Cluster-based Sampling and Ensemble for Bleeding Detection in Capsule Endoscopy Videos

Mohamed Abouelenien[a], Xiaohui Yuan[a]*, Balathasan Giritharan[a], Jianguo Liu[b], Shoujiang Tang[c]
[a] Department of Computer Science and Engineering, University of North Texas, Denton, TX, U.S.A.
[b] Department of Mathematics, University of North Texas, Denton, TX, U.S.A.
[c] University of Mississippi Medical Center, Jackson, MS, U.S.A.

Abstract – We present a cluster-based sampling and ensemble method to learn from large, imbalanced data set for bleeding detection in CE videos. Our method selects training examples randomly according to the data distributions derived from clustering. Multiple training sets are created such that data balance is restored. The sampling probability is proportional to the cluster distribution, and within each cluster the probability of a sample being selected is proportional to the distance to the center of the cluster. Classifiers are evaluated to compute performance-based weights and the prediction is made by aggregating decisions from the ensemble. Experiments were conducted using 8 annotated full-length videos. The cluster-based sampling provides training examples that preserve the innate data distribution with much less number of instances. Our experiments demonstrate that ensemble coupled with cluster-driven sampling achieves superior sensitivity and very competitive specificity. The one way ANOVA analysis reveals that our method greatly outperforms conventional SVM method.

Key Words – Capsule Endoscopy, Classification, Clustering, Video Analysis

I. Introduction

Capsule endoscopy (CE) is an imaging technology that has revolutionized our ability to visualize the entire small intestine non-invasively. The imaging component of this system is a vitamin-sized capsule that is composed of a color CMOS camera, a battery, a light source and a wireless transmitter. The camera acquires two pictures every second for approximately eight hours and generates 256×256 images transmitted to a recording device worn by the patient. It has been used to examine the entire small intestine non-invasively and is used mainly to diagnose lesions beyond the reach of conventional push endoscopy and colonoscopy. Its clinical applications have shown great improvement in diagnostic yield for bleeding sources in patients with obscure GI bleeding, and in diagnosing and localizing the source of blood loss. More information and clinical applications of CE can be found in [1] and the reference therein.

Among many efforts in computer aided diagnosis with CE videos, bleeding detection has been investigated the most due to its clinical importance. The "Suspected Blood Indicator" function by the Given Imaging, a CE manufacturer, provides the capability of detecting blood in video frames. A study by Liangpunsakul et al. [20] showed that the overall sensitivity and accuracy of SBI were 25% and 34.8%, respectively. It exhibits better performance for active bleeding lesions in the small bowel with reported sensitivity and accuracy of 81.2% and 83.3%. This deficiency motivated studies in automatic bleeding detection. Color feature is adopted in many detection studies [13, 10] and texture features are used in applications particularly for detecting heterogeneous objects, e.g., ulcer and polyps [2, 29, 4]. The combination of color and texture has also been heavily experimented [22, 8]. On the other hand, neural networks [27, 29, 21], Support Vector Machines (SVMs) [24, 23], and thresholding [29] are used to make decisions. Despite the encouraging improvement, many previous studies were evaluated with a small number of samples and to the best of our knowledge no performance was reported with respect to entire videos. An important question awaits investigation: "Given relatively small number of positive examples from CE videos, how to train learning algorithms to achieve minimal false negative detections?"

In this article, we present a novel method to learn from large, imbalanced data set for bleeding detection in CE videos. Our method uses a cluster-based sampling strategy to select training examples and create multiple distinct training sets such that data balance is restored. Using each training set, a classifier is built and evaluated with the rest of the examples to compute a performance-based weight. The prediction to a new instance is the weighted aggregation of decisions from all classifiers. With downsampling, the size of each training set is manageable by classifier. In

Corresponding Author: Xiaohui Yuan, xiaohui.yuan@unt.edu addition, since multiple training sets are created with randomly selected examples, the loss of information is suppressed such that the generalization performance is greatly improved.

Our contribution to the bleeding detection in CE videos is twofold: 1) a novel sample selection method that analyzes sample distribution and intelligently selects subsets for training such that a close representation of the data distribution is reached as well as data balance is recovered; and 2) a performance-driven ensemble learning strategy that circumvents possible loss of information due to downsampling by weighting trained base classifiers with their performance measures. Our method provides a framework that integrates multiple image features and addresses the imbalance problem in the real-world CE video analysis with a statistically plausible solution. From the experimental point of view, our extensive experiments conducted with 8 full-length videos reveal the possible drawbacks of training classifier with improper manually-selected data set and demonstrated a feasible remedy using cluster-based sampling and classifier ensemble.

The rest of this article is organized as follows: Section 2 reviews the state-of-the-art methods in bleeding detection from CE videos. Section 3 describes our method that uses cluster-based sampling and ensemble (CSE) to address the difficulties arose from large imbalanced data sets. Section 4 presents the experimental results using complete CE videos and discussions. Section 5 concludes the paper.

II. Related Work

Automatic detection of obscure bleeding in CE videos has been studied and Table 1 summarizes 12 related works. Despite different features and classification methods used, the experimental data and performances vary greatly. Among these studies, results in eight studies were generated from experiments using 1000 examples or less [18, 2, 29, 11, 22, 8, 12, 14, 15]. Four studies [17, 10, 19] used moderately larger number of examples. Comparing to the number of frames available in a CE video (approximately 50,000), however, the training data set size is much smaller. Ideally, if the training set is well-selected and unbiased, the classifier can achieve satisfactory generalization performance. It is unclear how the samples are selected and if the cohort formed represents the true data distribution.

Like many medical diagnosis applications, CE videos are full of negative examples and much less number of positive examples. That is CE videos are imbalanced in nature [29]. Classification with imbalanced data sets has been a well-known problem in many other fields of applications. The abundant examples from the majority class and significantly inadequate number of examples from the minority class affect the classification performance when applied to examine the entire video. The challenges lie in the misrepresented data distribution.

Methods have been developed to address the challenges of imbalanced classification from both data and algorithm aspects. Data-centered methods rely on resampling to achieve equal or approximately equal number of instances from both classes [7, 16]. The Synthetic Minority Oversampling Technique proposed by Chawla et al. gained much popularity in generating instances for the minority class [7]. The arguments, however, are the increase of data size which could potentially exceed the capacity of our modem computing power; whereas the downsampling techniques are facing critiques on possible loss of information.

In algorithm-centered methods, assumptions are made in favor of the minority class. There are many real-world applications that support such assumption. For instance, in medical diagnosis and surveillance, the rare cases (samples of the minority class) carry significantly greater values than the ordinary instances. To implement this assumption in algorithms, biased decision weights are commonly employed [9]. Another thrust of efforts is to construct classifiers using training instances from only the majority class, i.e., one-class learning [26]. The rationale is that ample instances from the majority class provide a well-defined class boundary. The difficulty lies in the subjectivity of the preference and the magnitude of the bias toward the minority class.

Combination of data-centered and algorithm-centered methods has also been investigated [3]. Research has explored generating multiple data sets and aggregating cost-sensitive classification. It was claimed improved performance in both handling large data set and overall accuracy. All submissions should follow the guidelines of this journal for submission.

III. Methodology

Our method uses cluster-based sampling and ensemble and consists of three steps: feature extraction, data rebalancing via cluster-based sampling, and ensemble classification. Fig. 1 illustrates the diagram of our method. The rationale of downsampling is that

Table 1 Experimental data and detection outcomes. ‘-’ indicates not reported in the paper.

	Data set size			Performance		
Reported studies	Total	Abnormal	Normal	Sensitivity	Specificity	Accuracy
Kodogiannis and Boulougoura [14]	140	35	35	-	-	95.7%
Kodogiannis and Lygouras [15]	140	35	35	-	-	97.1%
Vilarino et al. [29]	400	100	300	-	-	95.5%
Coimbra and Cunha [8]	1000	-	-	-	-	87%
Lau and Correia [17]	1705	577	1128	88.3%	-	-
Li and Meng [18]	60	30	30	65.2%	82.5%	-
Li and Meng [19]	3600	1800	1800	88.8%	84.2%	-
Li and Meng [22]	400	200	200	91%	93%	-
Jung et al. [10]	2000	1000	1000	92.8%	89.5%	-
Barbosa et al. [2]	204	100	104	98.7%	96.6%	-
Karargyris and Bourbakis [11]	-	20	30	75%	73.3%	-
Karargyris and Bourbakis [12]	50	10	40	100%	67.5%	-

samples of the majority class are of great number and are likely to be redundant. Downsampling the majority examples balances the two classes. The possible loss of information in the process of downsampling could be leveraged via bootstrap aggregating classifiers, which are trained with balanced examples from both classes.

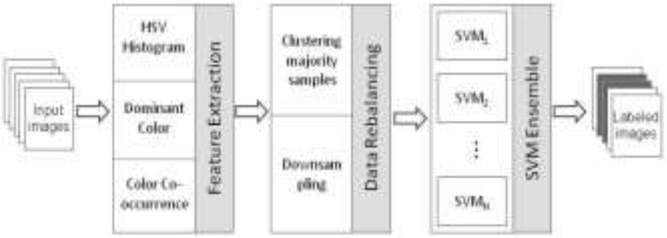

Fig. 1: Our method consists of three steps: Feature extraction, Data rebalancing, and SVM Ensemble. The arrows depict the data flow.

3.1 Feature Extraction

We employ three image features in our method: color histograms, dominant color, and color co-occurrence.

Color Histogram (CH)

Color histogram is widely used due to its concise representation of color information. Among many color spaces, HSV separates the luminance from chromaticness. It is usually represented with a hexacone, the central vertical axis of which denotes the luminance. Hue is defined as an angle relative to the red and ranges in [0, 2π]. Saturation is measured as a radial distance from the central axis of the hexacone. Its chromatic components describe color in a way that is most suitable to bleeding detection [23]. Hence, video frames are converted to HSV color space and each color component is normalized to [0, 1] and sampled with 256 bins.

Dominant Color (DC)

The dominant color consists of eight representative colors, variances for each color, and their percentages in the image [25]. The descriptor is presented as a vector in the following format and the total percentages of the colors in the image sum to 1.

$$DC = \{c_i, v_i, p_i\}, \text{and } i = \{1, \ldots, 8\} \qquad (1)$$

where c_i is the i-th dominant color, p_i is its percentage, v_i is the color variance.

For each video frame, colors are clustered and the mean color is used to represent each cluster. This results in a much smaller number of colors. The variance of dominant colors is computed for bleeding and non-bleeding frames. Despite possible information overlap with CH, DC delivers a more concise color description and suppresses the color variance as well as the number of colors.

Color Co-occurrence (CC)

The color co-occurrence matrix follows the classical computation of co-occurrence matrix and contains the frequency of color pair within a pre-defined distance, i.e., (Δx, Δy). In an 8-bit color image, there are possible 2^{24} colors. To reduce the matrix size, we quantize the color into a set of representative ones. In addition, to eliminate rotation variance in the image plane, we omit the direction of the spatial location of two pixels and only keep track of the pixel distance, i.e., $d = \Delta x^2 + \Delta y^2$. Because the matrix is symmetric with respect to the major diagonal line, our feature vector only uses the components in the upper triangle matrix.

3.2 Cluster-based Sampling

The imbalance ratio in CE videos is usually significant, which can be as much as 1000:1 (refer to Table 2 for examples). Randomly downsampling the majority class to rebalance the training data could lose critical instances; whereas upsampling the minority class results in much larger data set that exceeds the capacity of modern computers. We propose a downsampling strategy based on unsupervised clustering of the data set followed by a probability-driven sampling from each cluster to preserve the geometric structure of the data set with less number of instances. Our sampling strategy is inspired by the observation that CE video frames within a temporal neighborhood are highly correlated. That is, these data points are close to each other in feature space and, hence, form a cluster. Frames that contribute to a cluster are not necessarily temporally adjacent. Retaining a number of samples from each cluster could maximize the preservation of the original data distribution with a smaller set of data points. Our hypothesis is that the instances close to the centroid of the cluster are less influential to the classifier than the ones close to the boundary of the cluster. Hence, we sample each cluster according to its innate distribution area and the sample distance to the centroid. The sampling probability is proportional to the cluster distribution, and within each cluster the probability of a sample being selected is proportional to the distance from the sample to the cluster center. Let C_j be the j-th cluster with n_j samples. The probability of a sample c, $c \in C_j$, being selected is computed as follows:

$$\mathrm{p(c)} = \frac{\mathrm{n_j}}{\sum_{\mathrm{j=1}}^{\mathrm{J}} \mathrm{n_j}} \cdot \frac{\|\mathrm{c}-\bar{\mathrm{c}}_\mathrm{j}\|}{\sum_{\mathrm{i=1}}^{\mathrm{n_j}} \|\mathrm{c_i}-\bar{\mathrm{c}}_\mathrm{j}\|} \tag{2}$$

where $\bar{c}_j$ is the mean of the j-th cluster and J is the number of clusters. Function $\|c-\bar{c}_j\|$ computes the distance of c to $\bar{c}_j$. The first term in Eq. (2) gives the overall probability of samples selected from C_j. The second term decides the selection probability of samples inside C_j. The denominators ensure the sum of probabilities of all samples in the majority class is unit.

Depending on the nature of data distribution, the within-cluster probability can be modified. For instance, inverse multi-variant Gaussian function provides gradual descent of probabilities from the centroid. For simplicity, we use Euclidean distance in our probability modeling.

In our method, k-means clustering is employed. For each cluster C_j, a probability, p(c), is computed and assigned to each instance c. The cumulative probability function $P(C_j)$ of the cluster integrates the probability of all instances in C_j. In our sampling process, we compare a uniform random number r with $P(C_j)$ and select the sample that defines the range. Since the overall probability of a cluster is factored by the number of its instances, the random number is generated in the range of [0, T] and $T = \frac{n_j}{\sum_j n_j}$. An instance is removed from the cluster to prevent duplication. Removing an instance changes the cumulative probability P and the random number upper bound T. Hence P is updated in the iterations. Algorithm 1 summarizes our cluster-based sampling method.

Algorithm 1: Cluster-based Sampling

```
Generate J clusters from $\mathcal{C}$ using k-means algorithm
**for all** $C_j \subseteq C, j \in \{1, 2, \dots, J\}$
  Compute $n_j$ and $\bar{c}_j$ for cluster $C_j$
  **for all** $c_i \in C_j$
    Compute $p(c_i)$ using Eq. (2)
  **endfor**
  $\tilde{n}_j \leftarrow n_j\$$
  $P \leftarrow P(C_j)$ and $T \leftarrow \frac{\tilde{n}_j}{\sum_j n_j}$
  **for** $l \leftarrow 1$ to $\frac{N \cdot n_j}{\sum_j n_j}$
    Generate uniform random number $r \in [0, T]$
    $\tilde{C}(l) \leftarrow c_l$ such that $r \in P(c_l)$
    $C_j \leftarrow C_j - \{c_l\}$ and $\tilde{n}_j \leftarrow \tilde{n}_j - 1$
    $P \leftarrow P(C_j)$ and $T \leftarrow \frac{\tilde{n}_j}{\sum_j n_j}$
  **endfor**
**endfor**
```

Fig. 2 illustrates a visualization of our clustering result to the majority examples in one training video. Ten clusters are used in this example. Out of 4096 feature components, 2 are retained to accommodate 2-dimensional visualization space. Our feature selection is based on Fisher discriminant analysis and retains the two that give the best separation of the clusters. It is clear that the instances close to the origin are showing strong greenish color; whereas the far right of the x-axis depicts reddish color. The top of the space are filled with instances with wiggling folds and the lower part of the space is dominated with images with no significant texture.

3.3 Aggregation of Classifiers Trained with Multiple Features

A major concern of downsampling the majority class to rebalance the training set is the potential of missing critical instances and hence results in lower generalization performance. Without knowledge of the majority class distribution and the spatial relation of the two classes, a sampling process cannot guarantee that the downsized data set represents the information of the available data for the advantage of classification. Ensemble classifier is promising in that multiple dissimilar downsampled data sets provide balanced training set with comprehensive coverage of the majority class data distribution. The idea of bootstrap aggregating with SVM was broadly used in many problems and was originally developed to improve the estimation accuracy of weak classifiers [5,28]. The classifier ensemble labels an instance by aggregating decisions of all classifiers.

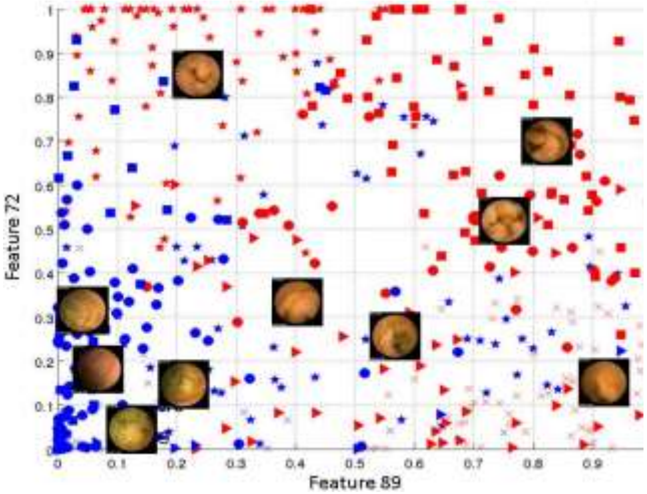

Fig. 2 2D distribution of ten clusters. One image is shown for each cluster.

In our classifier ensemble, several SVMs are trained independently using training set created by our sampling method. Let $V_j, j = \{1, \dots, M\}$, denotes a set of samples that are created using Algorithm 1 from a training set V. Let f_j be the discriminant function obtained through SVM learning using V_j. The label for an unseen instance, x_i, is computed by aggregating the decisions of the trained classifiers as follows:

$$\mathrm{L(x\_i)} = \mathrm{sign}\left(\sum_{j=1}^{M} \alpha_j \ f_j(x_i;\ w, b, \epsilon)\right) \qquad (3)$$

The α_j is a weight to the discriminant function f_j and is proportional to the generalization performance of f_j to the data set V:

$$\alpha_j = \frac{\mathcal{P}(f_j)}{\sum_{j=1}^{M} \mathcal{P}(f_j)} \qquad (4)$$

where $\mathcal{P}(f_j)$ denotes the metric function for evaluating the performance of f_j. The choices of $\mathcal{P}$ are many. In our implementation, we used sensitivity for $\mathcal{P}$.

An SVM is a hyperplane that maximizes the class margin. Suppose data points are represented as $\{(x_1, y_1), (x_2, y_2), \dots, (x_l, y_l)\}$, $y_i \in \{1, -1\}$, and each x_i is an N-dimensional vector. The hyperplane takes the form of $w \cdot x - b = 0$. In the linearly non-separable classification problems, soft margin SVM allows, but penalizes, examples that fall on the wrong side of the decision boundary. A general form of the quadratic programming problem with soft margin and nonlinear classifier is as follows:

$$\begin{gathered} \min \frac{1}{2}\|w\|^2 + C\xi^T e \\ \text{subject to } \mathrm{y_i}(\mathrm{w} \cdot \phi(\mathrm{x_i}) - \mathrm{b}) \geq 1 - \xi_\mathrm{i} \\ \text{and } \xi_i \geq 0, \qquad 1 \leq i \leq l \end{gathered} \qquad (5)$$

where ξ denotes training error and C provides a weighting between regularization term and the training error. The function ϕ is a mapping from $\Re^n$ to a higher dimensional space. Details on SVM and its implementation can be found in [6].

IV. Experimental Results and Discussion

4.1 Data Preparation

Eight CE videos were annotated by a gastroenterologist in our team. Frames were extracted from the raw video and converted into images in JPEG format. The frame size is 256×256 or 512×512 (see Table 2) with 8-bit color depth. The bounding black region in CE frames was removed, which is outside the field-of-view of the CE camera. A 3 by 3 average filter was used to suppress random noise.

Over-exposed (usually at the start of the video before the CE device enters the digestive tract) and very dark frames (usually the ones picturing feces) were removed since these frames are easily identifiable and could present erroneous information to the classifier training. Examples of such extreme frames are shown in Fig. 3. An empirical threshold was used: if more than 65% of the pixels within the field of view are either black or white, the frame is considered as an extreme frame. The elimination of extreme frames was applied to each video before image features were extracted.

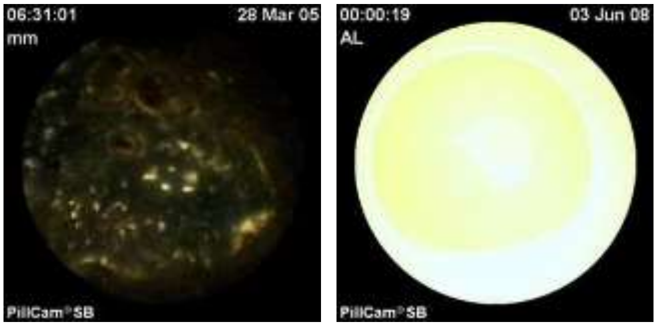

Fig. 3 Examples of over-exposed frame (a) and dark frame (b).

Table 2 lists the number of frames and imbalance ratio of the videos. Out of the 8 videos, frames from 2 videos were used as training data for classifiers and the other 6 videos were used as testing data. The data sets from the two training videos are representative in that both videos are highly imbalanced with a large number of negative examples (i.e., frames with no bleeding signs).

Table 2 Properties of the videos used in our experiments.

	Video index	**Total # frames**	**Bleeding frames #**	**Ratio**	**Frame size**
Test	1	53,526	52	1030:1	512 × 512
	2	55,460	691	80:1	256 × 256
	3	53,219	590	94:1	256 × 256
	4	56,721	51	1112:1	512 × 512
	5	54,522	588	93:1	256 × 256
	6	52,340	82	638:1	256 × 256
Train	7	51,450	465	111:1	256 × 256
	8	56,457	33	1711:1	256 × 256

For performance evaluation, we adopted widely used metrics: sensitivity ($\mathcal{S}_e$) and specificity ($\mathcal{S}_p$):

$$\mathcal{S}_\mathrm{e} = \frac{\mathrm{TP}}{\mathrm{TP+FN}} \text{ and } \mathcal{S}_\mathrm{p} = \frac{\mathrm{TN}}{\mathrm{TN+FP}} \qquad (6)$$

4.2 Effect of Manually Selected Small Training Samples

In our first experiment, we trained SVMs (based on libSVM [6]) with 800 images, in which 400 are CE frames showing obscure bleeding and the rest show normal tissues. Images of both classes were selected by our gastroenterologist. In these experiments, we used the color histogram and the raw pixel value in both RGB and HSV color spaces. Based on our previous studies [23], both polynomial and radial basis produced satisfactory results in classifying CE frames. In this study, we used radial basis function kernel with variance empirically selected as 0.0013.

Table 3 presents the results of our classifiers using manually selected balanced training data sets. The size denotes the percentile of examples used in training SVMs. The rest was used in testing. For example, for 80%, 640 images were used in training and about 160 images were used in testing. The training examples were selected randomly in each trial. The experiments were repeated for 10 trials, and the average results were reported together with standard deviations (STD).

The highlighted results illustrate the best outcomes. Between histogram and pixel value, there is no significant difference in performance. With the same color space, the two metrics give very comparable results. However, histogram-based feature provides more concise description of the view. The HSV color representation demonstrates better results than those of the RGB color representation. The mean sensitivity and specificity as well as the standard deviations are in the close range of 90%. If we aim to maximize the performance, the HSV histogram gives the best overall detection rates.

Table 3 The performance of bleeding detection using balanced training examples. The STDs are listed in parenthesis.

Feature	**Training data**	$\mathcal{S}_e$	$\mathcal{S}_p$
RGB Histogram	80%	93.8% (2.4)	82.6% (4.5)
	60%	92.6% (2.3)	78.3% (3.4)
	40%	93.1% (2.5)	78.0% (3.4)
RGB raw Pixel value	80%	91.8% (1.7)	80.7% (5.2)
	60%	91.8% (2.9)	80.7% (3.3)
	40%	90.4% (3.3)	77.6% (2.8)
Mean		92.5%	79.7%
HSV Histogram	80%	96.8% (1.8)	93.8% (1.4)
	60%	96.2% (0.9)	**93.9%** (2.9)
	40%	95.1% (1.1)	89.3% (5.0)
HSV raw Pixel value	80%	**97.5%** (2.9)	86.3% (1.8)
	60%	95.6% (3.4)	86.9% (3.5)
	40%	92.6% (4.6)	87.0% (2.1)
Mean		95.6%	89.5%

In our second experiment, we applied the trained classifier to the CE videos. Table 4 reports the performance of our classification. Six videos were used to test the previously trained classifier. The mean sensitivity is at 60.6%; whereas the specificity is at 88.1%. Comparing to the previous 95.6% (sensitivity) and 89.53% (specificity), the degradation is significant. A major factor is the misrepresentation of data distribution from the training data set.

Table 4 Results HSV histogram SVM classifier, without re-balancing.

Video	$\mathcal{S}_e$	$\mathcal{S}_p$
1	61.5%	88.3%
2	60.2%	88.1%
3	59.3%	89.1%
4	61.2%	87.9%
5	62.7%	86.9%
6	58.5%	88.4%
Mean (STD)	60.6% (1.5)	88.1% (0.7)

4.3 Performance Analysis

In constructing classifier ensemble, three sets of image features were extracted: Color histogram (CH), Dominate color (DC), and Color co-occurrence (CC). The parameters used for these image features are as follow: in color histogram feature, we used HSV (Hue-Saturation-Value) space with 256 bins for each component; in dominant color, 16 most prominent colors were selected to compute the feature vectors, and the radius used in color co-occurrence is selected as 5. For each set of features, three SVMs were generated. The selection of training images follows our cluster-based sampling method in Algorithm 1. The number of clusters was empirically chosen as 40. In this study, we employed radial basis function as the kernel for SVMs with the variance at 0.0024.

Table 5 lists the results of method. Each row presents the test performance with one entire video. The average performance and its standard deviation are reported. The sensitivity and specificity for each image feature are average of three individually trained SVMs. In contrast to the results in our second experiment (as shown in Table 4), using the same image feature and classifier, the outcomes are better with the sensitivity in the lower 70% and specificity in the lower 90%; whereas SVM trained with manually selected balanced data resulted sensitivity in the lower 60% and specificity in the upper 80%. The testing results using the other two image features are slightly better in sensitivity. The ensemble outperformed all the classifiers with average sensitivity at 81.4% and average specificity at 93.3%.

Table 5 Results from average SVM using rebalanced data sets.}

Video	CH		DC		CC		CSE	
	S_e	S_p	S_e	S_p	S_e	S_p	S_e	S_p
1	69.2	93.4	75.0	91.5	76.9	90.1	82.7	92.2
2	72.2	92.1	76.6	90.3	77.4	87.9	78.3	94.3
3	70.3	92.6	75.5	94.2	73.2	90.5	82.0	93.1
4	72.5	94.9	76.5	92.3	72.5	91.5	80.6	94.1
5	68.9	93.7	70.4	93.1	76.5	90.7	81.4	92.5
6	71.1	91.2	75.6	94.2	70.7	92.3	83.1	93.6
Mean	70.7	93.0	74.9	92.6	74.5	90.5	**81.4**	**93.3**
STD	2.3	1.7	5.3	2.4	7.7	2.2	3	0.7

Fig. 4 illustrates the box plots of the sensitivity and specificity between our method and the SVM trained with manually selected data set. It is interesting that even without constructing an ensemble the cluster-based sampling improves the classifier performance. The column CH and SVM was resulted from trained SVMs using the same image features (color histograms). The mean sensitivity and specificity are improved from 60.6% and 88.1% to 70.7% and 91.2%, respectively. Using one-way ANOVA analysis of the results using cluster-based sampling and manually selected balanced training data, the p-values of the sensitivity and specificity are 4.22E-7 and 1.27E-5, respectively. This indicates a fairly significant improvement originated from our sampling algorithm.

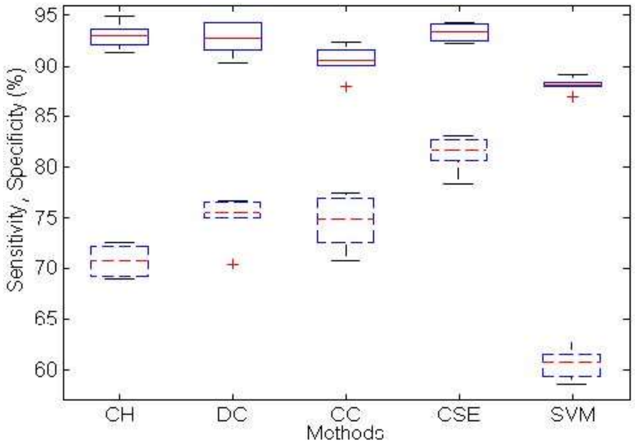

Fig. 4 Box plots of sensitivity (dash line) and specificity (solid line) of our method and the SVMs trained with manually selected data set.

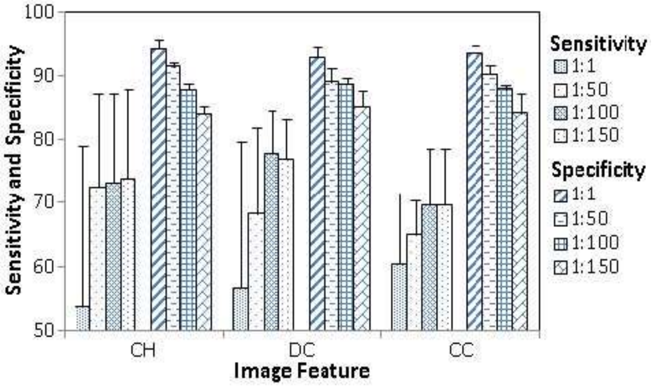

Fig. 5 Sensitivity and specificity of SVMs trained with weighted samples.

Fig. 5 shows the average performance over 6 videos of the conventional SVMs trained with weighted samples. Sensitivity and specificity are grouped for each image feature. The minority instances are assigned with greater weights to achieve a balance For instance, in the case of 1:50, all minority instances are weighted 50; whereas all majority instances are weighted 1. The average imbalance ratio in our data set is close to 1:150. With the increase of weights, sensitivity improves and specificity drops. When small weight is used, e.g., 1:1, the trained SVMs exhibit poor sensitivity and greatest variance. By applying large weight to minority instances, the sensitivity can be boosted to upper 70s with dominant color feature in sacrifice of specificity. In addition, it is evident that the improvement of sensitivity by increasing weight levels out as weight increases.

Table 6 reports the F-test results and p-values using pair-wise one-way ANOVA analysis. Each method is compared against the CSE method and SVM classifier. The results produced using CSE as reference show great improvement in sensitivity. The F-test and p-value in comparison to SVM are 479.79 and 8.8E-10, respectively. The specificity difference between CSE and the three features is not significant. However, as shown in Fig. 4, the specificity of CSE is more consistent than other methods. From the results produced using SVM as reference, it is clear that dominant color achieved most improvement in sensitivity (the F-test and p-value are 161.31 and 1.71E-7, respectively) and color histogram outperformed the other images features (the F-test and p-value are 64.15 and 1.17E-5, respectively), which is coincide with our observation from Fig. 4.

Table 6 F-test results and p-values of one way ANOVA analysis.

CSE as reference		**CH**	**DC**	**CC**	**SVM**
S_e	F-test	128.5	29.6	26.1	479.8
	p-value	5E-7	2.9E-4	4.6E-4	8.8E-10
S_p	F-test	0.3	1	15.9	129.4
	p-value	6.3E-1	3.6E-1	2.6E-3	4.8E-7

SVM as reference		**CH**	**DC**	**CC**	**SVM**
S_e	F-test	133.2	161.3	116.8	479.8
	p-value	4.2E-7	1.7E-7	7.8E-7	8.8E-10
S_p	F-test	64.2	41.4	12.3	129.4
	p-value	1.2E-5	7.5E-1	5.6E-3	4.8E-7

V. Conclusion

In this paper we describe a cluster-based sampling and ensemble method to learn from large, imbalanced data set for bleeding detection in CE videos that minimizes false negative decisions. Our method selects training examples randomly according to unsupervised clusters and creates multiple training sets such that data balance is restored. The sampling probability is proportional to the cluster size, and within each cluster the probability of a sample being selected is proportional to the distance to the center of the cluster. The prediction to a new instance is the weighted aggregation of decisions from all classifiers. With downsampling, the size of each training set is greatly reduced. In addition, since multiple training sets are created with randomly selected examples, the loss of information is greatly suppressed.

Based on our experiments, the following conclusions can be drawn. First, the cluster-based sampling provides training examples that preserves the innate data distribution with much less number of instances. Using the same number of training instances and the same image features, it is evident that the sampling algorithm contributes to the improvement of the classifier performance.

Second, the classifiers trained with different image features achieved much improved results using sampled data set. The dominant color and color co-occurrence give better sensitivity and the color histogram gives higher specificity.

Third, the ensemble integrates individually trained SVMs and achieves superior sensitivity and very competitive specificity. The one way ANOVA analysis illustrates that our method greatly outperforms conventional SVM method. The possible loss of information due to downsampling is successfully circumvented.

Last, we demonstrated the generalization degradation of using misrepresented training data set by constructing an SVM with manually selected, balanced data set of 800 images and applying the trained classifier to full-length videos. The testing performance degraded significantly. This is because the training data misrepresent the true data distribution of the CE video frames. Such misrepresentation is exaggerated when there is a large majority data set and only a small number of instances are selected for training.

REFERENCES

[1] D. G. Adler and C. J. Gostout, “Wireless capsule endoscopy,” Hospital Physician, vol. 39, no. 5, pp. 14–22, 2003.

[2] D. J. C. Barbosa, J. Ramos, and L. S. Carlos, “Detection of small bowel tumors in capsule endoscopy frames using texture analysis based on the discrete wavelet transform,” in the Int’l Conf. of the IEEE Engineering in Medicine and Biology Society, Aug 2008.

[3] G. E. A. P. A. Batista, R. C. Prati, and M. C. Monard, “A study of the behavior of several methods for balancing machine learning training data,” in SIGKDD Explorations, vol. 6, 2004, pp. 20–29.

[4] N. Bourbakis, “Detecting abnormal patterns in wce im¬ages,” in the 5th IEEE Symposium. on Bioinformatics and Bioenqineering, Oct 2005, pp. 223–238.

[5] L. Breiman, “Bagging predictors,” Machine Learning, vol. 24, no. 2, pp. 123–140, Aug 1996.

[6] C.C. Chang and C-J. Lin, “Libsvm: a library for support vector machines, software available at http: //www.csie.ntu.edu.tw/˜cjlin/libsvm,” 2010.

[7] N. V. Chawla, K. W. Bowyer, L. O. Hall, and P. Kegelmeyer, “SMOTE: Synthetic minority over-sampling technique,” Journal of Artificial Intelligence and Research, vol. 16, pp. 321–357, 2002.

[8] M. T. Coimbra and J. Cunha, “MPEG-7 visual descriptors contributions for automated feature extraction in capsule endoscopy,” IEEE Transactions on Circuits and Systems for Video Technology, vol. 16, no. 5, pp. 628–637, May 2006.

[9] Y. Huang and S. Du, “Weighted support vector machine for classification with uneven training class sizes,” the Int’l Conf. on Machine Learning and Cybernetics, vol. 7, pp. 4365–4369, 2005.

[10] Y. S. Jung, Y. H. Kim, D. H. Lee, and J. H. Kim, “Active blood detection in a high resolution capsule endoscopy using color spectrum transformation,” Int’l Conf. on BioMedical Engineering and Informatics, vol. 1, pp. 859–862, May 2008.

[11] A. Karargyris and N. Bourbakis, “Identification of polyps in wireless capsule endoscopy videos using log ga¬bor filters,” in the IEEE Life Science Systems and Applications Workshop, Apr 2009, pp. 143–147.

[12] ——, “Identification of ulcers in wireless capsule endoscopy videos,” in the IEEE Int’l Symposium on Biomedical Imaging: From Nano to Macro, Jun 2009, pp. 554–557.

[13] V. S. Kodogiannis, “Computer-aided diagnosis in clinical endoscopy using neuro-fuzzy systems,” in the IEEE Int’l Conf. on Fuzzy Systems, vol. 3, July 2004, pp. 1425–1429.

[14] V. S. Kodogiannis and M. Boulougoura, “Neural network-based approach for the classification of wireless-capsule endoscopic images,” in Neural Networks, vol. 4, Aug 2005, pp. 2423–2428.

[15] V. S. Kodogiannis and J. N. Lygouras, “A computerized diagnostic decision support system in wireless capsule endoscopy,” in the Int’l Conf. on Intelligent Systems, Sep 2006, pp. 638–644.

[16] M. Kubat and S. Matwin, “Addressing the curse of imbalanced training sets: One sided selection,” in the Int’l Conf. on Machine Learning, 1997, pp. 179–186.

[17] P. Y. Lau and P. L. Correia, “Detection of bleeding patterns in WCE video using multiple features,” in the Int’l Conf. of the IEEE Engineering in Medicine and Biology Society, Aug 2007, pp. 5601–5604.

[18] B. Li and M. Q. H. Meng, “Wireless capsule endoscopy images enhancement using contrast driven forward and backward anisotropic diffusion,” in the IEEE Int’l Conf. on Image Processing, vol. 2, Sep 2007, pp. 437–440.

[19] ——, “Computer aided detection of bleeding in capsule endoscopy images,” in Electrical and Computer Engineering Canadian Conference, May 2008, pp. 1963–1966.

[20] S. Liangpunsakul, L. Mays, and D. K. Rex, “Performance of given suspected blood indicator,” American Gastroenterology, vol. 98, no. 12, pp. 2676–2678, 2003.

[21] C. S. Lima, D. Barbosa, J. Ramos, A. Tavares, L. Monteiro, and L. Carvalho, “Classification of endoscopic capsule images by using color wavelet features, higher order statistics and radial basis functions,” in the Int’l Conf. of the IEEE Engineering in Medicine and Biology Society, Vancouver, Canada, Aug 2008.

[22] B. Liu and M. Q. H. Meng, “Computer-aided detection of bleeding regions for capsule endoscopy images,” IEEE Transactions on Biomedical Engineering, vol. 56, no. 4, pp. 1032–1039, Apr 2009.

[23] J. Liu and X. Yuan, “Obscure bleeding detection in endoscopy images using support vector machines,” Optimization and Engineering, vol. 10, pp. 289–299, 2009.

[24] M. Mackiewicz, J. Berens, M. Fisher, and D. Bell, “Color and texture based gastrointestinal tissue discrimination,” in the IEEE Int’l Conf. on Acoustics, Speech and Signal Processing, vol. 2, May 2006.

[25] B. Manjunath, J.-R. Ohm, V. Vasudevan, and A. Yamada, “Color and texture descriptors,” IEEE Transactions on Circuits and Systems for Video Technology, vol. 11, no. 6, pp. 703–715, Jun 2001.

[26] B. Raskutti and A. Kowalczyk, “Extreme re-balancing for SVMs: a case study,” SIGKDD Explorations Newslet¬ter, vol. 6, pp. 60–69, 2004.

[27] P. Spyridonos, F. Vilarino, J. Vitria, and P. Radeva, “Identification of intestinal motility events of capsule endoscopy video analysis,” Lecture Notes in Computer Science, vol. 3708, pp. 302–311, 2005.

[28] V. Vapnik, The Nature of Statistical Learning Theory. New York: Springer, 1999.

[29] F. Vilarino, L. I. Kuncheva, and P. Radeva, “ROC curves and video analysis optimization in intestinal capsule endoscopy,” Pattern Recognition Letters, vol. 27, pp. 875–881, 2006.