Text From Corners: A Novel Approach to Detect Text and Caption in Videos

Xu Zhao, Kai-Hsiang Lin, Yun Fu, *Member, IEEE*, Yuxiao Hu, *Member, IEEE*, Yuncai Liu, *Member, IEEE*, and Thomas S. Huang, *Life Fellow, IEEE*

***Abstract*—Detecting text and caption from videos is important and in great demand for video retrieval, annotation, indexing, and content analysis. In this paper, we present a corner based approach to detect text and caption from videos. This approach is inspired by the observation that there exist dense and orderly presences of corner points in characters, especially in text and caption. We use several discriminative features to describe the text regions formed by the corner points. The usage of these features is in a flexible manner, thus, can be adapted to different applications. Language independence is an important advantage of the proposed method. Moreover, based upon the text features, we further develop a novel algorithm to detect moving captions in videos. In the algorithm, the motion features, extracted by optical flow, are combined with text features to detect the moving caption patterns. The decision tree is adopted to learn the classification criteria. Experiments conducted on a large volume of real video shots demonstrate the efficiency and robustness of our proposed approaches and the real-world system. Our text and caption detection system was recently highlighted in a worldwide multimedia retrieval competition, *Star Challenge*, by achieving the superior performance with the top ranking.**

***Index Terms*—Caption detection, Harris corner detector, moving caption, optical flow, text detection, video retrieval.**

I. Introduction

THE AMOUNT of archival multimedia data is increasing dramatically with the extensive development of Internet and the broad use of digital video hardware. The availability of large databases leads to vast potential demands for efficient algorithms to retrieve, archive, index and locate desired content in large volume of image and video data. During the past decade, a lot of techniques have been put forward to extract the semantic and content information from image and videos [1]–[3]. Among the techniques, text detection based approaches are of particular interests to many applications due to the rich information contained in text and caption [4]–[8]. Specially, text and caption in images can provide direct high level semantic information such as program name, speaker name, speech content, special announcements, data, time, scene location, sports scores and so forth. These information usually is hard to acquire from other high level image features like human action, face, scene and vehicles, etc. Texts from images also enable useful real-world applications such as automatic annotation and indexing of images [9].

However, detecting text from videos is a challenging task which often suffers from appearance variations of text, low contrast and complex background [6]. Most existing approaches can be generally classified into three categories [4], [8], namely, *texture based methods*, *connected component based methods*, and *edge based methods*. Texture based methods [6], [7], [10]–[12] treat the text region as a special type of texture with distinct textural properties that can distinguish them from the background. The techniques that are used to extract texture features include support vector machine (SVM) [6], Wavelet [12], FFT [13], spatial variance [10], and neural networks [4], [5], [14], etc. Texture based approaches are efficient in dealing with complex background with dissimilar textual structure to the text regions. But the computational complexity restricts its applications in large databases.

Connected component based methods [15]–[17] segment an image into a set of connected components and successively merge the small components into larger ones. The final connected components are classified as text or background by analyzing their geometrical characteristics. This approach is built on the basic assumption that the text is represented with a uniform color, therefore, it sometimes appears with the color feature together [18]. While they are efficient when the background mainly contains uniform regions, this kind of approaches encounter difficulties when the text is noisy, multicolored and textured.

Edge based methods [19]–[23] utilize the structural and geometry properties of character and text. Since characters are composed of line segments and text regions contain rich edge information, usually, an edge extractor is applied for the edge detection and a smoothing operation or a morphological operator is used in the merging stage [8]. This kind of methods are effec-

Manuscript received January 28, 2009; revised June 15, 2009; accepted July 04, 2010. Date of publication August 19, 2010; date of current version February 18, 2011. This research was supported in part by the U.S. Government VACE program, the National Science Foundation under Grant CCF 04-26627, the National Basic Research Program of China 973 Program under Grant 2011CB302203, and the NSFC Program 60833009, 60975012. The associate editor coordinating the review of this manuscript and approving it for publication was Dr. Xuelong Li.

X. Zhao was with Beckman Institute for Advanced Science and Technology, University of Illinois at Urbana-Champaign, Urbana, IL 61801, USA. He is now with the School of Electronic, Information and Electrical Engineering, Shanghai Jiao Tong University, Shanghai 200240, China (e-mail: zhaoxu@sjtu.edu.cn; zhaoxuhong@gmail.com).

K.-H. Lin and T. S. Huang are with the Beckman Institute for Advanced Science and Technology, University of Illinois at Urbana-Champaign, Urbana, IL 61801 USA (e-mail: klin21@uiuc.edu; huang@ifp.uiuc.edu).

Y. Fu is with the Department of Computer Science and Engineering, SUNY at Buffalo, Buffalo, NY 14260 USA (e-mail: yunfu@buffalo.edu).

Y. Hu is with the Microsoft Live Search, Redmond, WA 98052 USA (e-mail: Yuxiao.Hu@microsoft.com).

Y. Liu is with the School of Electronic, Information and Electrical Engineering, Shanghai Jiao Tong University, Shanghai 200240, China (e-mail: whomliu@sjtu.edu.cn).

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TIP.2010.2068553

1057-7149/$26.00 © 2011 IEEE

Authorized licensed use limited to: University of New South Wales. Downloaded on August 17,2020 at 01:49:45 UTC from IEEE Xplore. Restrictions apply.

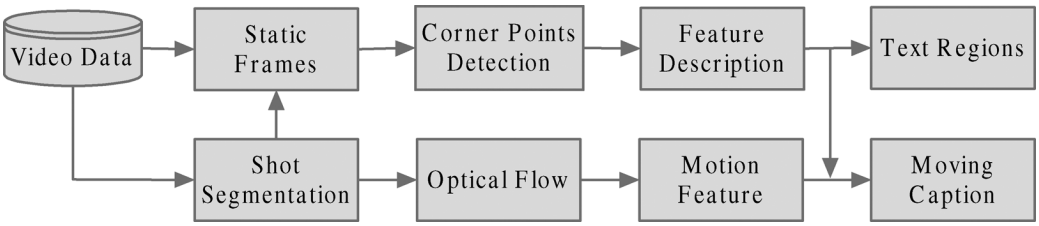


Fig. 1. Framework of our text and moving caption detection system.

tive in detecting text regions when the other parts in the image do not have too many strong edges. But for the scenario where the background has similar strong edge distributions as the text regions, they are not robust enough for a reliable performance.

Actually, in order to efficiently detect texts, we need to analyze the discriminative properties of text and its basic unit, character. As the special content in visual media, text and character originally are created to record and spread human ideas in writing systems.[1] They are printed in the visual media such as paper and screen, therefore, the readability is a basic requirement. Observing the structure of characters, no matter what language in consideration, we can find that there are many corners within the character especially in the array of characters. Taking along rich visual information, these salient points can attract readers' attention to a large extent.

In this paper, we propose a novel corner based approach to detect text and captions in video frames.[2] In nature, our approach is inspired by the observation that there exist dense and orderly presences of corner points in characters, especially in text and captions. In the most related work [25], the corners and edges are combined to locate the text regions in video frames. Although corner features are also introduced into text detection in [25], further analysis and description to the shape properties of detected regions are deficient. Our approach is quite different from the existing techniques. We propose to describe the text regions with the discriminative features, from which the nontext regions formed by the corner points appeared in the background can be filtered out efficiently. The flexibility of the feature extraction schemes allows smooth adaption of our approach to different applications and databases. Especially, language independence is an important advantage of our method.

Based upon the proposed text detection approach, we specially develop a novel method to detect moving captions in videos. In the algorithm, the motion features, extracted using optical flow, and text features are combined to detect the moving captions appeared in the video shots. The framework of our text and moving caption detection system is illustrated in Fig. 1. It consists of three main parts. Corner detection and feature description are the core parts, by which the text regions in static images are detected and located. The other two parts are serving for the moving caption detection. Optical flow based motion feature extraction describes the motion characteristic of the key frames within the video shots. The part for the combination of text feature and motion feature outputs the final detection results of moving captions.

[1]A writing system is a type of symbolic system used to represent elements or statements expressible in language [24].

[2]In this paper, caption is a kind of text actually. Caption is specially pointed out here as it is the most frequent text pattern compare to other text patterns.

Experiments conducted on a large volume of real video shots demonstrate the efficiency and robustness of our proposed approaches and the real-world system. It is worth mentioning that our prototype system was tested in the *Star Challenge*, a multimedia retrieval competition comprising over fifty teams from around the globe. We got the first and second place in the consecutive elimination rounds for the video retrieval tasks, respectively.

The remaining of this paper is organized as follows. We introduce the corner based features and criteria for text detection from static images in Section II. The algorithms for moving caption detection is described in Section III. The experimental results on different videos in a large database are shown in Section IV. We finally conclude the paper in Section V.

II. Features for Text Detection

In this section, we describe the features for text detection in videos. We choose the corner points as the essential feature by viewing the following three-fold advantages of corner points for text detection.

1) Corners are frequent and essential patterns in text regions. As an image feature, corner is more stable and robust than other low level features. Therefore, the impacts of background noises can be eliminated to a large extent.
2) The distributions of corner points in text regions are usually more orderly in comparison to the nontext regions. Therefore, the unordered nontext corner points can be filtered out according to our designed features.
3) The usage of corner points generates more flexible and efficient criteria, under which the margin between text and nontext regions in the feature space is discriminative.

In our algorithm, we use Harris corner detector to extract the corner points from images. We extract and describe the features by computing the shape properties of the regions containing the detected corner points. These shape properties are used to construct the basic criteria to detect text and captions. The flexibility of the criteria allows the adaption of our algorithm to different applications.

A. Corner Points Extraction

Corner points are the image features that are usually more salient and robust than edges for pattern representation. Corner detection is frequently used in motion detection, image matching, tracking, image mosaicking, 3-D modeling and so forth. A corner can be defined as the intersection of two edges or a point where there are two dominant and different edge directions in a local neighborhood of the point. In our

Authorized licensed use limited to: University of New South Wales. Downloaded on August 17,2020 at 01:49:45 UTC from IEEE Xplore. Restrictions apply.

(a) (b) (c) (d)

Fig. 2. Sample image frames with the corresponding detected corner points.

implementation, we use Harris corner detector [26] to extract the corner points.

The Harris corner detector is a popular interest point detector due to its strong invariance to rotation, scale, illumination variation, and image noise [27]. It is based upon the local auto-correlation function of a signal, which measures the local changes of the signal with patches shifted by a small amount in different directions. Suppose we have a gray scale 2-D image I. Consider taking an image patch over the window $W(u,v)$ and shifting it by $(\delta x, \delta y)$. The change produced by the shift is given by

$$E(\delta x, \delta y) = \sum_{W} [I(u+\delta x, v+\delta y) - I(u,v)]^2. \quad (1)$$

The shifted image is approximated by a Taylor expansion truncated to the first-order terms

$$I(u+\delta x, v+\delta y) \approx I(u,v) + [I_x(u,v)\, I_y(u,v)]\, [\delta x\, \delta y]^T \quad (2)$$

where I_x and I_y denote the partial derivatives in x and y directions, respectively. Substituting approximation (2) into (1) yields

$$E(\delta x, \delta y) = [\delta x \quad \delta y]\mathbf{C}[\delta x \quad \delta y]^T \quad (3)$$

where the Hessian matrix $\mathbf{C}$ captures the intensity structure of the local neighborhood

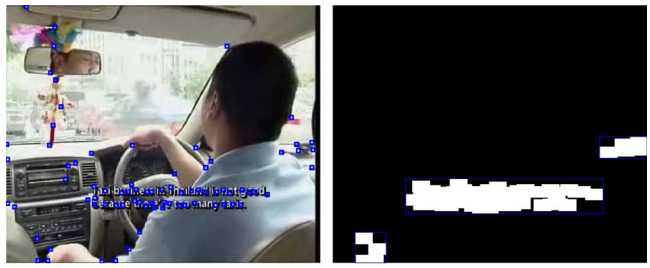
$$\mathbf{C} = \begin{bmatrix} \sum\limits_{W}(I_x(u,v))^2 & \sum\limits_{W} I_x(u,v)I_y(u,v) \\ \sum\limits_{W} I_x(u,v)I_y(u,v) & \sum\limits_{W}(I_y(u,v))^2 \end{bmatrix}. \quad (4)$$

Let λ_1, λ_2 be the eigenvalues of matrix $\mathbf{C}$. The eigenvalues form a rotationally invariant description. If both λ_1 and λ_2 are large and distinct positive values, a corner is determined as a detection target. In this scenario, the local auto-correlation function is sharply peaked and the shifts in any direction will result in a significant increase. To avoid the explicit eigenvalues decomposition, Harris and Stephens [26] design the response function

$$f_R = \lambda_1\lambda_2 - \kappa(\lambda_1+\lambda_2)^2 = \det \mathbf{C} - \kappa\, \mathrm{trace}^2(\mathbf{C}) \quad (5)$$

where κ is a tunable sensitivity parameter. As a measurement, f_R is positive in the corner region, negative in the edge region, and small in the flat region.

In Fig. 2, we show some image samples with detected corner points. It can be seen that the corners are the most frequent concurrent patterns in text and captions.



(a) (b)

Fig. 3. Area feature of the regions. (a) Corner points superimposed on the original image. (b) Top three regions measured with area feature.

B. Feature Description

After extracting the corner points, we need to compute the shape properties of the regions containing corner points, so that the system can make the decision to accept the regions as text or not. To this end, we firstly perform image morphology dilation on the binary corner image. In doing so, the separate corner points that are close to each other can be merged into a whole region. In the text and captions, the presence of corner points are dense because characters do not appear alone but together

Authorized licensed use limited to: University of New South Wales. Downloaded on August 17,2020 at 01:49:45 UTC from IEEE Xplore. Restrictions apply.

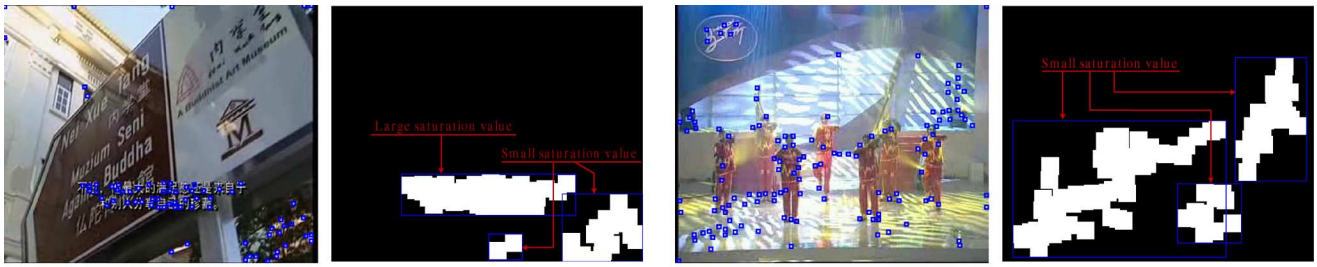


Fig. 4. Saturation feature of the regions. (a) Text region and nontext region with different R_{s} values. (b) Nontext regions with large R_{a} value and small R_{s} value.

with other characters and usually regularly placed in a horizontal string. Therefore, the text can be effectively detected by figuring out the shape properties of the formed regions. It should be pointed out here that dense corners may also appear in the nontext regions but they usually are unordered and can be filtered out using our designed features.

We use the following five region properties as the features to describe text regions: area, saturation, orientation, aspect ratio and position. For the convenience of description, we represent the five features as R_{a}, R_{s}, R_{o}, R_{as} and R_{c}, respectively. Another feature need to be clarified here is the bounding box. We define the bounding box as the smallest rectangular that completely encloses the corner points formed regions.

1) Area: The area of a region is defined as the number of foreground pixels in the region enclosed by a rectangle bounding box, see Fig. 3. Area is the basic feature for text detection. The small regions generated by the disorderly corner points can be easily filtered out according to the area measurement. Here, we define the disorderly corner points as those distributing randomly and irregular in image frames, therefore, will be hard to describe them with some criteria. In Fig. 3, we display the top three regions measured by the area. Except for the biggest one, another two regions are actually formed by nontext corner points, which can be easily discarded by measuring their areas.

2) Saturation: In our context, the saturation specifies the proportion of the foreground pixels in the bounding box that also belong to the region, which can be calculated by

$$R_{\mathrm{s}} = \frac{R_{\mathrm{a}}}{R_{\mathrm{B}}} \ (\in (0,1))$$

where R_{B} represents the whole region enclosed by the bounding box. This feature is very important for the cases where the nontext corner points can also generate the regions with relative large R_{a} values. Although with improper large R_{a}, but fortunately, this kind of regions usually have small R_{s} value originated from the disorderly distribution of detected corner points. Therefore, we can filter out the regions with the saturation feature. From Fig. 4, we can observe obvious difference of the saturation feature between text regions and nontext regions. In the interior of nontext regions, there are much more background pixels than the text regions.

3) Orientation: Orientation is defined as the angle (ranging from $-90°$ to $90°$) between the x-axis and the major axis of the ellipse that has the same second-moments as the region. In Fig. 5, the red ellipses illustrate the orientation of the regions.

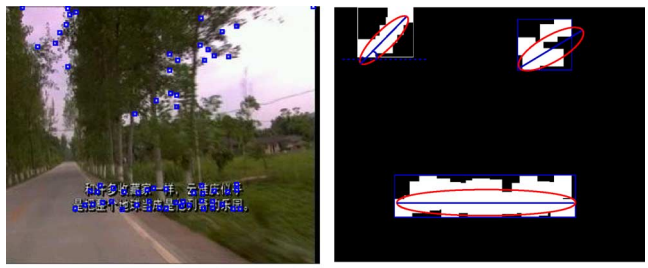


Fig. 5. Orientation feature of the regions. (a) Corner points superimposed on the original image. (b) Text and nontext regions with different orientation angles.

Fig. 6. Position information for the text regions. (a) Corner points superimposed on the original image. (b) Two text regions at the bottom with different position information. The lowest one is caption and another is the introductory text region.

This feature is useful where the nontext regions have relative large R_{a} and small R_{s}, as displayed in Fig. 4(b). It can be flexibly used in caption detection because the regular distribution of corner points in caption generates small $|R_{\mathrm{o}}|$ values.

4) Aspect Ratio: Aspect Ratio of a bounding box is defined as the ratio of its width to its height. In videos, text and captions usually are placed regularly along the horizontal axis. Therefore, a relative large R_{as} value indicates more the presence of captions than a small value. We can utilize this characteristic to filter out some of the false alarms.[3]

5) Position: We describe the position of a region with its centroid. The position information can be used to locate the text regions with specific type and style. For example, position is important to differentiate the captions from other text regions because captions are usually put at the bottom of the image, such as Fig. 6.

[3]In this paper, a false alarm occurs where a nontext object exceeds the detection criteria and is identified as a text object.

Authorized licensed use limited to: University of New South Wales. Downloaded on August 17,2020 at 01:49:45 UTC from IEEE Xplore. Restrictions apply.

Fig. 7. Our designed features for text detection are language independent. The first row shows the detected Korea and English text from the original image and the second row shows the Japanese and Chinese text. In the third row, the detected Arabic is shown.

Generally, we combine the previously mentioned five features to detect text and captions and filter out the false alarms. The flexibility of these features allow effective adaption of our algorithm to different applications. For example, these features can be used to train the classifiers and learning algorithms to find discriminative criteria automatically for different datum and applications. In this work, we did not use the learning based methods to determine feature parameters automatically because labeling the data in such a large database is very tedious and time consuming work. It is also worth noting that our corner formed features are language independent and can be used in the multilingual scenario. It can be seen in Fig. 7.

III. Moving Caption Detection

For some applications, detecting moving text and captions from videos is very important. For example, usually at the end of a movie or TV program, there appear the movie closing credits. A lot of useful information, such as the names of impersonators, the sponsors, the music producer and some advertisement related content, can be found in these captions. In content based video retrieval and classification, moving caption detection can also play an important role. In this section, we propose an approach to detect moving captions based upon the combination of text features introduced in Section II and motion features computed using optical flow. We apply the features to classify videos using decision tree method. The experimental results in Section IV verify the efficacy of our features and algorithm.

A. Optical Flow Based Motion Feature Extraction

In this paper, we use optical flow as our motion features. Here optical flow estimation is used to compute an approximation to the motion field from intensity difference of two consecutive frames.

There are several methods to implement optical flow estimation. In our implementation, we choose the multiresolution Lucas-Kanade algorithm [28], [29], in which there are four main components:

1) Gaussian pyramid construction;
2) motion estimation by Lucas-Kanade algorithm;
3) image warping;
4) coarse-to-fine refinement.

Lucas-Kanade algorithm makes use of the spatial intensity gradient of the images to find a good match using Newton–Raphson iteration. More detailed introduction about this algorithm can be found in [28], [29]. In order to preserve the spatial-temporal information as much as possible, we extract the optical flow feature every five frames for the test videos. These frames are called key frames in the following paragraph.

B. Feature Combination

Before combining the features, we first extract the text bounding box and optical flow for every key frames of each video shot. Then for each pixel on the key frames, we extract two kinds of features:

1) a binary value, which records whether this pixel is within a text region or not;
2) a motion vector of this pixel.

The two features are combined by multiplying. That is, if a pixel belongs to a text region, we record the feature of this pixel by its motion vector; otherwise, we record the feature of this pixel by a zero vector.

This process is shown in Fig. 8. Fig. 8(a) and (e) show the sample frames of input videos containing moving captions. Fig. 8(b) and (f) are the results of text detection for the original frames. The optical flow estimation of the original frames and the consequent frames are shown in Fig. 8(c) and (g). Actually, the output of optical flow estimation is a motion vector. In the figures, this motion feature vector is represented by the color image. Speed is represented by intensity and the motion orientation is represented by hue. In Fig. 8(c) and (g), the color of region containing text is green because the captions

Authorized licensed use limited to: University of New South Wales. Downloaded on August 17,2020 at 01:49:45 UTC from IEEE Xplore. Restrictions apply.

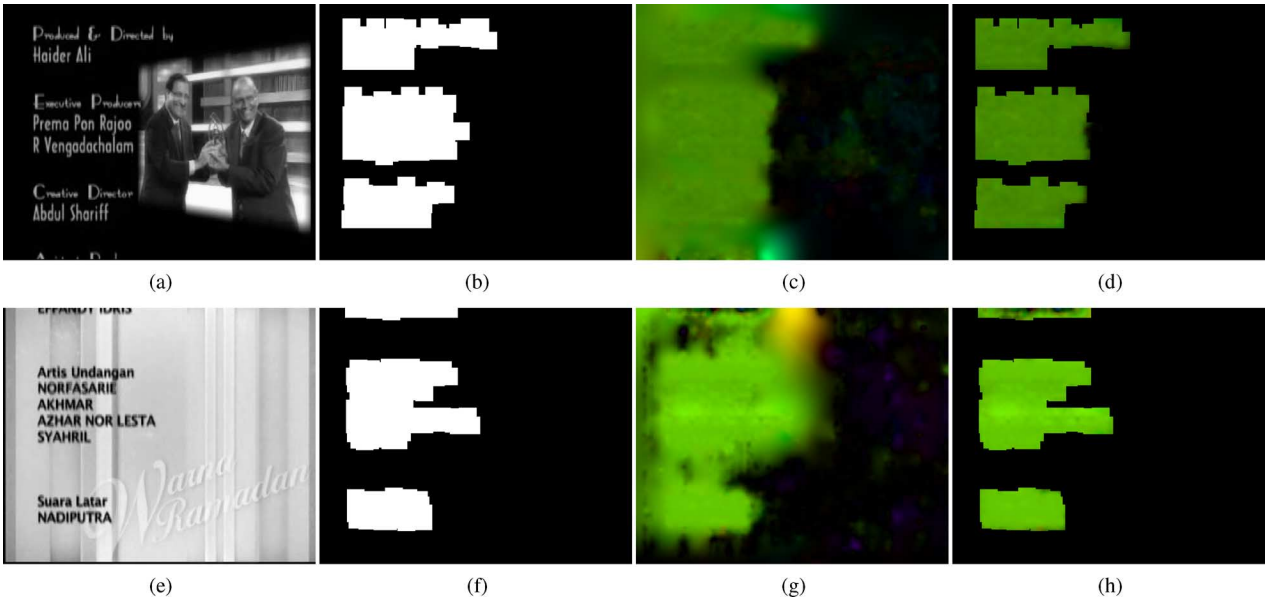


Fig. 8. Feature combination for moving caption detection. (a), (e) Sample original frames of the input videos. (b), (f) Results of text detection. (c), (g) Optical flow estimation of the original frames. (d), (h) Combination of motion feature and text features.

are moving upward which is described by green color in our representation. Fig. 8(d) and (h) show the combination results of Fig. 8(b), (f), (c), and (g), respectively.

C. Decision Tree for Caption Classification

Based upon the moving text features, some machine learning algorithms can be used to detect the moving captions. This is a two class classification problem. In this work, we use the decision tree as the classification algorithm.

Decision tree is a widely used method which works by partitioning the feature space into a set of rectangles and then assigning a simple model (e.g., a constant) in each one. There are several tree-based frameworks and we choose CART [30] in this work. This algorithm includes two main steps: grow the tree, and then prune the tree. The nodes of the tree are set by candidate feature summaries of each video. These features are statistical properties of area or position of the moving text. We use 10-fold cross-validation in our training data to find the best size of the tree and prune the tree in this best size.

In order to use the decision tree efficiently, moving text features of each video should be summarized by a feature vector. However, the features we are using are based upon each key frame rather than a video. Then how to summarize these image-based features into a feature vector which can be used to represent a video will be a problem. Fortunately, the advantages of decision tree allow us to try various statistical properties of image based features and the decision tree can choose the appropriate ones automatically. In our experiments, we have tried several statistical properties such as the mean and variance of locations of moving texts, mean and variance of speed of moving texts. And at last we find the most effective feature, total area of texts moving in the main direction. Therefore, we generate a one node decision tree, by which the threshold to distinguish moving captions is determined.

The procedure to extract total area of texts moving in the main direction is described as follows. Firstly, we quantize the image based motion features into four directions: upward, downward, rightward, and leftward. This quantization is made under the consideration that texts usually move in these four directions. Second, we measure the total area of moving texts in these directions, respectively. The direction with the largest area of moving text is considered as the main direction of this video shot. The total area of moving texts in this direction is the feature used in the decision tree. This feature is based upon two important properties of moving texts:

1) the direction of moving text is stable;
2) the total area of the moving text is usually much larger than the area of subtitles.

IV. Experiments

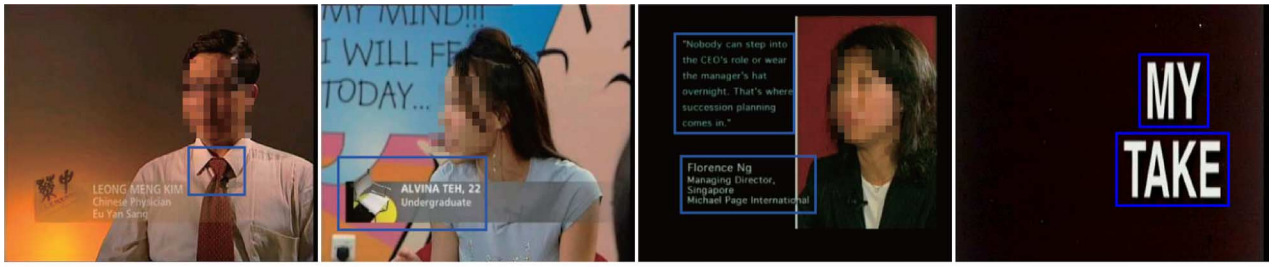
We evaluate the proposed text and caption detection approaches on the data set provided by the sponsor of the multimedia retrieval competition, Star Challenge. The content of the data set ranges over movie segments, TV news, other TV programs and so forth. The videos are encoded in MPEG format with a resolution of 352×288. The original videos are segmented as separate shots. We extract one real key frame and eight pseudo key frames from each shot. The format of the extracted images is JPEG. In the experiments, our text detection approach is tested in both static and dynamic scenario. The performance evaluations are carried out not only in image frames but also in video shots. The moving caption detection algorithm is designed for the dynamic scenarios, therefore, evaluated just on the video shots.

In the video retrieval tasks of this global competition, there are several classes of queries, such as the introductory caption and the moving ending credit, which are heavily dependent upon the results of text detection. In the consecutive elimination

Authorized licensed use limited to: University of New South Wales. Downloaded on August 17,2020 at 01:49:45 UTC from IEEE Xplore. Restrictions apply.

TABLE I
RANGE OF THE FEATURE VALUES FOR BOTH TEXT AND NON-TEXT REGIONS

Feature	R_{a} (Pixel)	R_{s}	R_{o} (Degree)	R_{as}	R_{c}
Text	> 1000	$[0.6, 1]$	$[-10°, 10°]$	> 2	$> 0.5H$ (for caption)
Non-text	$[100, 20000]$	[0.2,0.5]	$[-90°, -20°] \cup [20°, 90°]$	< 2	—



(a) (b) (c) (d)

Fig. 9. Image samples of (a), (b) missing frames and (c), (d) false alarm frames for the category of introductory caption.

rounds including previous text detection related video retrieval tasks, we took the first and second place, respectively.

A. Static Text and Caption Detection

In the static setting, the locations of text and captions in images are fixed and have no temporal motions. We conduct the experiments mainly on the data set with category label called introductory caption. In this category, there are 842 video shots and 7578 image frames in total. There is at least one key frame in a video shot with the captions on it.

In our algorithm, there are two important parameters need to be set carefully. One is the size of mask performing image dilation. Another is the threshold for corner points detection. Basically, these two parameters are determined by the resolution and quality of the videos. In our evaluation, the mask size and the corner threshold are set as 20×20 and 0.2, respectively. Under this setting, the general range of feature values for both text and nontext regions are summarized in Table I, where H means the height of the image. Because the captions usually appear in the bottom half of the image, the R_{c} is taken as $0.5H$. According to the value range, we can make the criteria to detect text and caption flexibly. We use all the five features in the experiments. Generally, we hope to detect as many text regions as possible with a reasonable precision, then remove the false alarms by adopting more strict criteria, which are usually determined by the applications.

We use recall and precision as the metrics to evaluate the performance. The results are summarized in Table II. Our approach detects 798 shots and misses 44 shots. Because all of the shots contain the text and captions, the false alarm is 0. At the frame level, we detect 4289 frames containing text and captions in total with 290 false alarms in it. There are 3289 fames that are recognized as noncaption frames. The number of missing frames is 625. The missing detection is mainly caused by the blur and low contrast quality of text regions, where corner points are seldom detected [see Fig. 9(a)]. Another reason for the missing frames is that the text region and nontext region are connected occasionally, therefore, the shape properties of the text region are changed. In Fig. 9(b), the caption region enclosed by the blue rectangle is not recognized as introductory caption because the R_{s} feature is not qualified. In the experiments, the number of false alarms is relatively small and most of them are caused by the semantic confusion of the category labels. In Fig. 9(c) and (d), although the texts are correctly detected, the frames are identified as false alarms because their labels are not introductory captions. Fig. 10 shows some image samples with the bounding box of detected text and captions.

To further evaluate our approach, we make comparisons with texture based approach. The selected benchmark approach is an SVM based approach [6], in which SVM is used to determine whether the input texture pattern represented by the intensities of raw pixels is text or not. We randomly select 300 images from the category of introductory caption as the initial training samples while others as test samples. Nontext training examples are collected from 200 nontext images in other categories. The size of scan window is set as 13. The results of comparison are shown in Table III. Our approach outperforms a little bit on precision than the texture based approach because the number of false alarm frames is smaller. However, the performance of texture based approach on recall rate is much better than our approach. This is because the additional training process makes the SVM recognize more textual patterns than our approach. Therefore the number of missing samples is reduced. As far as the computational cost is concerned, our approach outperforms almost 15 times than the texture based approach. This is because the SVM based approach needs to scan the images with a sliding window and then input the features to SVM to determine its label. Hence, the time cost is more expensive compared to our approach. The average time spent on each image frame is 0.25 s on a computer with Pentium 4 CPU, 3G Hz clock speed and 1G memory using un-optimized Matlab code. This advantage is very important when handling large scale data set.

B. Moving Caption Detection

For the experiments of moving caption detection, in total 1593 video shots are involved in the training and testing process. They are labeled into two categories: 1) videos contain

Authorized licensed use limited to: University of New South Wales. Downloaded on August 17,2020 at 01:49:45 UTC from IEEE Xplore. Restrictions apply.

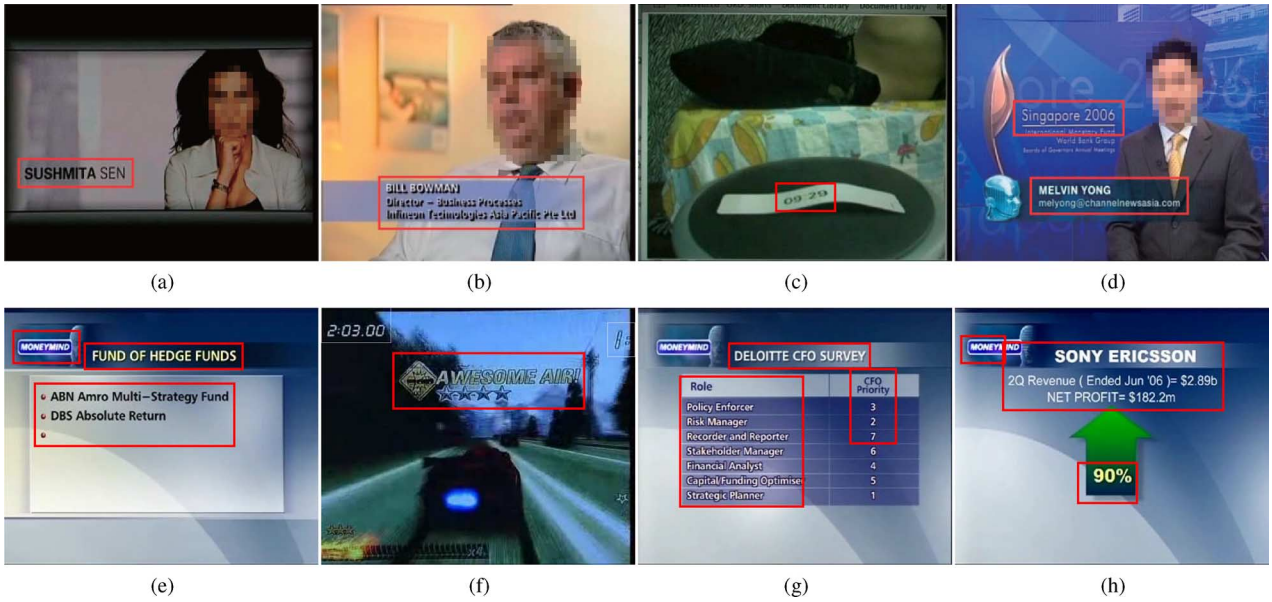


Fig. 10. Image samples with the bounding box of detected text and caption.

TABLE II
PERFORMANCE OF THE STATIC TEXT AND CAPTION DETECTION

	Hits	Misses	False alarms	Precision	Recall
Shot	798	44	0	1	94.77%
Frame	3999	625	290	93.24%	86.48%

TABLE III
PERFORMANCE COMPARISON BETWEEN OUR APPROACH AND TEXTURE BASED APPROACH

	Precision	Recall	Time cost (sec. /frame)
Our approach	93.24%	86.48%	0.25
Texture based	91.35%	93.10%	3.80

moving captions and 2) videos have no moving captions. There are 45 video shots in category 1) and 1548 video shots in category 2). The data set is divided into two parts evenly. Half of the video shots are used for training and another half for testing. Because the numbers of the video shots in these two categories in training set are extremely biased (22 versus 774), we further sub sample category 1) with factor 10 to make the number of training samples in category 1) and 2) are in the same scale (22 versus 78). These training samples are then used to determine the threshold of training features using decision tree, as described in Section III.

For moving caption detection, we have to consider the affect from background. We differentiate background and caption by speed and direction of the optical flow. We only calculate optical flow in four directions, namely, left, right, up and down. It is reasonable because caption seldom moves along other directions in a movie or TV program. However, the background may move towards all directions. Even if the directions are the same, the moving speed of caption is fixed but that of background is often random. We can also utilize this phenomenon to differentiate background and caption.

Table IV shows the confusion matrix of the detection result by applying the threshold learned from the training data. The

TABLE IV
CONFUSION MATRIX OF THE MOVING CAPTION DETECTION

	Video without moving caption	Video with moving caption
Video without moving caption	95.3%	4.6%
Video with moving caption	8.7%	91.3%

remaining 797 video shots are used as the test samples. It can be seen that the detection ratio of moving captions is over 90%. But the miss-detection ratio (8.7%) of moving caption shots is much higher than that (4.6%) of the nonmoving caption shots. Most of this situation is caused by the blurred image in which the corner points can not be effectively detected. Another reason is that sometimes the motion direction of the captions is irregular and the moving speed is too fast.

V. CONCLUSION

We have developed an automatic video text and caption detection system. Viewing the corner points as the fundamental feature of character and text in visual media, the system detects video text with high precision and efficiency. We built up several discriminative features for text detection on the base of the corner points. These features can be used flexibly to adapt different applications. We also presented a novel approach to detect moving captions from video shots. Optical flow based motion feature is combined with the text features to detect the moving caption. Over 90% detection ratio is attained. The results are very encouraging. Most of the algorithms presented in this paper are easy to implement and can be straightforwardly applied to caption extraction in video programs with different languages. Our next focus will be on the word segmentation and text recognition based on the results of text detection.

ACKNOWLEDGMENT

The authors would like to thank Singapore A*Star for their generous contribution of the data set.

Authorized licensed use limited to: University of New South Wales. Downloaded on August 17,2020 at 01:49:45 UTC from IEEE Xplore. Restrictions apply.

REFERENCES

[1] Y. Rui, T. S. Huang, and S. F. Chang, "Image retrieval: Current techniques, promising directions, and open issues," *J. Vis. Commun. Image Represent.*, vol. 10, no. 1, pp. 39–62, 1999.

[2] A. W. M. Smeulders, M. Worring, S. Santini, A. Gupta, and R. Jain, "Content-based image retrieval at the end of the early years," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 22, no. 12, pp. 1349–1380, Dec. 2000.

[3] Y. A. Aslandogan and C. T. Yu, "Techniques and systems for image and video retrieval," *IEEE Trans. Knowl. Data Eng.*, vol. 11, no. 1, pp. 56–63, Jan./Feb. 1999.

[4] X. Tang, X. Gao, J. Liu, and H. Zhang, "A spatial-temporal approach for video caption detection and recognition," *IEEE Trans. Neural Netw.*, vol. 13, no. 4, pp. 961–971, Jul. 2002.

[5] H. Li, D. Doermann, and O. Kia, "Automatic text detection and tracking in digital video," *IEEE Trans. Image Proces.*, vol. 9, no. 1, pp. 147–156, Jan. 2000.

[6] K. Kim, K. Jung, and J. Kim, "Texture-based approach for text detection in images using support vector machines and continuously adaptive mean shift algorithm," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 25, no. 12, pp. 1631–1639, Dec. 2003.

[7] Y. Zhong, H. Zhang, and A. K. Jain, "Automatic caption localization in compressed video," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 22, no. 4, pp. 385–392, Apr. 2000.

[8] K. Jung, K. I. Kim, and A. K. Jain, "Text information extraction in images and video: A survey," *Pattern Recognit.*, vol. 37, no. 5, pp. 977–997, 2004.

[9] F. Idris and S. Panchanathan, "Review of image and video indexing techniques," *J. Vis. Commun. Image Represent.*, vol. 8, no. 2, pp. 146–166, 1997.

[10] Y. Zhong, K. Karu, and A. K. Jain, "Locating text in complex color images," *Pattern Recognit.*, vol. 28, no. 10, pp. 1523–1535, 1995.

[11] V. Wu, R. Manmatha, and E. M. Riseman, "Textfinder: An automatic system to detect and recognize text in images," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 21, no. 11, pp. 1224–1229, Nov. 1999.

[12] W. Mao, F. Chung, K. K. M. Lam, and W. Sun, "Hybrid chinese/english text detection in images and video frames," in *Proc. 16th Int. Conf. Pattern Recognit.*, 2002, vol. 3, pp. 1015–1018.

[13] B. K. Sin, S. K. Kim, and B. J. Cho, "Locating characters in scene images using frequency features," in *Proc. 16th Int. Conf. Pattern Recognit.*, 2002, vol. 3, pp. 489–492.

[14] Y. Hao, Z. Yi, H. Zengguang, and T. Min, "Automatic text detection in video frames based on bootstrap artificial neural network and ced," in *Proc. 11th Int. Conf. Central Eur. Comput. Graph., Vis. Comput. Vis.*, 2003, vol. 11, pp. 4246–4251.

[15] A. K. Jain and B. Yu, "Automatic text location in images and video frames," *Pattern Recognit.*, vol. 31, no. 12, pp. 2055–2076, 1998.

[16] J. Ohya, A. Shio, and S. Akamatsu, "Recognizing characters in scene images," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 16, no. 2, pp. 214–220, Feb. 1994.

[17] R. Lienhart and F. Stuber, "Automatic text recognition in digital videos," in *Proc. Image Video Process. IV*, pp. 180–188.

[18] W. Kim and C. Kim, "A new approach for overlay text detection and extraction from complex video scene," *IEEE Trans. Image Process.*, vol. 18, no. 2, pp. 401–411, Feb. 2009.

[19] M. A. Smith and T. Kanade, Video Skimming for Quick Browsing Based on Audio and Image Characterization Carnegie Mellon Univ., School Comput. Sci., Pittsburgh, PA, Tech. Rep. CMU-CS-95-186, 1995.

[20] L. Agnihotri and N. Dimitrova, "Text detection for video analysis," in *Proc. IEEE Workshop Content-Based Access Image Video Libraries*, Washington, DC, 1999, pp. 109–113.

[21] C. Garcia and X. Apostolidis, "Text detection and segmentation in complex color images," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, 2000, vol. 6, pp. 2326–2329.

[22] X. Liu and J. Samarabandu, "Multiscale edge-based text extraction from complex images," in *Proc. Int. Conf. Multimedia Expo.*, 2006, pp. 1721–1724.

[23] M. R. Lyu, J. Song, and M. Cai, "A comprehensive method for multilingual video text detection, localization, and extraction," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 15, no. 2, pp. 243–255, Feb. 2005.

[24] F. Coulmas, *The Blackwell Encyclopedia of Writing Systems*. Oxford, U.K.: Blackwell, 1996.

[25] X. S. Hua, X. R. Chen, L. Wenyin, and H. J. Zhang, "Automatic location of text in video frames," in *Proc. ACM Workshops Multimedia: Multimedia Inf. Retriev.*, 2001, pp. 24–27.

[26] C. Harris and M. Stephens, "A combined corner and edge detector," in *Proc. Alvey Vis. Conf.*, 1988, pp. 147–151.

[27] C. Schmid, R. Mohr, and C. Bauckhage, "Evaluation of interest point detectors," *Int. J. Comput. Vis.*, vol. 37, no. 2, pp. 151–172, 2000.

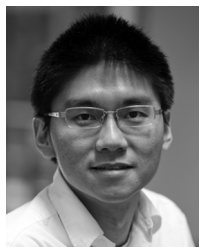
[28] B. D. Lucas and T. Kanade, "An iterative image registration technique with an application to stereo vision," in *Proc. Int. Joint Conf. Artif. Intell.*, 1981, pp. 674–679.

[29] J. R. Bergen, P. Anandan, K. J. Hanna, and R. Hingorani, "Hierarchical model-based motion estimation," in *Proc. Eur. Conf. Comput. Vis.*, 1992, vol. 588, pp. 237–252.

[30] L. Breiman, *Classification and Regression Trees*. London, U.K.: Chapman & Hall, 1998.

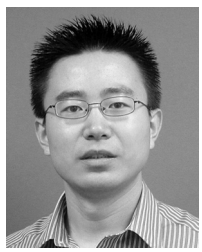
Xu Zhao received the M.S. degree in electrical engineering from China Ship Research and Develop Academy, Beijing, China, in 2004, and is currently pursuing the Ph.D. degree at the Institute of Image Processing and Pattern Recognition, Shanghai Jiao Tong University, Shanghai, China.

He was a visiting student at the Beckman Institute for Advanced Science and Technology, University of Illinois, Urbana-Champaign, from 2007 to 2008. His research interests include visual analysis of human motion, machine learning, and image/video processing.



Kai-Hsiang Lin received the B.S. degree from the Department of Engineering Science and Ocean Engineering, National Taiwan University (NTU), Taipei, Taiwan, in 2002, the M.S. degree from the Institute of Electrical and Control Engineering, National Chiao-Tung University (NCTU), Hsinchu, Taiwan, in 2004, and is currently pursuing the Ph.D. degree at the Department of Electrical and Computer Engineering, University of Illinois, Urbana-Champaign (UIUC), Urbana.

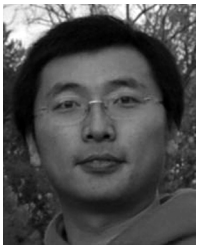
His research interests include image processing, computer vision, and machine learning.



Yun Fu (S'07–M'08) received the B.Eng. degree in information engineering in 2001 and the M.Eng. degree in pattern recognition and intelligence systems in 2004, both from Xi'an Jiaotong University (XJTU), China. He received the M.S. degree in statistics in 2007 and the Ph.D. degree in electrical and computer engineering in 2008, both from the University of Illinois at Urbana-Champaign (UIUC). He was a research intern with Mitsubishi Electric Research Laboratories, Cambridge, MA, in summer 2005, and with Multimedia Research Lab of Motorola Labs, Schaumburg, IL, in summer 2006. He joined BBN Technologies, Cambridge, MA, as a Scientist in 2008. He held a part-time Lecturer position at the Department of Computer Science, Tufts University, Medford, MA, in the spring of 2009. He joined the Department of Computer Science and Engineering, SUNY at Buffalo, as an Assistant Professor in 2010.

Dr. Fu's research interests include Applied Machine Learning, Human-Centered Computing, Pattern Recognition, and Intelligent Vision System. He is the recipient of the 2002 Rockwell Automation Master of Science Award, Edison Cups of the 2002 GE Fund Edison Cup Technology Innovation Competition, the 2003 Hewlett-Packard (HP) Silver Medal and Science Scholarship, the 2007 Chinese Government Award for Outstanding Self-financed Students Abroad, the 2007 DoCoMo USA Labs Innovative Paper Award (IEEE ICIP'07 best paper award), the 2007-2008 Beckman Graduate Fellowship, the 2008 M. E. Van Valkenburg Graduate Research Award, the ITESOFT Best Paper Award of 2010 IAPR International Conferences on the Frontiers of Handwriting Recognition (ICFHR), and the 2010 Google Faculty Research Award. He is a life member of the Institute of Mathematical Statistics (IMS) and Beckman Graduate Fellow.

Authorized licensed use limited to: University of New South Wales. Downloaded on August 17,2020 at 01:49:45 UTC from IEEE Xplore. Restrictions apply.



Yuxiao Hu (M'08) received the Ph.D. degree in electrical and computer Engineer from the University of Illinois, Urbana-Champaign.

He is currently a Research Software Developer Engineer in Microsoft Bing Multimedia Search, Redmond, WA. Before that, He worked as an Assistant Researcher in Microsoft Research Asia during 2001 and 2004. His research interests include multimedia search (image and video analysis), computer vision, and machine learning.

Yuncai Liu (M'94) received the Ph.D. degree from the Department of Electrical and Computer Science Engineering, University of Illinois, Urbana-Champaign, in 1990.

From 1990 to 1991, he was an Associate Researcher with Beckman Institute of Science and Technology, Urbana, IL. Since 1991, he had been a System Consultant and then a Chief Consultant of Research with Sumitomo Electric Industries Ltd., Tokyo, Japan. In 2000, he joined the Institute of Image Processing and Pattern Recognition, Shanghai Jiao Tong University, Shanghai, China, and is currently a Distinguished Professor. His current research interests include image processing and computer vision, especially in motion estimation, feature detection and matching, and image registration. He also made many progresses in the research of intelligent transportation systems.

Thomas S. Huang (S'61–M'63–SM'76–F'79–LF'01) received the B.S. degree in electrical engineering from National Taiwan University, Taipei, Taiwan, R.O.C., and the M.S. and D.Sc. degrees in electrical engineering from the Massachusetts Institute of Technology (MIT), Cambridge.

He was on the Faculty of the Department of Electrical Engineering, MIT, from 1963 to 1973, and on the Faculty of the School of Electrical Engineering and Director of its Laboratory for Information and Signal Processing at Purdue University, West Lafayette, IN, from 1973 to 1980. In 1980, he joined the University of Illinois at Urbana-Champaign, where he is now the William L. Everitt Distinguished Professor of Electrical and Computer Engineering and Research Professor at the Coordinated Science Laboratory and Head of the Image Formation and Processing Group at the Beckman Institute for Advanced Science and Technology and Co-Chair of the Institute's major research theme Human Computer Intelligent Interaction. His professional interests lie in the broad area of information technology, especially the transmission and processing of multidimensional signals. He has published 20 books and over 500 papers in network theory, digital filtering, image processing, and computer vision.

Dr. Huang is a Member of the National Academy of Engineering; a Foreign Member of the Chinese Academies of Engineering and Sciences; and a Fellow of the International Association of Pattern Recognition and the Optical Society of American. He has received a Guggenheim Fellowship, an A. V. Humboldt Foundation Senior U.S. Scientist Award , and a Fellowship from the Japan Association for the Promotion of Science. He received the IEEE Signal Processing Society's Technical Achievement Award in 1987 and the Society Award in 1991. He was awarded the IEEE Third Millennium Medal in 2000. Also in 2000, he received the Honda Lifetime Achievement Award for "contributions to motion analysis". In 2001, he received the IEEE Jack S. Kilby Medal. In 2002, he received the King-Sun Fu Prize, International Association of Pattern Recognition, and the Pan Wen-Yuan Outstanding Research Award. In 2005, he received the Okawa Prize. In 2006, he was named by IS&T and SPIE as the Electronic Imaging Scientist of the year. He is a Founding Editor of the *International Journal Computer Vision, Graphics, and Image Processing* and Editor of the Springer Series in Information Sciences, published by Springer Verlag.

Authorized licensed use limited to: University of New South Wales. Downloaded on August 17,2020 at 01:49:45 UTC from IEEE Xplore. Restrictions apply.