ISSN 1598-0170 (Print)
ISSN 2287-1136 (Online)
http://www.jksii.or.kr

모바일 비디오기기 위에서의 중요한 객체탐색을 위한 문맥인식 특성벡터 선택 모델

Context Aware Feature Selection Model for Salient Feature Detection from Mobile Video Devices

이재호[1] 신현경[2*]
Jaeho Lee Hyunkyung Shin

요 약

모바일 기기를 사용한 실시간 비디오 영상처리분야의 중요 객체탐색 및 추적의 문제에 있어서 난제는 복잡한 배경속에서 전경을 구분해 내는 일이다. 본 논문에서는 기계학습을 위한 특성벡터 선정의 문제를 위한 문맥인식 모델을 제시하여 잡음제거를 위한 기계학습기반의 구분자를 구현하였다. 수학적으로 NP-hard로 알려진 가장 가까운 이웃을 사용한 문맥인식 특성벡터 선정 알고리즘의 구현에 있어서, 본 논문은 연산횟수를 줄인 유사방법론에 대해 자세히 거론하였다. 또한, 문맥인식 성격을 가미한 특성벡터 선정을 통해 얻어진 특성 공간에서의 향상된 분리성에 대해 주성분 분석을 통해 엄밀한 분석결과를 제시하였다. 전반적인 성능 향상의 정도를 계측하기 위해 다양한 기계학습 방법론, 예를 들어, 다층신경망, 지원벡터기계, 나이브베이지안, 회귀분석 등을 사용해 비교결과를 제시하였다. 본 논문에서 제시한 방법론의 성능과 계산상 자원사용에 대한 내용을 결론으로 서술하였다.

☞ 주제어 : 특징벡터선택, 가장가까운근방탐색, 주성분분석, 중요객체탐색, 기계학습.

ABSTRACT

Cluttered background is a major obstacle in developing salient object detection and tracking system for mobile device captured natural scene video frames. In this paper we propose a context aware feature vector selection model to provide an efficient noise filtering by machine learning based classifiers. Since the context awareness for feature selection is achieved by searching nearest neighborhoods, known as NP hard problem, we apply a fast approximation method with complexity analysis in details. Separability enhancement in feature vector space by adding the context aware feature subsets is studied rigorously using principal component analysis (PCA). Overall performance enhancement is quantified by the statistical measures in terms of the various machine learning models including MLP, SVM, Naïve Bayesian, CART. Summary of computational costs and performance enhancement is also presented

☞ keyword : feature vector selection, nearest neighbor search,principal component analysis, salient feature detection, machine learning

1. Introduction

Mobile devicesare undergone a speedy transformation to intelligent stimulus-response (S/R) agent reacting in real time to the sensor captured sonic and the optical information [1]. As for sonic stimulus S/R agent, Siri and Shazam are the tangible examples. As for camera captured optical stimulus S/R agent, problems of face, gesture, and character recognition from natural scene have been the main target framework from various research projects on moving object detection and tracking [2,3] exemplified by product scanning, business card reader, and international travel aid with foreign language translator.

In this paper we focus on the issues of optical S/R agent against the texts embedded in video frames acquired through optical sensor of mobile device. For clarity in writing, we assume an ideal sensor conditions with no noises due to climate, optical sensor surface, motion blur, and digitization error. For text extraction, presence of the cluttered background is the most significant obstacle. Various preprocessing methods

[1] Department of Computer Education, Gyeongin National University of Educatio(jjlee@ginue.ac.kr)
[2] Department of Mathematical Science, Gachon University
[*] Corresponding author (hyunkyung@gachon.ac.kr)

[Received 1December 2014, Reviewed 3December 2014, Accepted 6Decmeber2014]

http://dx.doi.org/10.7472/jksii.2014.15.6.117

are used such as background normalization [4], statistical block region classification [5], and salient color area extraction [6]. With the preprocessed image, gradient based edge retrieval is a basis of illumination invariant object detection where the retrieved edge provides candidate location of the object under consideration. The candidate edge objects are then passed through a filter implemented by problem domain specific pattern classification. The general tools for filtering are the supervised learning based statistical models such as MLP(Multi-layer perceptron),SVM(Support vector machine), Naïve Bayesian, Decision Tree, and Boosting, which apply different training rules but share the common requirement, the input feature vector labeled with ground truth value. At this stage the factual impediment caused by presence of cluttered background emerges. The cluttered background, e.g., tree leaves near traffic signs, generates the surrounding edge structures to obscure decision making by the classifiers. Gao et al. [7] showed a discriminant bottom-up saliency model with center-surround stimulus. In this paper we propose a feature selection model including the neighborhood information to guarantee strong separability of feature vector space between the non-text objects and the character.

The remainder of this paper is organized as follows: in section 2 an overview of the related work is given with respect to the three standard methods of text detection for from document and natural scene. In section 3 the proposedcontext aware feature selection model is explained in details. In section 4 we present the results of experiments and the qualitative assessment on the reference data used for the study of this paper. Section 5 gives a short conclusion.

2. Related Works

Text extraction problem from image or video frame captured from natural scene has been approached by the three types of methods: the region based, the edge based, and the texture based. The region based method is top-down and classifies a block of pixels using various forms of region's energy estimated by analysis of pixel value distribution within the region. Ohya et al. [8] used gray level difference, Shim et al. [9] used homogeneity of intensity, Lienhart et al. [10] used the mean absolute difference, Chun et al. [11] uses FFT(Fast fourier transform), Qian et al. [12] used DCT coefficients, Chen et al. [13] and Lie et al. [11] used Gabor, Mao et al. [14] used wavelet coefficients, Clark et al.[15], Kim et al.[16], and Ekin[15] used spatial variance.

The block by block energy values are used as input features for a classifier to decide candidate text containing region. For the classifiers, a rule based is applied by [18], a heuristics based is applied by [19,34], MLP is applied by [11], SVM is applied by [16], and CRF (conditional random field) is applied by [17]. Pixel block classification methods are fast and stable: fastness comes from the simplicity of pixel value operations deduced by pre-defined mathematical formula, and stability comes from the stability of well-defined statistical measures. But block classification methods are suffered from high type-II error level (false negative). In text detection problem, false positive (incorrectly classifies non-texts to texts) is tolerable but false negative (incorrectly classifies texts to non-texts) is not.

The edge based method is bottom-up approach, which retrieves edge structures from input image and rules out non-text candidates. For the filtering criterion, geometrical and statistical characteristics of the connected components of edge (contour) are used: Smith et al. [19] uses aspect ratio, fill factor, and size of bounding box; Yassin et al. [20] used gray level homogeneity in addition to the three features; Chen et al. [21] used Gabor filter for feature selection and used aspect ratio for filtering; Wang et al. [22] used color segmentation to locate text contours; Zhang et al.[23] introduced MRF(Markov random field) for scene text learning system; Kim et al. [24] used color continuity, gray level variation and color variation as the features to build text strokes; Liu et al. [25] used wavelet coefficients; Lyu et al. [26] used aspect ratio; Takahashi et al. [27] used representative color, positions, area, and aspect ratio; Epshtein et al. [3] was more focused attention on retrieval of edge structure using MSER (maximally stable extremal regions) and used uniformity of stroke width as its feature.

The features associated with individual contours are taken as input of classifier to perform filtering: [19, 20] used rule-based predicators, [21] used MLP, [24] used SVM for the filtering.

Finally, the texture based methods are characterized by pixel based (not region based) and neighborhood region's energy evaluation (not edge based). Zhong et al. [29] used 1x21 horizontal spatial variance; Wu et al. [30] used nine second-order gaussian derivatives; Sin et al. [31] and Mao et al. [32] used frequency domain analysis of FFT and wavelet, respectively. Frequency domain transformations are computationally expensive, applying the transformations at each pixel is not practically applicable, which is the reason that researchers choose coarse-to-fine approach like region based as described above.

3. Context Aware Feature Selection Model

Suppose the contours of candidate text objects are retrieved from natural scene image, classification on the contours is best attained by machine learning method. For the purpose of construction of training data feature vector selection is achieved by estimating two dimensional descriptors of the contours. For instances, perimeter length ratio to area (Arc/Area = P^2/A) for circularity, occupancy ratio (Occupancy = mM/A) for density, ratio of minimum and maximum axis length of bounding rectangle (Aspect Ratio = m/M) for aspect ratio, and seven count of central momentum (μ_{02}, μ_{03}, μ_{11}, μ_{12}, μ_{20}, μ_{21}, μ_{30}) of a given contour are the standard measurements. Refer to the Figure 1for the notations.

All the 10 features described above have the commonality that they are the descriptors of single contour without any observation on their neighborhood contours. As can be seen at Figure 3 below, the graphs of interquartile range show that these 10 features do not assure obvious separability between text and noise. This behavior is as expected since a contour retrieved from tree leaf or brick of building does not show much difference in the shape with a contour from text.

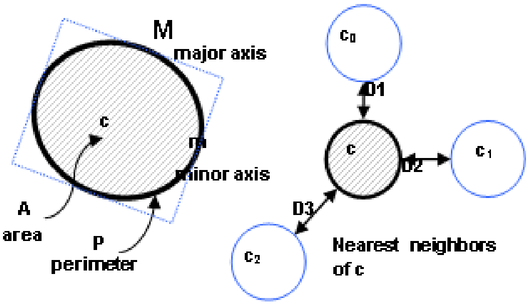


Figure 1. Notations used in feature selection

The discrepancy between texts and noises lies in the spatial distribution which cannot be assessed by a single contour. Qualitative examination shows that, as seen in Fig. 2, tree leaves' locations are distributed irregularly and bricks of building have repetition of exact same pattern. Based on this observation, we propose an enhanced feature selection model which exercises configurations of neighborhood contours. As seen in the Fig. 1, for a given contour c the three of nearest neighbor contours (c_0, c_1, c_2) are considered.

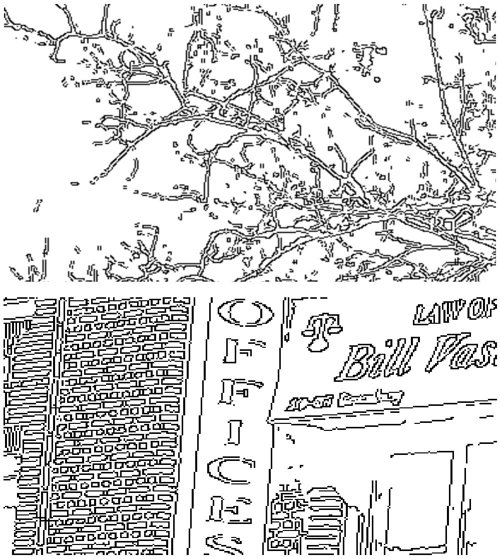


Figure 2. Shape of contours retrieved from tree leaves (top) and bricks of building (bottom)

Searching for the nearest neighbors is a NP hard problem [28], which can cause the proposed model unnecessarily complicated bottleneck and unusable for fast real time processing environment. We adopt the methods from FLANN[33]. For the construction of the index class, L2 is used for the distance measure between features and kd tree structure is used for indexing. The nearest neighborhood searching performance analysis shows that *112.46* average tick counts with variance of *32.68* in CPU with *3.10*GHz, which amounts to average *0.037* milliseconds and variance *0.011* milliseconds. The contour contour distances **D1**, **D2**, and **D3** are newly added to the existing feature selection, which grants context aware property for feature selection. We should mention that contour contour distance is scale dependent. For the purpose of maintaining scale invariant property, we normalize **D1**, **D2**, and **D3** by the height of the contour **c**, i.e.

D1 = |**c** – **c**$_0$| / height of **c**
D2 = |**c** – **c**$_1$| / height of **c**
D3 = |**c** – **c**$_2$| / height of **c**

As a result, feature vector selection issummarized at the Table1.

Table 1. Formation of feature vector selection.F(10) F(12) are context aware feature components

0	1	2	3	4	5	6	7	8	9	10	11	12
μ_{02}	μ_{03}	μ_{11}	μ_{12}	μ_{20}	μ_{21}	μ_{30}	P^2/A	Mm/A	m/M	D1	D2	D3

Interquartile range analysis and principal component analysis on the selected features acquired from certain data set are performed at the following section.

4. Experimental Results

For the validation of the proposed feature vector selection, we sampled the two groups of image data as the following way. For the one group, we collected *32* public images from the English language training data used by 'tesseract OCR'. For the other group, we took *32* pictures at the street of New York City using a cell phone. *16* of *32* pictures contain the texts and the rest do not contain the texts. Refer to the Table 2 for summary.

Table 2.Organization of ground truth data set

Set 1	*32* document images containing texts only, which are the English training data for googletesseract	
Set 2	A	*16* natural scene images containing no texts
	B	*16* street images containing texts

From the Set 1, about *600,000* contours were obtained and most of the contours of which sizes are greater than the threshold of *32* are considered as text representing contours, on the other hand, from the Set 2 A, about *10,000* contours are picked as the noise representing contours.

Each of *13* components of the proposed feature vector was measured by using randomly sampled contours from the ground data set in Table 2, approximately *15%*. The graphs in Fig. 3 below present the interquartile range of the values. For example, the top left graph can be read as follows: the interquartile range of μ_{02} of 'text representing contours' is [*0.09, 0.19*] and that of 'noise representing contours' is [*0.06, 0.16*].

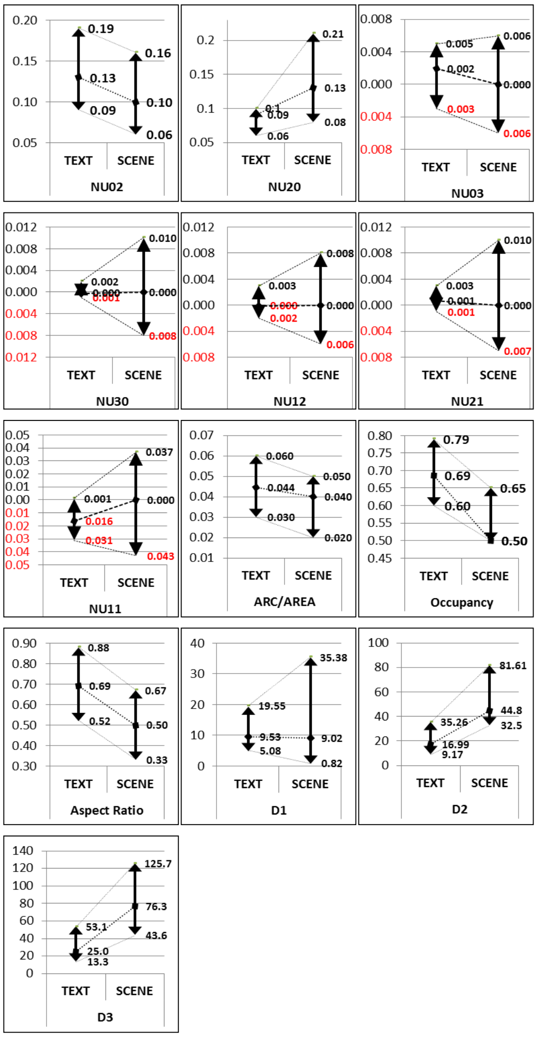
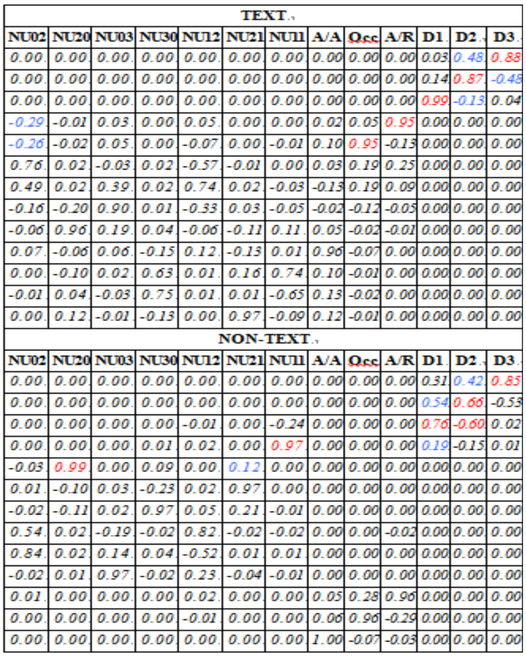


Figure 3. Interquartile range of the features for text and non text. Red colored letter indicates negative number

The interquartile ranges of central momentum show that the medians of text feature fall within the interquartile range of the noises, which indicates no significant difference between them. For the case of *2* dimensional spatial edge boundary descriptors (**ARC/AREA**, **Occupancy**, and **Aspect Ratio**), except **ARC/AREA**, there are stronger difference. Among the newly added features, **D2** and **D3** have extremely strong significant difference between the texts and the noises.

Principal component analysis (PCA) is applied to the covariance matrix of realizations of normalized feature vector space and is summarized in Table 3. The table demonstrates two sets of eigenvectors, the one from text feature space (titled as TEXT) and the other from non text feature space (titled NON TEXT). The first and second maximal magnitudes of the five maximum eigenvector are colored red and blue, respectively. PCA result shows that the maximal eigenvectors are similar, the second and the third eigenvectors are slightly different, and the fourth eigenvectors are orthogonally different. This affords linear separability of the feature vector space between the text and the non text.

Table 3. List of eigenvectors from PCA on covariance matrix of feature vectors for the text and the noise

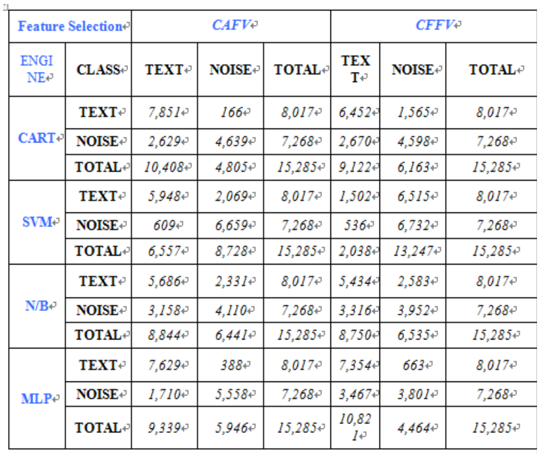


TEXT												
NU02	NU20	NU03	NU30	NU12	NU21	NU11	A/A	Occ	A/R	D1	D2	D3
0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.03	0.48	0.88
0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.14	0.87	-0.48
0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.99	-0.13	0.04
-0.29	-0.01	0.03	0.00	0.05	0.00	0.00	0.02	0.05	0.95	0.00	0.00	0.00
-0.26	-0.02	0.05	0.00	-0.07	0.00	-0.01	0.10	0.95	-0.13	0.00	0.00	0.00
0.76	0.02	-0.03	0.02	-0.57	-0.01	0.00	0.03	0.19	0.25	0.00	0.00	0.00
0.49	0.02	0.39	0.02	0.74	0.02	-0.03	-0.13	0.19	0.09	0.00	0.00	0.00
-0.16	-0.20	0.90	0.01	-0.33	0.03	-0.05	-0.02	-0.12	-0.05	0.00	0.00	0.00
-0.06	0.96	0.19	0.04	-0.06	-0.11	0.11	0.05	-0.02	-0.01	0.00	0.00	0.00
0.07	-0.06	0.06	-0.15	0.12	-0.13	0.01	0.96	-0.07	0.00	0.00	0.00	0.00
0.00	-0.10	0.02	0.63	0.01	0.16	0.74	0.10	-0.01	0.00	0.00	0.00	0.00
-0.01	0.04	-0.03	0.75	0.01	0.01	-0.65	0.13	-0.02	0.00	0.00	0.00	0.00
0.00	0.12	-0.01	-0.13	0.00	0.97	-0.09	0.12	-0.01	0.00	0.00	0.00	0.00
NON-TEXT												
NU02	NU20	NU03	NU30	NU12	NU21	NU11	A/A	Occ	A/R	D1	D2	D3
0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.31	0.42	0.85
0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.54	0.66	-0.53
0.00	0.00	0.00	0.00	-0.01	0.00	-0.24	0.00	0.00	0.00	0.76	-0.60	0.02
0.00	0.00	0.00	0.01	0.02	0.00	0.97	0.00	0.00	0.00	0.19	-0.15	0.01
-0.03	0.99	0.00	0.09	0.00	0.12	0.00	0.00	0.00	0.00	0.00	0.00	0.00
0.01	-0.10	0.03	-0.23	0.02	0.97	0.00	0.00	0.00	0.00	0.00	0.00	0.00
-0.02	-0.11	0.02	0.97	0.05	0.21	-0.01	0.00	0.00	0.00	0.00	0.00	0.00
0.54	0.02	-0.19	-0.02	0.82	-0.02	-0.02	0.00	0.00	-0.02	0.00	0.00	0.00
0.84	0.02	0.14	0.04	-0.52	0.01	0.01	0.00	0.00	0.00	0.00	0.00	0.00
-0.02	0.01	0.97	-0.02	0.23	-0.04	-0.01	0.00	0.00	0.00	0.00	0.00	0.00
0.01	0.00	0.00	0.00	0.02	0.00	0.00	0.05	0.28	0.96	0.00	0.00	0.00
0.00	0.00	0.00	0.00	-0.01	0.00	0.00	0.06	0.96	-0.29	0.00	0.00	0.00
0.00	0.00	0.00	0.00	0.00	0.00	0.00	1.00	-0.07	-0.03	0.00	0.00	0.00

With the feature vector selection model as proposed in this paper, we perform classification using CART, SVM, Naïve Bayesian (N/B), and MLP. For the purpose of demonstration on enhancement by new feature selection, we also perform classification with the feature vector model without **D1**, **D2**, and **D3**. For clarity of writing, we denote ***CAFV*** (context aware feature vector) and ***CFFV***(context free feature vector) as follows: ***CAFV*** = {(***F***[*0*], ⋯, ***F***[*12*]) | ***F*** is defined in Table 1}and***CFFV*** = {(***F***[*0*], ⋯, ***F***[*9*]) | ***F*** is defined in Table 1}

From the ground truth data, two third of them are used for training and one third of them are used for validation. The number of data is so large that cross validation is not necessary. The results of classification are summarized in Table 4. The number of text representing contours is *8,017* and the number of noise (non text) contours is *7,268*. As an example, CART classifier with ***CAFV*** determined *7,851* number of TEXT correctly out of *8,017* but a number *166* of texts were classified as NOISE. It determined *2,629* counts of NOISE out of *7,268* as TEXT. We use a standard parameter for CART as follows: maximum depth is *80*; minimum sample count for a node *3*; regression accuracy is *0.002*; surrogation of node, 1se rule, and tree pruning are allowed, and priors are set by uniform vector. For SVM, we use RBF(radial basis function) kernel. For MLP, we use back propagation rule for error correction learning and *2* hidden layers with *13* and *5* nodes, respectively, for multi layered structure.

Table 4. Classification results from the four different classifiers and the two types of feature vectors



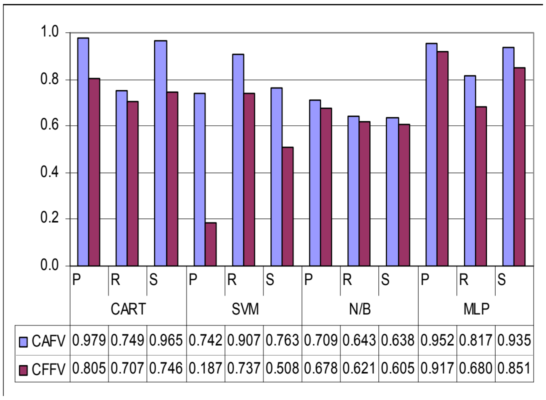
Feature Selection		*CAFV*			*CFFV*		
ENGINE	CLASS	TEXT	NOISE	TOTAL	TEXT	NOISE	TOTAL
CART	TEXT	*7,851*	*166*	*8,017*	*6,452*	*1,565*	*8,017*
	NOISE	*2,629*	*4,639*	*7,268*	*2,670*	*4,598*	*7,268*
	TOTAL	*10,408*	*4,805*	*15,285*	*9,122*	*6,163*	*15,285*
SVM	TEXT	*5,948*	*2,069*	*8,017*	*1,502*	*6,515*	*8,017*
	NOISE	*609*	*6,659*	*7,268*	*536*	*6,732*	*7,268*
	TOTAL	*6,557*	*8,728*	*15,285*	*2,038*	*13,247*	*15,285*
N/B	TEXT	*5,686*	*2,331*	*8,017*	*5,434*	*2,583*	*8,017*
	NOISE	*3,158*	*4,110*	*7,268*	*3,316*	*3,952*	*7,268*
	TOTAL	*8,844*	*6,441*	*15,285*	*8,750*	*6,535*	*15,285*
MLP	TEXT	*7,629*	*388*	*8,017*	*7,354*	*663*	*8,017*
	NOISE	*1,710*	*5,558*	*7,268*	*3,467*	*3,801*	*7,268*
	TOTAL	*9,339*	*5,946*	*15,285*	*10,821*	*4,464*	*15,285*

To analyze the results, we use the three standard statistical measurements: precision, recall, and specificity defined as follows:

- Precision (***P***) = ***prob***{Classified as TEXT | All TEXT}.
- Recall (***R***) = ***prob***{Classified as TEXT & ALL TEXT | Classified as TEXT }.
- Specificity (***S***) = ***prob***{Classified as NOISE | ALL NOISE} .

The three measurements are presented in Fig 4 by using the data in Table 4. As seen in the table, the precisions are significantly improved by using CAFV. SVM shows failure in classification with CFFV but becomes successful with CAFV.

The recall rates are all good and more stabilized with CAFV. Specificity is a test of negativity, which is not so important as precision and recall since losing text to noise is serious but losing noise to text is not serious defect.

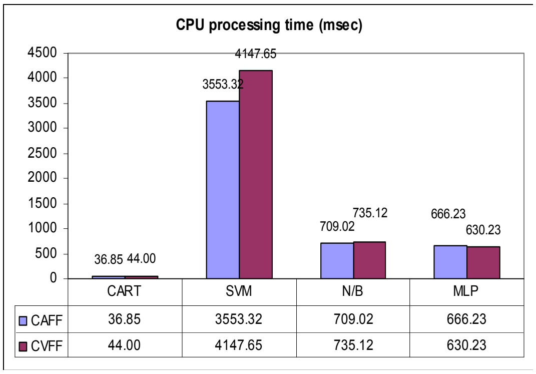


	CART			SVM			N/B			MLP		
	P	R	S	P	R	S	P	R	S	P	R	S
▢ CAFV	0.979	0.749	0.965	0.742	0.907	0.763	0.709	0.643	0.638	0.952	0.817	0.935
■ CFFV	0.805	0.707	0.746	0.187	0.737	0.508	0.678	0.621	0.605	0.917	0.680	0.851

Figure 4. Precision, recall, and specificity of classification results based on CAFV and CFFV

For the conciseness of paper, we drop the data on the training error.

Finally, Fig. 5 illustrates the processing time in milliseconds via CPU tick counts for each classifier. We evaluated the tick counts during the tests to create the data presented in Table 4. The result indicates that CART is the fastest classifier and 18 times faster than MLP.



	CART	SVM	N/B	MLP
▢ CAFF	36.85	3553.32	709.02	666.23
■ CVFF	44.00	4147.65	735.12	630.23

Figure 5. Time complexity of classifiers for classification

Interestingly, classification using ***CAFV*** (with 13 dimensions) is faster than using ***CFFV*** (with 10 dimensions) except MLP. As seen in PCA analysis, newly added ***D2*** and ***D3*** are the most significant factors for formation of the feature vector space of $\boldsymbol{R}^{13}$. Presence of such components may shorten the depth of the node tree (CART), the process of support vector finding (SVM). On the other hand, for the case of MLP, *23%* increase of feature vector dimension (from *10* to *13*) results in *5.4%* more process time (from *630.23* to *666.23* milliseconds) which may be caused by increase of input nodes.

5. Conclusions

We proposed a new feature vector selection model for supervised learning based contour classification problem, a main preprocessing stage of video OCR on natural scene since rectification projection and OCR is subject to be applied only on the contours categorized as text. A standard method of contour classification creates the feature vector using two dimensional geometrical and statistical descriptors of a single contour given individually. We addressed a question that additional contextual information on the nearest neighborhood contours, for example normalized distances between contours, would result in performance enhancement of classification process.

In section 4, we demonstrated huge improvement on the classifiers: *17.45%*, *55.45%*, *3.14%*, and *3.43%* for CART, SVM, N/B, and MLP, respectively. The new feature selection also improved the performance in recall, and specificity. The statistical measure 'type II error', can be regarded as 1

’precision’, is extremely important in this type of problem: we do not want to lose ’text’ by mis classification to ’noise’, while the opposite (losing ’noise’ by mis classification to ’text’) would be acceptable, which implies the precision and the recall are relatively more important than the specificity.

The results show that the non linear discriminant methods (CART and MLP) have superior performance to the linear discriminant methods (SVM and N/B). CART and MLP show extremely well behavior on the precision and the recall rates. The result implies that MLP performs very well without additional context aware features. We would mention that MLP is 18 times slower than CART. For the video OCR requiring real time process speed and for the simplicity of its structure, CART would be the choice.

The objective of this paper is not to find the best feature vector selection model per se for a specific problem but to prove significant improvement as a result of context awareness.

Reference

[1] B. Chen, H.H. Cheng, and J. Palen, “Integrating Mobile Agent Technology with Multi-Agent Systems for Distributed Traffic Detection and Management System,” *Transportation Research Part C*, 17, 2009, pp. 1-10, 200

[2] D.M. Gavrila, “The Visual Analysis of Human Movement: A Survey,” *Computer Vision and Image Understanding*, Vol. 73, No. 1, pp. 82-98, 1999.

[3] B. Epshtein, E. Ofek, and Y. Wexler, “Detecting text in natural scenes with stroke width transform,” in *CVPR*, pp. 2963 – 2970, 2010.

[4] W. Boussellaa, A. Zahour, A. Alimi, "A methodology for the separation of foreground / background in Arabic historical manuscripts using hybrid methods," *Proceeding SAC* ’07, pp. 605 – 609, 2007.

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

[6] L. Zhang, T.K. Marks, M.H. Tong, H. Shan, and G.W. Cottrell, “SUN: a Bayesian framework for saliency using natural statistics”, *Journal of Vision*, 8(7):32, 1-20, 2008.

[7] D. Gao, V. Mahadevan, and N. Vasconcelos. “The discriminant center-surround hypothesis for bottom-up saliency,” in *Proc. of NIPS*, pp. 497 – 504, 2007.

[8] J. Ohya, A. Shio and S. Aksmatsu, ”Recognizing characters in scene images,” *IEEE. Trans. PAMI*, vol. 16, pp. 214-224, 1994.

[9] J.C. Shim, C. Dorai, and R., Bolle, "Automatic text extraction from video for content-based annotationand retrieval," *Proc. 14th Int. Conf. on PR*, vol. 1:16-20 pp. 618 – 620, 1998.

[10] R. Lienhart and W. Effelsberg, “Automatic Text Segmentation and Text Recognition for Video Indexing,” *TR-98-009, PraktischeInformatik IV*, University of Mannheim, 1998.

[11] B.T. Chun, Y. Bae, and T.Y. Kim, “Automatic Text Extraction in Digital Videos Using FFT and Neural Network,” *Proc. IEEE Int. Fuzzy Sys. Conf.* Seoul, Korea, 2:1112-1115, 1999.

[12] Qian, X., Liu, G., “Text Detection, Localization and Segmentation in Compressed Videos,” *Proc. Int. Conf. on Acoustics, Speech and Signal Processing*, 2:385-388, 2006.

[13] D. Chen, K. Shearer, and H. Bourlard, “Text Enhancement with Symmetric Alter for Video OCR,” *Proc. Int. Conf. on Image Analysis and Recognition*, pp.192- 197, 2001.

[14] W. Mao, F. Chung, K. Lanm, and W. Siu, “Hybrid Chinese/ English Text Detection in Images and Video Frames,” *Proc. Int. Conf. on CVPR*, 3:1015-1018, 2002.

[15] P. Clark and M. Mirmehdi, “Finding Text Regions Using Localized Measures,” *Proc. 11th British Machine Vision Conference*, pp.675-684, 2000.

[16] K.I. Kim, K. Jung, and J.H. Kim, “Texture-based approach for text detection in images using support vector machine sand continuously adaptive mean shift algorithm,” *IEEE Trans. PAMI*, 25(12):1631-1639, 2003.

[17] J. Weinman, A. Hanson, and A. McCallum, "Sign Detection In Natural Images With Conditional Random Fields," *IEEE Int. Work. on Machine Learning for Signal Processing*, Brazil, Sep. 2004.

[18] S. Messelodi and C.M. Modena, “Automatic Identification and Skew Estimation of Text Lines in Real Scene Images,” *Pattern Recognition*, 32 (1992) 791-810.

[19] M.A. Smith and T. Kanade, “Video Skimming for Quick Browsing Based on Audio and Image Characterization”, Carnegie Mellon University, *Technical Report* CMU-CS-95-186, July 1995.

[20] Y.M.Y. Hasan and L.J. Karam, “Morphological Text Extraction from Images,” *IEEE Transactions on Image Processing*, 9 (11) pp. 1978-1983, 2000.

[21] D. Chen, K. Shearer, and H. Bourlard, “Text Enhancement with Asymmetric Filter for Video OCR,” *Proc. of Int. Conf. on Image Analysis and Processing*, pp. 192-197, 2001.

[22] Wang, K.Q., Kangas, J.A., “Character location in scene images from digital camera,” *Pattern Recognition*, 36(10): pp. 2287-2299, 2003.

[23] D.Q. Zhang and F.H. Chang, “Learning to Detect Scene Text Using a Higher-Order MRF with Belief Propagation,” *Proc. Int. Conf. on CVPR*, p.101-107, 2004.

[24] K.K. Kim and Y.K. Chung, “Scene Text Extraction in Natural Scene Images Using Hierarchical Feature Combining and Verification,” *Proc. Int. Conf. on CVPR*, 2: pp. 679-682, 2004.

[25] C. Liu, C. Wang, and R. Dai, “Text Detection in Images Based on Unsupervised Classification of Edge-based Features,” *ICDAR*, 2005.

[26] 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.*,15(2):243-255, 2005.

[27] H. Takahashi and M. Nakajima, “Region Graph Based Text Extraction from Outdoor Images,” *Proc. 3rd Int. Conf. on Info. Tech. and App.*, 1:680-685, 2005.

[28] Zhang, Statman, and Shasha, “On the editing distance between unordered labeled trees,” *Information Processing Letters*, 42:133-139, 1992.

[29] Y. Zhong, K. Karu, and A.K. Jain, "Locating text in complex images," *Pattern Recognition*, 28(10), 1523-1535, 1995.

[30] V. Wu, R. Manmatha, and E.M. Riseman, "TextFinder: An Automatic System to Detect and Recognize Text in Images," *IEEE Trans. on PAMI*, 21(11), pp. 1224-1229, 1999.

[31] B. Sin, S. Kim, and B. Cho, "Locating Characters in Scene Images using Frequency Features,” *Proc. of Int. Conf. on Pattern Recognition*, 2002.

[32] W. Mao, F. Chung, K. Lanm, and W. Siu, "Hybrid Chinese/English Text Detection Technology on Image Sequence", *Proc. of Int. Conf. on Pattern Recognition*, vol. 3, pp. 1015-1018, 2002.

[33] D. Arthur and S. Vassilvitskii. “k-means++: the advantages of careful seeding,” *In Proc. of the eighteenth annual ACM-SIAM symposium on Discrete algorithms*, pp. 1027, 2007.

[34] T Plötz , NY. Hammerla and POlivier, “Feature learning for activity recognition in ubiquitous computing", *Pro of the Twenty-Second international joint Conf. on Artificial Intelligence*, pp. 1729-1734, 2011

◐ 저자소개 ◑

이재호 (Jeeho Lee)

1989~1996 Senior research staff at the Electronics and Telecommunications Research Institute
1996~Professor of the Department of Computer Education at Gyeongin National University of Education in
2014 President of the Korean Society for Creative Information Culture and Chief Director
Area of research interests: Education on computer science, invention, and convergence for the gifted.
Email: jhlee@ginue.ac.kr

신현경 (Hyunkyung Shin)

2012 Ph.D. State University of New York at Stony Brook. Department of Applied Mathematics and Statistics
2007~ Associate Professor Gachon University, Korea.
Area of research interests: Image Processing, Computer vision. Machine Learning based on Neural Networks Knowledge Representation and Processing
Email: hyunkyung@gachon.ac.kr