

Text region in a video frame always has three following features [3, 4, 9].

1. Edge-strong
   Text or characters are composed by strokes, which are edge-strong, so the text regions are edge-strong.
2. Salience
   The function of text is to deliver message, so the text regions are always high contrast against the background, some kind of saliency, to absorb attention.
3. Edge-intensive
   Characters do not appear alone but together with other characters, so the text regions, which are composed by edge-strong characters, are edge-intensive.

The caption saliency map generated by our proposed approach is following the three features mentioned above.

A. Caption Saliency Map Generation. The caption saliency map is generated by phase-only Fourier synthesis and Gaussian lowpass filter. After graying, the gray image is applied the Discrete Fourier Transform (DFT, Eq. 1),

$$\begin{aligned} F(u,v) &= DFT(f(x,y)) \\ &= R(u,v) + jI(u,v). \end{aligned} \tag{1}$$

Where $j=\sqrt{-1}$, *f(x, y)* is the input gray image, and *F(u, v)* is the DFT of the *f(x, y)*. *R(u, v)* and *I(u, v)* are real and imaginary part of *F(u, v)* respectively.

After the DFT, the image is transformed from spatial domain into frequency domain. Keep the spectral phase and normalize the spectral magnitude. Then apply a Gaussian lowpass filter on the normalized spectral magnitude. At last, do

the Inverse Discrete Fourier Transform (IDFT) on Gaussian filtered phase-only frequency spectrogram, and the caption saliency map is generated (Eq. 2),

$$\begin{aligned} &P_{caption}(u,v) = \arctan\left(I(u,v)/R(u,v)\right), \\ &M_{caption}(u,v) = 1 \times G(u,v), \\ &F_{caption}(u,v) = M_{caption}(u,v) \times e^{jP_{caption}(u,v)}, \\ &f_{caption}(x,y) = IDFT(F_{caption}(u,v)). \end{aligned} \tag{2}$$

Where $P_{caption}(u,v)$ is the phase of the caption saliency map, which equals to the phase of the $F(u,v)$. $M_{caption}(u,v)$ is the magnitude of the caption saliency map, which equals to Gaussian filtered normalized magnitude. $G(u,v)$ is a Gaussian lowpass filter, and $F_{caption}(u,v)$ is Gaussian filtered phase-only frequency spectrogram. $f_{caption}(x,y)$ is the caption saliency map, which is the IDFT of $F_{caption}(u,v)$.

Fig. 3. Phase-only Image and Caption Saliency Map. (a) input image, (b) phase-only image, (c) caption saliency map.

Comparing with the phase-only Fourier synthesis, our approach applies an additional Gaussian filter on the normalized spectral magnitude. After filtering, the spatial pixel contains its Gaussian weighted neighboring information, and this process has many advantages as follows,

1. Because the text regions are edge-intensive, the characters are enhanced by their neighboring characters;
2. Because the nontext is not so edge-intensive as the text, the nontext regions are weaken by their neighboring low response regions;
3. Some noise in the phase-only image are removed.

As shown in Fig. 3, the text regions have a significantly higher response in the caption saliency map (Fig. 3(c)) than the phase-only image (Fig. 3(b)), and the differences between text and nontext are more obvious in the saliency map. So the candidate caption regions are extracted from the caption saliency map instead of the phase-only image in the following step. For convenient process, the saliency map is mapped to 0-255.

B. Candidate Caption Region Extraction. Use K-means to classify the caption saliency map into two clusters with the center *C1* and *C2*. The larger center (*Larger_Center*) between *C1* and *C2* indicates the text, and the smaller one (*Smaller_Center*) indicates the nontext. If the distance between the two centers is large (Eq. 3), the text exists. If not, the text does not exist.

$$Smaller\_Center \leq dist \times Larger\_Center, dist \in (0, 1). \tag{3}$$

Because K-means is a relatively time-consuming algorithm, it is not suitable for real-time caption detection. A fast adaptive binarization method OTSU [13] can replace K-means because of the significant difference between the two regions. The pixels greater than the threshold of OTSU belong to the text region, and the others belong to the nontext region.

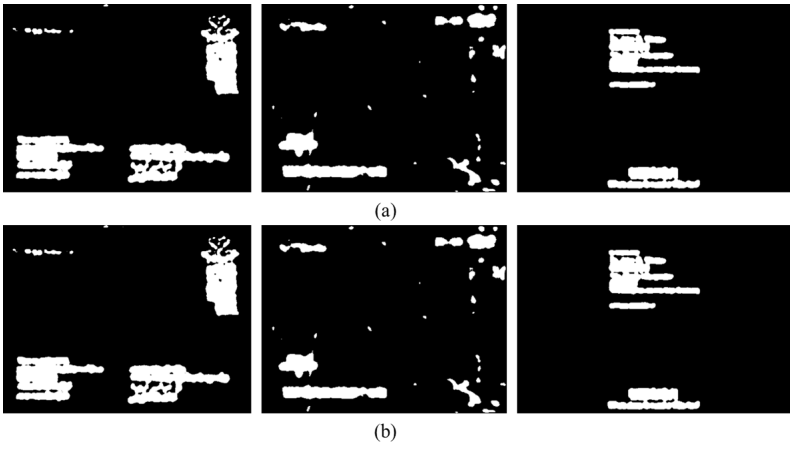


Fig. 4. Caption Region Extraction. (a) using K-means, (b) using OTSU.

Fig. 4(a) and Fig. 4(b) are the caption extraction results of Fig. 3 using K-means and OTSU respectively. The results of these two methods are almost

the same, that's because the differences between text and nontext are obvious in the saliency map. The computing speed of OTSU is much faster (detailed in Section 3.2). But for the nontext frames, K-means gives a lower misclassification rate (detailed in Section 3.1).

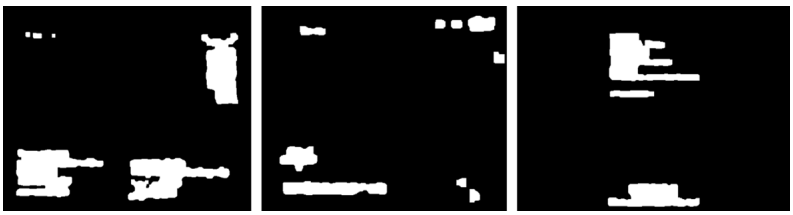


Fig. 5. After Morphological Operation.

At last, the morphological operation is used to remove small false alarms, and fill the holes (Fig. 5).

2.2 Candidate Caption Region Refinement

Among the candidate caption regions, there are still some false alarms. The following four text region shape features [6, 9] are used in the refinement,

1. Area (R_a): R_a is the area of the candidate region, and equals to the amount of "One" pixels of the region. For better visualization, the area of the caption regions is always large, so the region with a small R_a is removed.
2. Saturation (R_s): R_s equals to R_a divided by R_b (Eq. 4), and R_b is the area of the bounding box of the corresponding candidate region. The caption region is often a nearly rectangle shape, so the R_a is close to R_b, and the R_s of caption region is close to 1.

$$R_s = R_a / R_b. \tag{4}$$

3. Orientation (R_o): R_o is defined as the angle between the x-axis and the major axis of the ellipse that has the same second-moments as the region. Because the caption region is often a horizontal or vertical bar, the $|R_o|$ of caption region approximately equals to 0 degree or 90 degree.
4. Edge Density (R_e): R_e equals to $Edge$ divided by R_a (Eq. 5), and the $Edge$ is the sum of the binary Sobel edge map within the corresponding region. Because the caption region is edge-strong and edge-intensive, the R_e of caption region is always high, and the region with a small R_e is removed.

$$R_e = Edge / R_a. \tag{5}$$

After the refinement, nontext regions can be removed from the candidate regions effectively, while the text regions are kept (Fig. 6).

It is clear that our proposed algorithm is language-independent and can be applied in the multilingual video. The left-most frame of Fig. 6, which contains at least 11 different kinds of languages, is a good example to prove it.

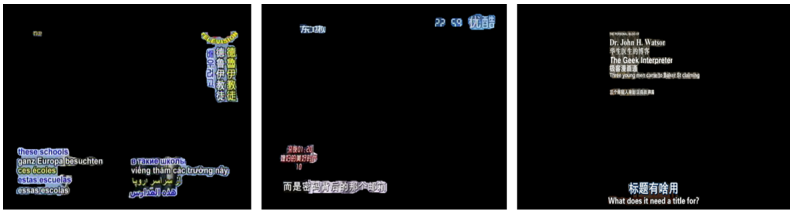


Fig. 6. Final Results After the Refinement.

3 Experimental Results

There is no standard public benchmark dataset specialized for caption detection [2, 6], so a variety of video frames are selected for the performance comparison experiments. Most of frames are from the VideoData subset of public dataset—Research Asia Multimedia 1.0 (MSRA-MM 1.0) [14], and the rest are collected from the various kinds of videos. Our dataset contains news, movies, cartoons, sports programs, etc. The languages are multilingual, including 319 Chinese Character frames (like Chinese and Japanese, called Chinese for short), 321 Latin Alphabet frames (like English and French, called Latin for short), 307 mixed language frames and 200 nontext frames. The amount of frames is 1147.

The comparison experiments are implemented with two existing methods, one is *Laplacian_Filter* based method [6], the other is *Harris* based method [9]. Our proposed approach has two different applications, one is *Phase_K-means*, the other is *Phase_OTSU*. The parameter "*dist*" used in *Phase_K-means* is 0.35. For a better comparison, all the four methods use the same caption refinement method mentioned in Section 2.2. And the Table 1 details the range of the four text region shape features.

Table 1. Range of Four Text Region Shape Features

Feature	$\boldsymbol{R_a}$	$\boldsymbol{R_s}$	$\boldsymbol{\lvert R_o \rvert}$	$\boldsymbol{R_e}$
Range	≥ 50	≥ 0.5	$[0°, 15°] \| \| [75°, 90°]$	≥ 1.5

The block level performance measurements [6] are used in comparison experiments, and the following categories are definitions of the detected block,

- Actual Text Blocks (ATB): The number of true text blocks (count manually);
- Truly Detected Block (TDB): A block that contains at least one character;
- Falsely Detected Block (FDB): A block that does not contain any character;
- Text Block with Missing Data (MDB): A block that misses more than 20% of a text line.

The performance measurements are defined as follows,

- Recall (R) = TDB / ATB;
- Precision (P) = TDB / (TDB+FDB);
- F-measure (F) = 2*P*R / (P+R);
- Misdetection Rate (MDR) = MDB / TDB.

3.1 Experiment on Caption Detection Performance

Fig. 7 shows the performance of the four methods on the each subsets. In the three texts-exist subsets, for the recall, both of our proposed phase-based methods are higher than the other two existing methods. *Phase_OTSU* has the highest recall in Chinese Subset. *Phase_K-means* has the highest recall in both Latin and Mixed Language Subset. For the precision, the two proposed methods rank the second and the third, but the gap with the highest *Laplacian_Filter* is slight. For the F-measure, the two proposed methods are both higher than the other two existing methods. And *Phase_K-means* has the highest F-measure in both Chinese and Latin Subset. *Phase_OTSU* has the highest F-measure in Mixed Language Subset. For the MDR, *Phase_K-means* is the lowest in all three texts-exist subsets. And *Phase_OTSU* ranks the second. Fig. 8 shows some results of the four methods on the three texts-exist subsets respectively.

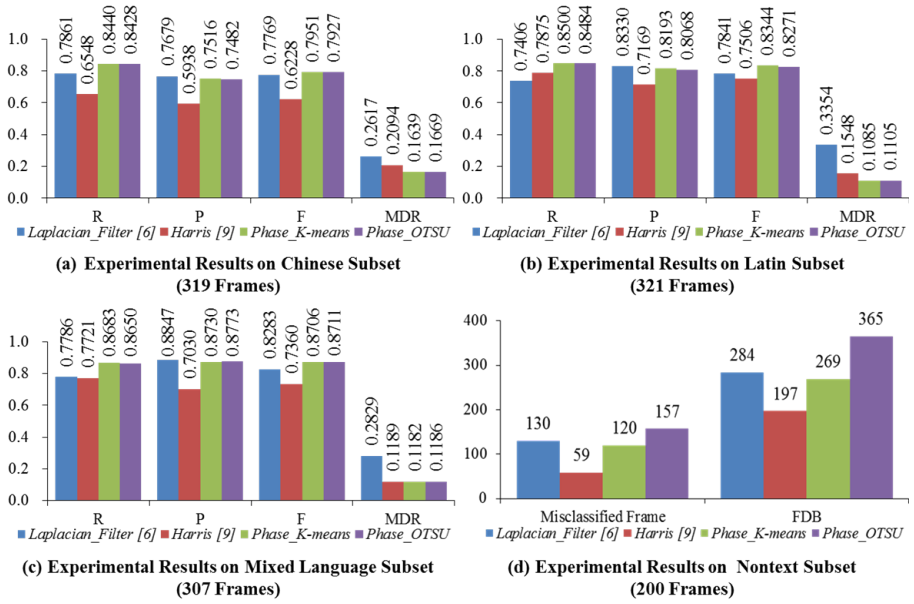
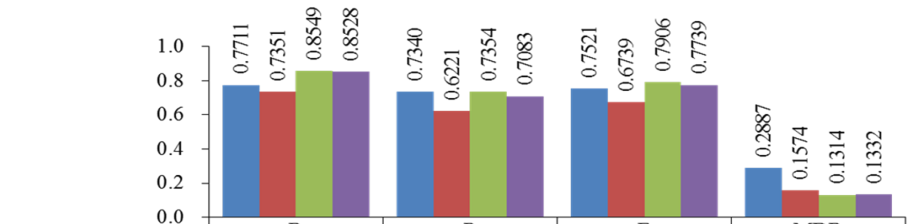


Fig. 7. Experimental Results on Each Subset.

For performance on the nontext subset, misclassified frame means mistaking the nontext frame as the texts-exist frame. Among the four methods, the *Harris*

Fig. 8. Example of Comparison Experiments on three Texts-exist Subsets. (a) input frames, (b) *Laplacian_Filter* [6], (c) *Harris* [9], (d) *Phase_K-means*, (e) *Phase_OTSU*.

has the lowest misclassified frame rate and FDB. The *Phase_K-means* ranks the second, that's because *Phase_K-means* has a parameter "*dist*" to measures the distance between the centers of text and nontext clusters. And this parameter can identify whether the text contains or not.



	R	P	F	MDR
Laplacian_Filter [6]	0.7711	0.7340	0.7521	0.2887
Harris [9]	0.7351	0.6221	0.6739	0.1574
Phase_Kmeans	0.8549	0.7354	0.7906	0.1314
Phase_OTSU	0.8528	0.7083	0.7739	0.1332

Fig. 9. Experimental Results on the Whole Dataset (1147 frames).

Fig. 9 shows the performance of the four methods on the whole dataset. The proposed *Phase_K-means* gives the best comprehensive performance (F-measure) and the lowest MDR. Fig. 10 shows some other results of the two proposed approaches on the whole dataset.

3.2 Experiment on Time Cost

The CPU of the test computer is Intel Core i5-2400, 3.10GHz, and the RAM is 2GB, 1333MHz. All the four methods are programming in Matlab code. Table 2 shows the average time cost. The *Harris* and proposed *Phase_OTSU* are apparently faster than the other two K-means using methods (*Laplacian_Filter* and *Phase_K-means*), and the *Phase_OTSU* is a little slower than the *Harris* method, but not obvious. The two K-means using methods are slow because K-means is a relatively time-consuming method. And our proposed *Phase_K-means* is a little faster than the existing *Laplacian_Filter*.

Table 2. Average Time Cost

Method	Time Cost (second)
***Laplacian_Filter* [6]**	0.2769
***Harris* [9]**	0.0625
Phase_K-means	0.2511
Phase_OTSU	0.0651

Fig. 10. Some Results of the two Proposed Approaches. (a) input frames, (b) *Phase_K-means*, (c) *Phase_OTSU*.

4 Conclusion and Future Works

In this paper, a novel phase-based caption detection approach is proposed. In the detection step, the candidate caption regions are extracted from the caption saliency map mainly generated by phase-only Fourier synthesis. In the refinement step, the nontext regions are refined by four shape features. The comparison experimental results with two existing methods have verified the effectiveness and efficiency of our proposed approach. Furthermore, our approach is language-independent, and can be applied in the multilingual videos.

Our proposed approach has two different applications, one is *Phase_K-means*, the other is *Phase_OTSU*. *Phase_K-means* offers a better comprehensive performance (F-measure), and a lower misclassified frame rate in the nontext frames. However, *Phase_OTSU* offers a faster speed, and be suitable for captioned video frames, like news and sports programs.

In the future, the proposed phase-based method can be extended on moving caption detection or on document image analysis.

Acknowledgement. This work was supported by NSF of China (Grand No. 90920008). We would like to thank Dr. Li Ce for many helpful suggestion.

References

1. Kwang In Kim, Keechul Jung, Jin Hyung Kim: Texture-Based Approach for Text Detection in Images Using Support Vector Machines and Continuously Adaptive Mean Shift Algorithm. IEEE Transactions on Pattern Analysis and Machine Intelligence **25** (2003) 1631–1639
2. Xiaoqian Liu, Weiqiang Wang: Robustly Extracting Captions in Videos Based on Stroke-Like Edges and Spatio-Temporal Analysis. IEEE Transactions on Multimedia **14** (2012) 482–489
3. Palaiahnakote Shivakumara, Weihua Huang, Chew Lim Tan: An Efficient Edge based Technique for Text Detection in Video Frames. The Eighth IAPR Workshop on Document Analysis Systems (2008) 307–314
4. Palaiahnakote Shivakumara, Trung Quy Phan, Chew Lim Tan: Video Text Detection Based on Filters and Edge Features. IEEE International Conference on Multimedia and Expo (2009) 514–517
5. Trung Quy Phan, Palaiahnakote Shivakumara, Chew Lim Tan: A Laplacian Method for Video Text Detection. International Conference on Document Analysis and Recognition (2009) 66–70
6. Palaiahnakote Shivakumara, Trung Quy Phan, Chew Lim Tan: A Laplacian Approach to Multi-Oriented Text Detection in Video. IEEE Transactions on Pattern Analysis and Machine Intelligence **33** (2011) 412–419
7. Ilktan Ar, M. Elif Karsligil: Text Area Detection in Digital Documents Images Using Textural Features. International Conference on Computer Analysis of Images and Patterns (2007) 555–562
8. Qixiang Ye, Qingming Huang, Wen Gao, Debin Zhao: Fast and Robust Text Detection in Images and Video Frames. Image and Vision Computing **23** (2005) 565–576
9. Xu Zhao, Kai-Hsiang Lin, Yun Fu, Yuxiao Hu, Yuncai Liu, Huang T.S.: Text From Corners: A Novel Approach to Detect Text and Caption in Videos. IEEE Transactions On Image Processing **20** (2011) 790–799
10. A.V.Oppenheim, J.S.Lim: The Importance of Phase in Signals. Proceedings of the IEEE **69** (1981) 529–541
11. Chenlei Guo, Qi Ma, Liming Zhang: Spatio-temporal Saliency Detection Using Phase Spectrum of Quaternion Fourier Transform. IEEE Computer Society Conference on Computer Vision and Pattern Recognition (2008) 295–302
12. Dror Aiger, Hugues Talbot: The Phase Only Transform For Unsupervised Surface Defect Detection. IEEE Conference on Computer Vision and Pattern Recognition (2010) 295–302
13. N. Otsu: A Threshold Selection Method From Gray Level Histograms. IEEE Transactions on Systems, Man and Cybernetics **9** (1979) 62–66
14. W.Meng, L.J Yang, X.S. Hua: MSRA-MM: Bridging Research and Industrial Societies for Multimedia Information Retrieval. Microsoft Technical Report, MSR-TR-2009-30 (2009)