Detecting Informative Frames from Wireless Capsule Endoscopic Video Using Color and Texture Features

M.K. Bashar^{1,3}, K. Mori^{2,3}, Y. Suenaga^{2,3}, T. Kitasaka^{2,3}, and Y. Mekada^{3,4}

¹ Graduate School of Engineering, Nagoya University, Japan
kbashar@suenaga.m.is.nagoya-u.ac.jp

² Graduate School of Information Science, Nagoya University, Japan

³ MEXT Innovative Research Center for Preventive Medical Engineering, Nagoya University, Japan

⁴ School of Life System Science and Technology, Chukyo University, Toyota, Japan

Abstract. Despite emerging technology, wireless capsule endoscopy needs high amount of diagnosis-time due to the presence of many use-less frames, created by turbid fluids, foods, and faecal materials. These materials and fluids present a wide range of colors and/or bubble-like texture patterns. We, therefore, propose a cascade method for informative frame detection, which uses local color histogram to isolate highly contaminated non-bubbled (HCN) frames, and Gauss Laguerre Transform (GLT) based multiresolution norm-1 energy feature to isolate significantly bubbled (SB) frames. Supervised support vector machine is used to classify HCN frames (Stage-1), while automatic bubble segmentation followed by threshold operation(Stage-2) is adopted to detect informative frames by isolating SB frames. An experiment with 20,558 frames from the three videos shows 97.48 % average detection accuracy by the proposed method, when compared with methods adopting Gabor based-(75.52%) and discrete wavelet based features (63.15%) with the same color feature.

1 Introduction

Wireless Capsule Endoscopy (WCE), which is aimed for investigating the gastrointestinal tract, consists of a micro-camera located in the capsule. The capsule contains a whole illumination set, a battery, and a radio frequency emitter which sends a video signal of its trip throughout the gut to an external device in a typical rate of two frames per second. Once the clinical analysis is completed, the recorded video is downloaded into a work station for its visualization by the specialists. The journey time is about eight hours and the stored data consists of approximately 50,000 frames. One of the most salient features of this technique relates to the lack of need of hospitalization or specialized stuff, overcoming most of the main drawbacks related to classical endoscopy [1]. On the other hand, the major weakness of the WCE is the requirement of significant diagnosing time

Fig. 1. (a) Non-informative frames in WCE video. Frames with (i)(ii) bubbles and turbid fluids (a little), and (iii) bubbles only, (iv) faecal, bubbles and turbid fluids, (v) food and turbid fluids, (vi)(vii) faecal material only, (viii) turbid fluid only, (ix) turbid fluid with intestinal lumen, and (x) the processed food. (b) Block diagram of the cascade method.

(approximately 2 hours) and close concentration of an expert clinician, driving this clinical routine not feasible for certain clinical scenarios [2].

The term “non-informative frames” may be defined as having invisible (to normal vision) tissues, folds, and/or lumen in WCE frames. Presence of many such frames is one reason for the large visualization time. In WCE, the good visibility of the internal tissue, folds or organ lumen of the GI tract is usually obstructed by the intestinal juices, which can be visualized as a semi-opaque turbid liquid accompanied by bubbles and other artifacts related to the flux of different fluids into the gut. Residual or unabsorbed foods, and faecal materials mix up with these secretions and complicate the visibility issues. As a result, the correct visualization of the gut is hindered. These useless frames, which have wide range of colors from brown to yellow, can be categorized into two classes: highly contaminated non-bubbled (HCN) frames and significantly bubbled (SB) frames. HCN class consists of color dominated non-bubbled frames, highly contaminated with residual foods and/or faecal materials, while SB class consists of frames associated with bubble patterns of different shapes and sizes. Figure 1(a) shows various non-informative frames as found from three WCE videos of healthy persons.

Recently, Vilarino et al.[1] applied Gabor based texture feature to segment intestinal juices from WCE frames. However, the effects of residual foods, and faecal materials remain unexplored except bubble-patterns in the intestinal juices. We, therefore, adopt texture as well as color features for detecting informative frames from WCE videos. The rest of the paper is organized as follows. Section 2 describes the proposed cascade scheme and features for informative frame detection. Section 3 presents experimental results and the performance comparison including discussion. Finally, the research is concluded with some future remarks through section 4.

2 Cascade Detection Method

2.1 Overview of the System

We propose a two-stage cascade approach for automatic detection of informative frames. Figure 1(b) shows a simple block diagram. In Stage-1, a support vector machine (SVM) classifier is trained with the color feature vectors, computed from the training set of HCN and Non-HCN classes from all videos. Testing set from each video is applied to Stage-1, where HCN frames are separated to output informative and bubbled frames. These output frames, which preserve almost all informative frames, are input to Stage-2, which uses multiresolution texture feature for automatic bubble segmentation and SB frames isolation to produce final informative frames.

2.2 Color and Texture Features

Non-informative frames have very complex patterns of color and textures. A statistical color and multiresolution texture descriptors will be useful to isolate them.

Color Feature: Color features can be extracted directly in the RGB domain, which is not perceptually uniform. In our study, we chose HSV [3] color space for its approximate perceptual uniformity. The main attraction of this space is, however, the separation of chromaticity (hue and saturation) from the luminance (intensity or value). While there are a number of color descriptors (e.g., color moments, color layout, color coherence vector and color correlogram), color histogram [4] is simple but efficient in representing color distribution of images. However, global color histogram lacks of spatial information, which is important for similarity comparison. We, therefore, divide each HSV component (each of size 288×288) into nine non-overlapping blocks (each 32×32) and compute local color histogram from each. H, S, and V components are independently quantized into 12, 5, and 8 uniform bins, respectively. Therefore, 25 bins per spatial-block comprise a total of 225 bins histogram. If N_b is the total number of pixels in a block, then the color histogram can be given by

$$H_{hsv}^b(i) = \frac{n_i^b}{N_b}, \quad (1)$$

where n_i^b is the frequency of i -th bin in the b -th block and $i = 1, 2, \dots, 25$; $b = 1, 2, \dots, 9$. The color features described above will be used with SVM classifier to isolate non-informative frames in Stage-1 of the proposed method.

Texture Feature: Texture feature is used for bubble-patterns segmentation. Laguerre Gauss circular harmonic function (LG-CHFs) in the GLT domain are chosen for their similarity and high correlation with bubble patterns. Moreover,

they have self-steering property to produce rotation-invariant features [5,6]. In GLT, image decomposition is usually performed by the convolution operator:

$$I_k^n(x, y, t) = I(x, y) * \mathcal{L}_k^n(x, y, t), \quad (2)$$

where LG-CHFs, $\mathcal{L}_k^n(x, y, t)$, may be given by using polar notations as:

$$\begin{aligned} \mathcal{L}_k^n(x, y, t) &= \mathcal{L}_k^n\left(\frac{\rho}{\sigma}, \theta\right) = \mathcal{L}_k^n(t, \theta) \text{ and} \\ \mathcal{L}_k^n(t, \theta) &= (-1)^k 2^{\frac{(|n|+1)}{2}} \pi^{\frac{|n|}{2}} \left[\frac{k!}{(|n|+k)!} \right]^{1/2} t^{|n|} L_k^n(2\pi t^2) e^{-\pi t^2} e^{jn\theta}, \end{aligned} \quad (3)$$

where n is the order, k is the degree, and $L_k^n(t)$ is the generalized Laguerre polynomial, defined by the Rodrigues formula:

$$L_k^n(t) = \sum_{h=0}^k (-1)^h \binom{n+k}{k-h} \frac{t^h}{h!}, \text{ with } t = \frac{\rho}{\sigma}. \quad (4)$$

LG-CHFs (Eq. 3) is a complex filter but we used real part to maintain uniformity with other similar methods. The local absolute energy at each pixel of the response images is computed as texture features as given in Eq. 5.

$$I_f^p(x, y, t) = \sum_{(x_0, y_0) \in W} |I_k^n(x + x_0, y + y_0, t)|. \quad (5)$$

Here $|\cdot|$ indicates the norm-1 energy within the local neighborhood W of each response image and $p = 1, 2, \dots, N_f$, where $N_f = (k \times n \times N_\sigma)$. N_σ is experimentally fixed to 2, which corresponds to two discrete values of σ , i.e., $\sigma = \{3.0, 5.0\}$. The same set of filters is repetitively applied to a reduced (by a factor of two) lowpass version of the image at each level. A separable cubic B-spline filter [7] is chosen as lowpass filter and is used before downsampling operation to avoid aliasing.

2.3 Classification by Support Vector Machine (SVM)

SVM classifier [8] is used in Stage-1 of the proposed system to isolate HCN frames from the input images. Because of the wide color variations between HCN and Non-HCN classes, a non-linear classification is preferred. The associated radial basis kernel function (RBF) is given by:

$$K_{rbf}(x, x_i) = \exp \frac{-|x - x_i|^2}{2\sigma^2}, \gamma = \frac{1}{2\sigma^2}, \quad (6)$$

where γ is the kernel parameter. Kernel based optimization also includes a regularization parameter C . It is unknown beforehand which C and γ are the best for one problem. We, therefore, adopted a "grid-search" strategy using 10 fold cross-validations on the training data. Several training sets were tried with SVM to ensure a high cross-validation accuracy, which is needed to avoid overfitting problem. Finally, testing sets were tested with the trained classifier.

2.4 Bubble Structures Segmentation

The algorithm for **bubble patterns segmentation** is as follows:

- **Input:** A wireless capsule endoscopic image.
 - **Output:** Corresponding segmented image.
1. Generate a set of 2D LG-CHF kernels using multiple values of n , k , and σ . Choose also a 1D cubic B spline function as a lowpass filter. The sizes of LG-CHF and cubic B spline filters are chosen to 11×11 and 5×1 , respectively.
 2. Execute following steps using lowpass image (original image at level-1) at each level of decomposition:
 - (a) Convolve the lowpass image with the filter banks and downsample the responses. This will create a set of responses, $I_k^n(x, y)$, including a lowpass one. This lowpass response will be used for the next level of decomposition.
 - (b) Compute initial features, $I_f^p(x, y)$, from all but the lowpass response by Eq.5, where $p = 1, 2, \dots, N_f$, and N_f is the number of initial features.
 - (c) Select effective number of features $N_{ef} < N_f$ by selecting the lower, Th_1 , and upper bound, Th_2 , in the sorted energy-contrast map by analyzing its consistency.
 - (d) Expand features to the original image dimension by linear interpolation [9].
 3. Combine all expanded features by pixel-wise averaging. This will create a combined feature image, $I_{cf}(x, y)$.
 4. Apply Otsu's method [10] on the feature image for automatic threshold detection.
 5. Segment the combined feature image using detected threshold. This will create final segmented image, $I_{seg}(x, y)$.

3 Datasets and Experiments

3.1 Datasets

Three WCE videos (63495, 56431 and 40753 frames) of three persons were collected from Olympus Corporation, Japan. Image size is 288×288 . However, experiments were done on reduced subsets. Each video is manually divided into two (training and testing) independent sets, i.e., (15749, 15850) for Video-1, (12721, 13160) for Video-2, and (12185, 13530) for Video-3. Three training sets were merged into a combined set (40655 frames). A unique random algorithm is then applied to the initial sets to come up with the final training (5717 frames) and testing sets (5085 from Video-1, 5049 from Video-2, and 4707 from Video-3). "Holdout method" [9] is used to verify classification accuracy in Stage-1.

Table 1. Isolating HCN frames by HSV color histogram (Stage-1)

Data	Test DB Size (Org. info. frames)	Output frames (Info. frames)	Preserve ratio(%)	Classification accuracy(%)
Video-1	5085(4647)	4781(4647)	100.00	99.95
Video-2	5049(2936)	3443(2932)	99.86	99.10
Video-3	4707(3819)	4397(3775)	98.84	97.15

Fig. 2. Segmentation of bubble structures by the proposed GLT feature. Results are obtained by (a) Video-1, (b) Video-2, and (c) Video-3, respectively.

3.2 Results and Performance Analysis

HCN frames are isolated in Stage-1 using local color histogram with SVM classifier as mention in 2.3. Kernel based SVM uses 10 fold cross-validation over the combined training set with 99.82% accuracy including $C = 8.0$ and $\gamma = 2.0$. This eliminates overtraining problem. Table 1 and Fig. 3(a) show classification results. Clearly, the local color histogram preserves more than 98% of the original informative frames during Stage-1 processing. Note that we excluded food frames without turbid liquid (Fig.1(x)) from the training process because of their limited numbers. SB frames are isolated in Stage-2 using segmentation and a threshold operation. Various parameters in the segmentation scheme were: $n = \{1, 2, 3, 4, 5\}$, $k = \{0, 1\}$, $\sigma = \{3.0, 5.0\}$, $L = 3$, Th_1 , Th_2 (automatically), and Th_{area} (manually). Figures 2(a), (b), and (c) show segmentation results by the proposed method. Obviously, GLT based energy feature is effective for segmenting bubble patterns of different sizes and shapes at different backgrounds, e.g., turbid, normal tissue, and lumen. Segmented area of each frame is thresholded to decide if the frame is useful or not. Two evaluation criteria are used to justify the detection performance:

$$\text{Preserve ratio, } R_p = \frac{N_{di}^1}{N_{ti}} \times 100\% \quad (7)$$

$$\text{Recall, } R = \frac{N_{di}^2}{N_{ti}} \times 100\% \quad (8)$$

where N_{di}^1 and N_{di}^2 are the detected informative frames in Stage-1 and Stage-2 and N_{ti} is the true number of informative frames in a single test dataset. Two

Table 2. Informative frame detection by GLT based norm-1 energy feature(Stage-2)

Video	Stage-1 output frames by HSV local color histogram	Detection Performance Recall, R(%), and False Positive(FP)					
		Wavelet		Gabor		GLT	
		FP	R	FP	R	FP	R
Video-1	4781	30	15.04	30	35.03	35	99.01
Video-2	3443	169	91.09	187	94.98	166	95.91
Video-3	4397	201	83.33	233	96.55	217	97.52
Average performance		133	63.15	150	75.52	139	97.48

Fig. 3. Detection Performance. Feature- and dataset-wise (a) Preserve ratio/HCN isolation accuracy, (b) Recall rate, (c) Average false positive, and (d) Average recall, and (e) Segmentation results of informative frames for Video-1 images.

conventional approaches based on Gabor [7] and wavelet [11] transform were also adopted in Stage-2. The detection results and comparison are shown in Table 2 and Fig. 3(b), (c), and (d). Clearly, the proposed method obtains the highest average recall (97.48 %), when compared with Gabor-(75.52 %) and wavelet features (63.15%) in combination with same color feature.

3.3 Discussion

The proposed method works well, while it requires some parameters (n , k , σ , and L) setting empirically. Observation shows that higher degree filters cause unreliable segmentation than that with higher order filters. While the proposed GLT feature consistently produce better segmentation, Gabor and wavelet features cause erroneous segmentation for Video-1 (Fig. 3(e)) frames. This is because GLT usually produces high contrasted features due to high correlation with bubble patterns. This makes Otsu’s [10] method to select clear threshold for automatic bubble patterns segmentation even in case of smoother frames (Video-1). Note that all methods use similar parameters (real filters, decomposition depth ($L = 3$), feature neighborhood ($W = 16 \times 16$), orientation (4 to 5)) for fair comparison.

4 Conclusion

We developed a novel cascade scheme for informative frame detection from WCE images. HSV local color histogram and GLT based norm-1 energy feature were proposed for isolating non-informative frames. SVM classifier is used in Stage-1 to isolate HCN frames with high preserve ratio ($> 98\%$) of the informative frames. SB frames were isolated in Stage-2 by automatic segmentation and thresholding. Three datasets were selected from the three videos and validated the classifier performance using Holdout method. Experiments with approximately 5000 frames from each video showed 97.48%, 75.52%, and 63.15% average recall, and 130, 150, and 139 frames of average false positives by the proposed, and other two methods (Gabor- and wavelet based). Possibility of using color and texture features into a single framework, and the unsupervised technique will be explored in future.

References

1. Vilarino, F., Spyridonos, P., Pujol, O., Vitria, J., Radeva, P.: Automatic detection of intestinal juices in wireless capsule video endoscopy. In: Proc. IEEE/ICPR, vol. 4, pp. 719–722 (2006)
2. Vilarino, F., Kuncheva, L., Radeva, P.: Roc curves and video analysis optimization in intestinal capsule endoscopy. Pattern Recognition Letters 27, 875–881 (2006)
3. Manjunath, B.S., Ohm, J.-R., Vasidevan, V.V., Yamada, A.: Color and texture descriptors. IEEE Trans. circuit and systems for video technology 11(6) (June 2001)
4. Swain, M., Ballard, D.: Indexing via color histograms. Int. J. Comput. Vision 7, 11–32 (1991)
5. Cimminiello, L.S.N., Neri, A.: Keypoints selection in the gauss laguerre transformed domain. In: Proc. BMVC, vol. II, pp. 539–548 (2006)
6. Jacovitti, G., Neri, A.: Multiresolution circular harmonic decomposition. IEEE Trans. Signal Process 48, 3242–3247 (2000)
7. Nestares, O., Navarro, R., Portilla, J., Tabernero, A.: Efficient spatial domain implementation of a multiscale image representation based on gabor functions. J. Electronic Imaging 7, 246–248 (1998)
8. Mercier, G., Lennon, M.: Support vector machines for hyperspectral image classification with spectral-based kernels. In: Proc. IEEE symposium in Geoscience and Remote Sensing, July 21–25, 2003, vol. 1, pp. 288–290 (2003)
9. Gonzalez, R., Woods, R., Eddins, S.L.: Digital Image Processing using Matlab, 3rd edn. Pearson Prentice Hall, New Jersey (2004)
10. Otsu, N.: A threshold selection method from gray-level histograms. IEEE Trans. Systems Man and Cybernetics 9, 62–66 (1979)
11. Mallat, S.: The theory for multiresolution signal decomposition: the wavelet representation. IEEE Trans. Patt. Anal. Mach. Intell. 11 (1989)