# Defocus-aware Dirichlet particle filter for stable endoscopic video frame recognition

Tsubasa Hirakawa[a,*], Toru Tamaki[a], Bisser Raytchev[a], Kazufumi Kaneda[a], Tetsushi Koide[b], Shigeto Yoshida[c], Yoko Kominami[d], Shinji Tanaka[d]

[a] *Department of Information Engineering, Graduate School of Engineering, Hiroshima University, 1-4-1 Kagamiyama, Higashi-Hiroshima, Hiroshima 739-8527, Japan*
[b] *Research Institute for Nanodevice and Bio Systems (RNBS), Hiroshima University, 1-4-2 Kagamiyama, Higashi-Hiroshima 739-8527 Japan*
[c] *Department of Gastroenterology, Hiroshima General Hospital of West Japan Railway Company, 3-1-36 Futabanosato, Higashiku, Hiroshima 732-0057, Japan*
[d] *Department of Endoscopy, Hiroshima University Hospital,1-2-3 Kasumi, Minami-ku, Hiroshima 734-8551, Japan*

**Abstract**

Background and Objective: A computer-aided system for colorectal endoscopy could provide endoscopists with important helpful diagnostic support during examinations. A straightforward means of providing an objective diagnosis in real time might be for using classifiers to identify individual parts of every endoscopic video frame, but the results could be highly unstable due to out-of-focus frames. To address this problem, we propose a defocus-aware Dirichlet particle filter (D-DPF) that combines a particle filter with a Dirichlet distribution and defocus information.
Methods: We develop a particle filter with a Dirichlet distribution that represents the state transition and likelihood of each video frame. We also incorporate additional defocus information by using isolated pixel ratios to sample from a Rayleigh distribution.
Results: We tested the performance of the proposed method using synthetic and real endoscopic videos with a frame-wise classifier trained on 1,671 images of colorectal endoscopy. Two synthetic videos comprising 600 frames were used for comparisons with a Kalman filter and D-DPF without defocus information, and D-DPF was shown to be more robust against the instability of frame-wise classification results. Computation time was approximately 88 ms/frame, which is sufficient for real-time applications. We applied our method to 33 endoscopic videos and showed that the proposed method can effectively smoothen highly unstable probability curves under actual defocus of the endoscopic videos.
Conclusion: The proposed D-DPF is a useful tool for smoothing unstable results of frame-wise classification of endoscopic videos to support real-time diagnosis during endoscopic examinations.

*Keywords:* Particle filter, Dirichlet distribution, Defocus information, Endoscopy, Colorectal cancer

## 1. Introduction

Colorectal endoscopy (or *colonoscopy*), i.e. endoscopic examination using a narrow-band imaging (NBI) system, is widely used to diagnose colorectal cancer [1]. A computer-aided diagnosis system for colonoscopy would be an extremely helpful tool for supporting diagnosis during examinations. Processing a video stream plays an important role in providing such support because endoscopists typically specify the region of a tumor and capture a video frame to diagnose the tumor's condition. However, intra/inter-observer variability [2–4] shows that diagnosis can be subjective and depends on the endoscopist's experience. Hence, a computer-aided system that provides an objective measure for diagnosis on a screen would be of great assistance [5]. One straightforward means of providing an objective real-time diagnosis might be *frame-wise classification*, i.e., using a machine-learning-based classifier trained off-line with training image patches to recognize a part of every endoscopic video frame and showing classification results (labels or probabilities) on a screen. However, the problem then arises that we do not see when we independently classify training image patches. Figure 1 shows a typical result obtained from a frame-wise classification with three classes. The three curves of posterior probabilities represent the classification results of every frame and are shown to visualize the confidence of the classifier. Although this video sequence continues to capture the same tumor, the classification results are highly unstable, and it would be difficult for endoscopists to understand the output during an examination.

One of the principal causes for this instability is scene blur, or *defocus*, due to the narrow depth of field (see Figure 2). Since operating an endoscope requires expert skill, and the intestinal wall continues moving, it is difficult to

*Corresponding author.
*Email address:* hirakawat@hiroshima-u.ac.jp (Tsubasa Hirakawa)
*URL:* http://home.hiroshima-u.ac.jp/hirakawat/ (Tsubasa Hirakawa)

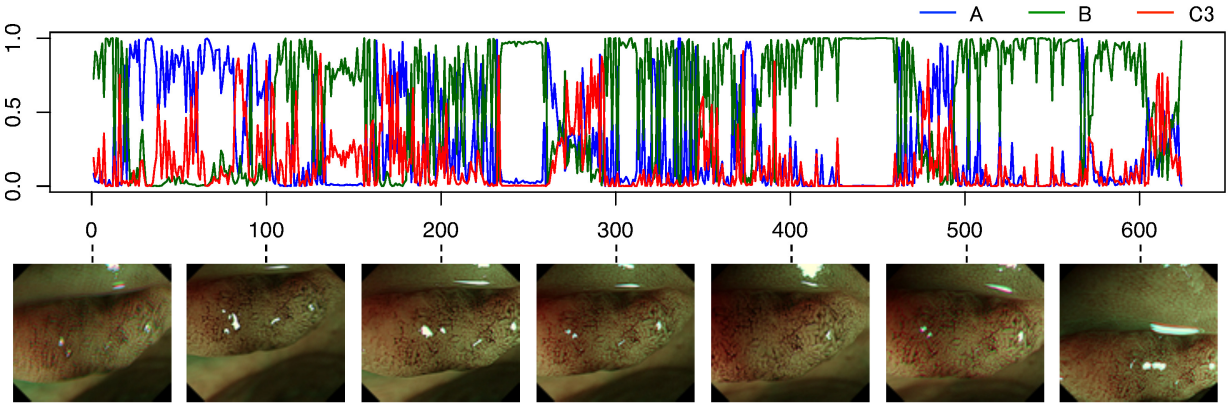



Figure 1: Example of frame-wise classification results from an NBI video [6] with snapshots. For each frame, a patch of size $200 \times 200$ at the center of the frame is classified by a frame-wise classifier to obtain posterior probabilities as a result of a 3-class classification problem. These three classes, type A, B, and C3, correspond to certain diagnostic criteria for a tumor (details are described in Section 2.1). In the upper row, the three curves of posterior probabilities represent the classification results obtained in each frame: the horizontal axis shows frame number and the vertical axis the classification probabilities for the three classes of type A (blue), B (green), and C3 (red). The bottom row shows frames of the video at every 100 frames.

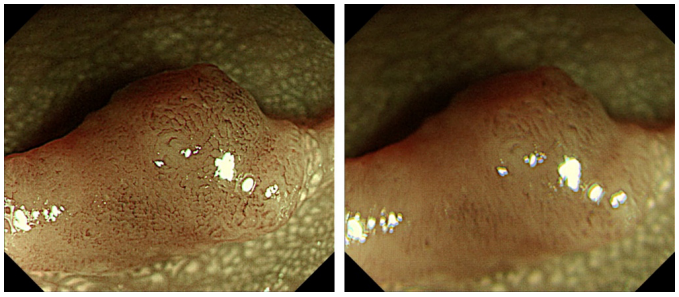


Figure 2: Example of defocus. The tumor is captured in focus in one frame (left), but is defocused in another frame (right).

maintain focus on a tumor for a long time. Features extracted from defocused frames cause unstable results because the classifier has not been trained with such features. Our preliminary experiments also demonstrate that classifying defocused image patches performs worse than classifying well-focused ones (see Section 5.2). Removing defocused frames from a video stream [7] would not be helpful in such an application because results on the screen would frequently stop or disappear.

To overcome this problem, we propose a method for smoothing probability curves, or sequences of posterior probabilities, such as those shown in Figure 1, by incorporating information representing the degree of defocus from each frame in the framework of particle filtering with a Dirichlet distribution [8]. We call the proposed method the *defocus-aware Dirichlet particle filter* (D-DPF). There are two reasons why we need to develop our own method for smoothing probability curves.

First, smoothing techniques use only given signals; therefore, it is difficult to recover from failures in frame-wise classification owing to the defocus of frames. In such a case, it is reasonable to use additional information which represents defocus of each frame; smoothing results tend to follow the observation of the current frame if the frame is in focus, and to keep the results from the previous frames otherwise. Our proposed method uses *isolated pixel ratio* (IPR) (see Section 4.2.1) as defocus information in the likelihood of the particle filter so as to show the confidence of the classification result at each frame.

Second, smoothed results obtained by existing smoothing methods must typically be renormalized at each frame to sum to one, leading to inconsistency between frames as this has no probabilistic significance. Our system outputs confidence values at each frame, i.e., posterior probabilities for the results when classifying a patch in each frame into three classes (type A, B, and C3), on the basis of *NBI magnification findings* [4, 9] (see Section 2.1). Therefore, we developed a probabilistic framework with a particle filter to perform "smoothing of probabilities" using the Dirichlet distribution (see Section 3.2) in such a way that defocus information is incorporated.

The remainder of this paper is organized as follows. Section 2 reviews the relevant medical background regarding colorectal cancer, NBI magnification findings, and related work. We formulate the problem as a particle filter with a Dirichlet distribution in Section 3. Section 4 introduces additional defocus information into the likelihood of the particle filter. Section 5 shows several experimental results using real and synthetic data. We conclude the paper in Section 6.

## 2. Colorectal cancer and related work

### *2.1. Colorectal cancer and colonoscopy*

Colorectal cancer is one of the major causes of cancer related deaths worldwide [10]. Colonoscopy is the most

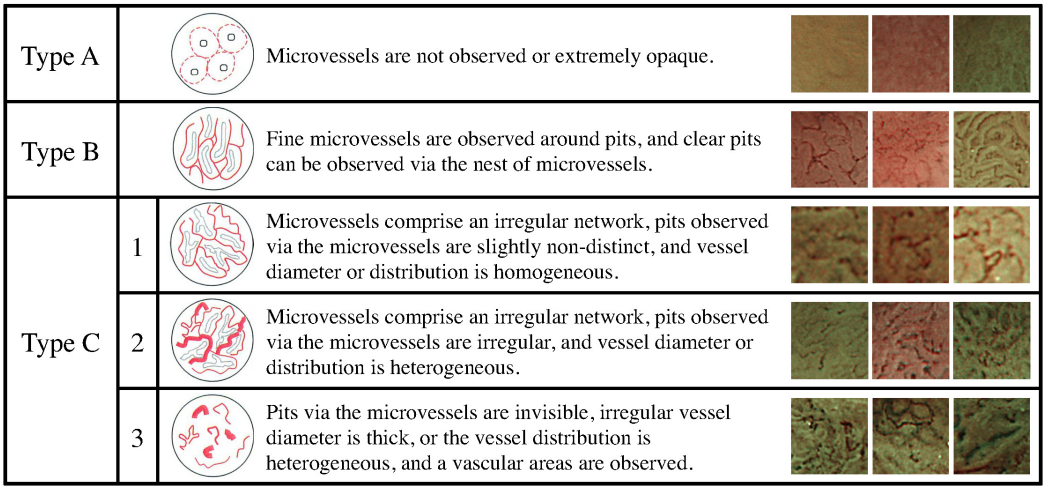



Figure 3: NBI magnification findings [9].

popular and widely used inspection method for such cancer. During colonoscopy, endoscopists observe a tumor displayed on a monitor to determine whether treatment and resection are necessary. Recently, the development of NBI [11–13] has enabled endoscopists to perform examinations in less time. NBI enables the enhancement of microvessel structures by using two light sources of specific wavelengths that are absorbed by hemoglobin in the blood vessels. *NBI magnification findings* have been proposed as a diagnostic criterion by the Hiroshima University Hospital [4, 9], which categorizes appearances of tumors into types A, B, and C, with type C further subclassified into C1, C2, and C3 on the basis of microvessel structures (see Figure 3). Based on the NBI magnification findings, Tamaki et al. [6] proposed a recognition method that classifies NBI images into three types, A, B, and C3.[1]

Tamaki et al. [6] proposed a prototype computer-aided diagnosis system for NBI endoscopy image patches. They used the bag-of-visual words (BoW) framework with a scale invariant feature transform (SIFT), followed by support vector machine (SVM) classifiers. In their experiments, they achieved a recognition rate of 96 % on 908 NBI image patches with 10-fold cross validation. Through the use of this method, a *frame-wise classification* for NBI videoendoscopy, which classifies a part of each endoscopic video frame in a frame-by-frame manner, could be developed. One instance of such frame-wise classification is shown in Fig. 1, which is unstable as mentioned above. In order to make the classification results stable over time, in this paper we propose a method for smoothing sequences of the frame-wise classification results.

[1]The reason to exclude types C1 and C2 is the inherent difficulty to distinguish between subtypes C1, C2, and C3 due to large inter/intra-observer variability [4]. Therefore a poor classification performance is obtained for a five-class classification problem (see [6] for details).

### *2.2. Related work*

Polyp detection has been the most widely performed and studied task in colorectal videoendoscopy in the past two decades. Maroulis et al. [14] proposed a detection system of colorectal lesions in endoscopic videos using neural networks, and Karkanis et al. [15] used color wavelet features. There have also been various other efforts [16–19].

Surprisingly, classification of endoscopic videos has been scarcely investigated. One possible reason might be that a frame-wise classification could be developed by simply applying patch-based classification to video streams frame by frame. In fact, many patch-based classification methods for endoscopic images have been proposed for pit-pattern [20–32] and NBI-endoscopic images [6, 33–35]. Such a simple application of frame-wise classification to video frames was proposed by Manivannan et al. [36]. They classified video frames into normal and abnormal using patch statistics and Gaussian scale-space.

Later, they proposed a *video-specific SVM* (V-SVM), training with video frames to independently classify images or frames [37]. This approach involves the following problems. First, each video frame must have a label assigned by endoscopists, which is a very expensive task. Second, an endoscopic video frame contains many unnecessary parts such as dark background, defocused parts, and highlights. Therefore, using entire frames for learning would lead to a deterioration of classification performance. Third, training a classifier with an entire video is more expensive than training with image patches. Selecting representative image patches is much more efficient when training a classifier or constructing a training dataset. Hence, we employ a more practical strategy–smoothing as post-processing of a frame-wise classification.

Several methods have been proposed to detect and exclude defocused frames from endoscopic videos. Oh et al. [7] attempted to classify video frames into informative and non-informative ones. They proposed two methods, i.e.,

edge- and clustering-based methods. As an edge-based method, they apply a Canny edge detector to each frame and calculate the IPR, the ratio of isolated edge pixels to all edge pixels. They then classify each frame by thresholding the IPR. As a clustering-based method, they extract seven texture features from gray-level co-occurrence matrices of discrete Fourier transform magnitude images and then classify each frame by k-means clustering. To detect indistinct frames, Arnold et al. [38] used the $L^2$-norm of the detail coefficients of a wavelet conversion. Liu et al. [39] proposed robust tracking by detecting and discarding endoscopic video frames that lack features useful to track by using the blurry image detection algorithm proposed by Liu et al. [40].

In our method, we use Oh's IPR because it is inexpensive to compute and incorporate into a particle filter.

## 3. DPF

In this section, we develop DPF, which is a particle filter with a Dirichlet distribution [8].

### *3.1. Particle filters*

Particle filtering [41–44] is a method for estimating the internal states of a state-space model sequentially. In particle filters, a probability distribution is represented by Monte Carlo approximation, i.e., by a set of $K$ random samples, or *particles*, to handle non-linear and non-Gaussian problems. A particle filter estimates posterior probabilities of the internal states conditioned on observations through the following two steps. A *prediction step* estimates the prior probabilities at time $t$ from observations before time $t-1$ and a state transition. An *update step* estimates the posterior probabilities at time $t$ from the prior probabilities and a likelihood. To avoid confusion of terms, we hereafter refer to *classification probability* as the discrete (posterior) probability obtained from a frame-wise classification, and to *posterior probability* as the smoothed probability obtained by the particle filter.

Observation $\boldsymbol{y}_t$ is the classification probability obtained at time (or frame) $t$ and has a unit $L_1$ norm: $\|\boldsymbol{y}_t\|_1 = 1$. Let $\boldsymbol{y}_{1:t} = \{\boldsymbol{y}_1, \ldots, \boldsymbol{y}_t\}$ be a series of observations obtained prior to time $t$. The internal state $\boldsymbol{x}_t$ is the posterior probability that should be estimated at time t by smoothing. Hence $\|\boldsymbol{x}_t\|_1 = 1$, and we use the notation $\boldsymbol{x}_{1:t} = \{\boldsymbol{x}_1, \ldots, \boldsymbol{x}_t\}$ in the same manner as that for $\boldsymbol{y}_{1:t}$.

A prediction step estimates the prior probability $p(\boldsymbol{x}_t \,|\, \boldsymbol{y}_{1:t-1})$ using the following integral:

$$p\left(\boldsymbol{x}_t \,|\, \boldsymbol{y}_{1:t-1}\right) = \int p\left(\boldsymbol{x}_t \,|\, \boldsymbol{x}_{t-1}\right) p\left(\boldsymbol{x}_{t-1} \,|\, \boldsymbol{y}_{1:t-1}\right) d\boldsymbol{x}_{t-1}, \quad (1)$$

where $p\left(\boldsymbol{x}_t \,|\, \boldsymbol{x}_{t-1}\right)$ is the *state transition probability* between states at times $t-1$ and $t$. Once $p(\boldsymbol{x}_t \,|\, \boldsymbol{y}_{1:t-1})$ is computed, an update step computes the posterior probability $p(\boldsymbol{x}_t \,|\, \boldsymbol{y}_{1:t})$ as follows:

$$p\left(\boldsymbol{x}_t \,|\, \boldsymbol{y}_{1:t}\right) \propto p\left(\boldsymbol{y}_t \,|\, \boldsymbol{x}_t\right) p\left(\boldsymbol{x}_t \,|\, \boldsymbol{y}_{1:t-1}\right), \quad (2)$$

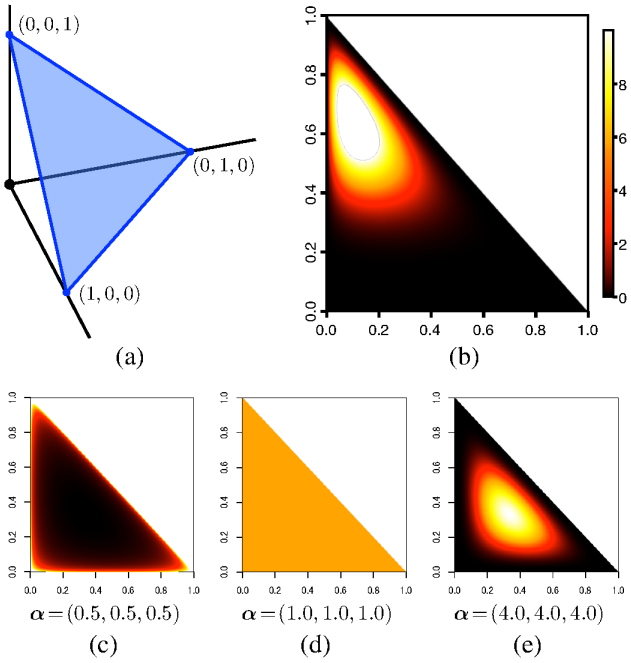



Figure 4: Examples of 3D Dirichlet distributions. (a) Support of the Dirichlet distribution. (b) Probability density of a Dirichlet distribution (darker the pixels, lower the density). (c, d, and e) Typical probability density shapes for different parameters $\boldsymbol{\alpha}$.

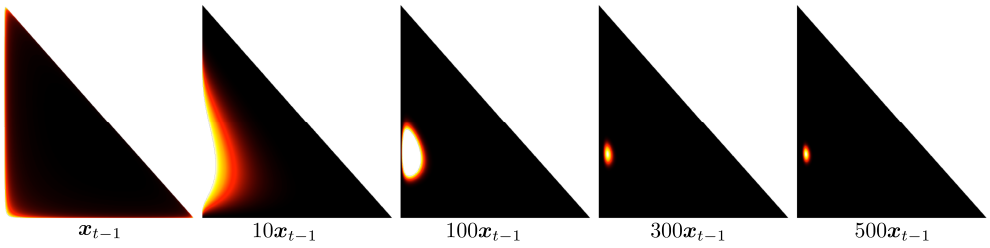
where $p\left(\boldsymbol{y}_t \,|\, \boldsymbol{x}_t\right)$ is the *likelihood*. Repeating the prediction and update steps from time zero yields a sequence of estimates of $\boldsymbol{x}_t$.

### *3.2. Dirichlet distribution*

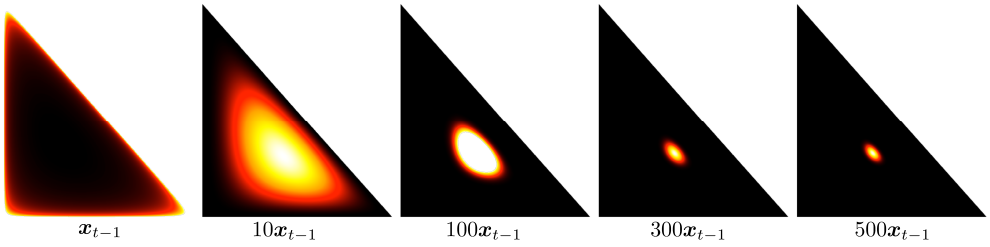
To represent probability distributions of $\boldsymbol{x}_t$, we propose to use the $N$-dimensional Dirichlet distribution [45] with parameter $\boldsymbol{\alpha} = (\alpha_1, \ldots, \alpha_N), \alpha_i > 0$ defined by

$$\mathrm{Dir}_{\boldsymbol{x}}[\boldsymbol{\alpha}] = \frac{\Gamma(\sum_{i=1}^{N} \alpha_i)}{\prod_{i=1}^{N} \Gamma(\alpha_i)} \prod_{i=1}^{N} x_i^{\alpha_i - 1}, \quad (3)$$

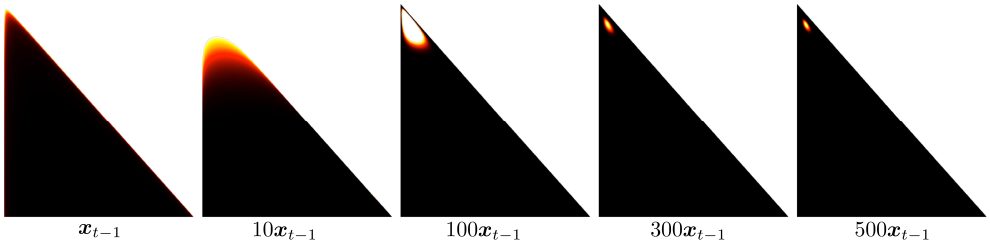
where $\Gamma$ is the gamma function, and $\boldsymbol{x} = (x_1, \ldots, x_N)$ is a random variable with $\|\boldsymbol{x}\|_1 = 1$. The parameter $\boldsymbol{\alpha}$ controls the shape of the distribution. Figure 4 shows examples of three-dimensional (3D) Dirichlet distributions. When $\alpha_i < 1$ for all $i$, the density has greater probabilities around the vertices, as shown in Figure 4(c). When $\alpha_i = 1$ for all $i$, the density flattens. Otherwise, the density has a peak at the mode $\frac{\boldsymbol{\alpha}-\mathbf{1}}{\|\boldsymbol{\alpha}\|_1 - N}$, where $\mathbf{1}$ is a vector of ones, as shown in Figure 4(e). Additionally, the larger the parameter values are, the steeper the peak becomes. In our three-class classification problem ($N = 3$), the support of the probability density is a two-dimensional triangle in a 3D space. Using the Dirichlet distribution enables us to formulate the state transition and the likelihood of the model.




(a) $\boldsymbol{x}_{t-1} = (0.050, 0.300, 0.650)$




(b) $\boldsymbol{x}_{t-1} = (0.400, 0.300, 0.300)$




(c) $\boldsymbol{x}_{t-1} = (0.050, 0.900, 0.050)$

Figure 5: Examples of state transition probabilities modeled by a Dirichlet distribution.

### *3.3. State transition*

A prediction step models the relationship between internal states at times $t-1$ and $t$. To define a state transition probability, we remember that the internal state of our problem is a posterior probability $\boldsymbol{x}$ that satisfies $\|\boldsymbol{x}\|_1 = 1$. The support of $\boldsymbol{x}$ is exactly the same as that of the Dirichlet distribution; thus, it can be used to define the state transition. Intuitively, internal state $\boldsymbol{x}_t$ has a large probability if it is similar to $\boldsymbol{x}_{t-1}$. Therefore, it is natural to use the probability density $p(\boldsymbol{x}_t \mid \boldsymbol{x}_{t-1})$ having a peak around $\boldsymbol{x}_{t-1}$. We propose to define the state transition probability with a Dirichlet distribution as follows:

$$p\left(\boldsymbol{x}_t \mid \boldsymbol{x}_{t-1}\right) = \mathrm{Dir}_{\boldsymbol{x}_t}[\boldsymbol{\alpha}\left(\boldsymbol{x}_{t-1}\right)], \tag{4}$$

where parameter $\boldsymbol{\alpha}$ is now a function of $\boldsymbol{x}_{t-1}$. To constrain the density (4) to have a peak around $\boldsymbol{x}_{t-1}$, we assume linearity between $\boldsymbol{\alpha}$ and $\boldsymbol{x}_{t-1}$:

$$\boldsymbol{\alpha} = A\boldsymbol{x}_{t-1} + \boldsymbol{b}, \tag{5}$$

where $A$ is an $N \times N$ matrix and $\boldsymbol{b}$ is an $N$-vector. Throughout the remainder of this paper, we further simplify the linear function as $A = aI$ and $\boldsymbol{b} = b\boldsymbol{1}$, and use the following simplified notation:

$$\boldsymbol{\alpha}\left(\boldsymbol{x}_{t-1}, a, b\right) = a\boldsymbol{x}_{t-1} + b\boldsymbol{1}. \tag{6}$$

Now, we reformulate Eq. (4) as follows:

$$p\left(\boldsymbol{x}_t \mid \boldsymbol{x}_{t-1}, \theta\right) = \mathrm{Dir}_{\boldsymbol{x}_t}[\boldsymbol{\alpha}(\boldsymbol{x}_{t-1}, \theta, 0)]. \tag{7}$$

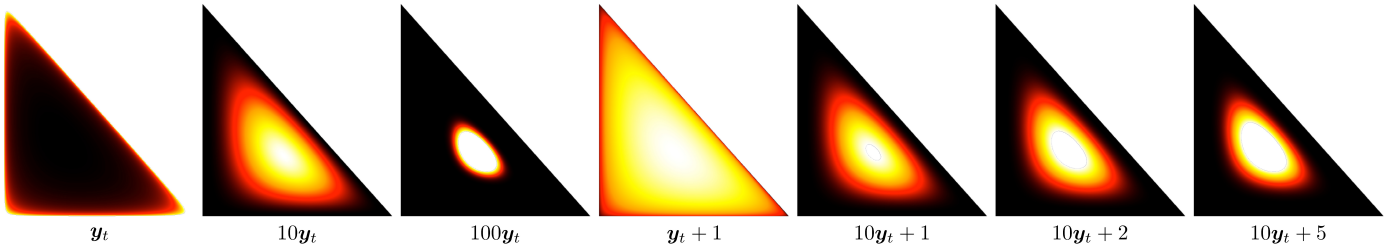
Here, we set $b = 0$ to make the mean of the density coincide with $\boldsymbol{x}_{t-1}$:

$$E[\boldsymbol{x}_t] = \frac{\boldsymbol{\alpha}}{\|\boldsymbol{\alpha}\|_1} = \frac{a\boldsymbol{x}_{t-1}}{\|a\boldsymbol{x}_{t-1}\|_1} = \frac{a\boldsymbol{x}_{t-1}}{a\|\boldsymbol{x}_{t-1}\|_1} = \boldsymbol{x}_{t-1}, \tag{8}$$

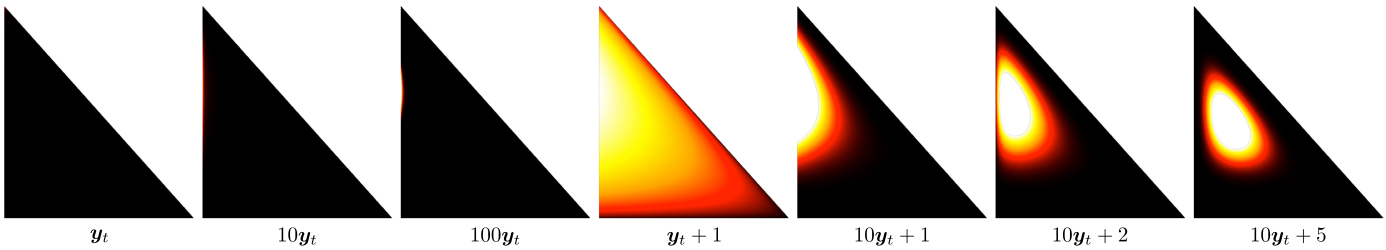
where we use the scale-invariant property of the L1-norm for the third equation and $\|\boldsymbol{x}_{t-1}\|_1 = 1$ for the last equation. Figure 5 shows examples of the state transition probability density function of a 3D Dirichlet distribution. According to our observations when changing the range of the parameter $\theta$, 100 or greater is a typical choice for the value of $\theta$.

### *3.4. Likelihood*

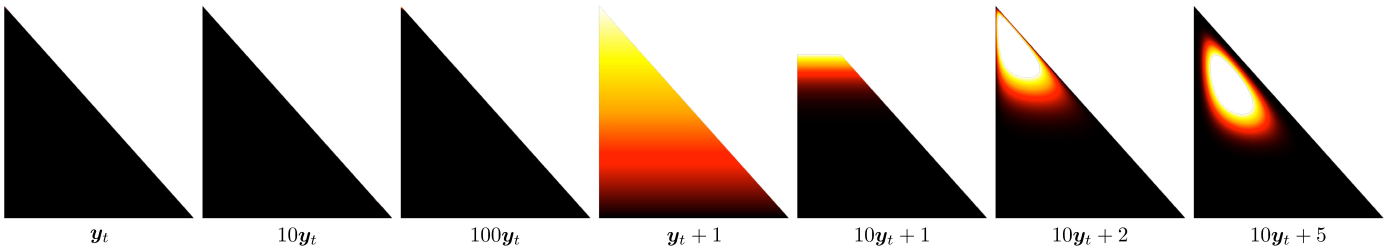
In an update step, particles representing prior distribution $p(\boldsymbol{x}_t \mid \boldsymbol{y}_{1:t-1})$ are weighted by a likelihood, and the posterior probability $p(\boldsymbol{x}_t \mid \boldsymbol{y}_{1:t})$ is then estimated on the basis of the weighted particles. In our problem, we assume that the likelihood has a peak at $\boldsymbol{y}_t$ and propose to define the likelihood by using a Dirichlet distribution as follows:




(a) $\boldsymbol{y}_t = (0.4000, 0.3000, 0.3000)$




(b) $\boldsymbol{y}_t = (0.0001, 0.6000, 0.3999)$




(c) $\boldsymbol{y}_t = (0.0001, 0.9998, 0.0001)$

Figure 6: Examples of likelihood functions modeled by Dirichlet distribution.

$$p\left(\boldsymbol{y}_t \,|\, \boldsymbol{x}_t, \gamma\right) = \mathrm{Dir}_{\boldsymbol{x}_t}[\boldsymbol{\alpha}\left(\boldsymbol{y}_t, \gamma, b\right)]. \tag{9}$$

Here, we use the probability distribution of $\boldsymbol{x}_t$ as the likelihood of $\boldsymbol{y}_t$. We wish to make the likelihood have a broad peak at $\boldsymbol{y}_t$ because the smoothing effect might decrease if the peak is sufficiently steep such that only particles near $\boldsymbol{y}_t$ have extremely large weights. To this end, we set $b = 1$ instead of $b = 0$, because the zero bias leads to a steep peak when the value of $\boldsymbol{y}_t$ is extremely close to zero. Figure 6 shows examples with and without the bias term (i.e., $b = 0$ or 1). When some values in $\boldsymbol{y}_t$ are negligible as in Figures 6(b) and (c), the likelihood with $b = 0$ has a steep peak at the edge of the triangular support. In contrast, the likelihood with $b = 1$ has a reasonably broad peak inside the triangle. Based on our observations, typical values of $\gamma$ and $b$ should be 10 (or less) and on 1 (or greater), respectively. Particularly, setting $b = 1$ makes the likelihood have a peak at exactly $\boldsymbol{y}_t$ because:

$$\begin{aligned}
\mathrm{mode}[\boldsymbol{x}_t] &= \frac{\boldsymbol{\alpha} - \mathbf{1}}{\|\boldsymbol{\alpha}\|_1 - N} \\
&= \frac{(\gamma \boldsymbol{y}_t + \mathbf{1}) - \mathbf{1}}{\|\gamma \boldsymbol{y}_t + \mathbf{1}\|_1 - N} \\
&= \frac{\gamma \boldsymbol{y}_t}{\gamma \|\boldsymbol{y}_t\|_1 + \|\mathbf{1}\|_1 - N} \\
&= \boldsymbol{y}_t,
\end{aligned} \tag{10}$$

where $\|\boldsymbol{y}_t\|_1 = 1$ and $\|\mathbf{1}\|_1 = N$.

## 4. D-DPF

Here, we incorporate additional defocus information at each frame to develop D-DPF. The DPF discussed in the previous section assumes that each observation $\boldsymbol{y}_t$ is generated by the true state $\boldsymbol{x}_t$ with a Dirichlet distribution with parameter $\gamma$. The graphical model of DPF is shown in Figure 7(a), with factor nodes (black squares) representing potential functions of $\boldsymbol{y}_t, \boldsymbol{x}_t$, and the deterministic parameter $\gamma$.

We assume that observation $\boldsymbol{y}_t$ is influenced by the true state $\boldsymbol{x}_t$ as well as a temporal hidden variable $\gamma_t$, which

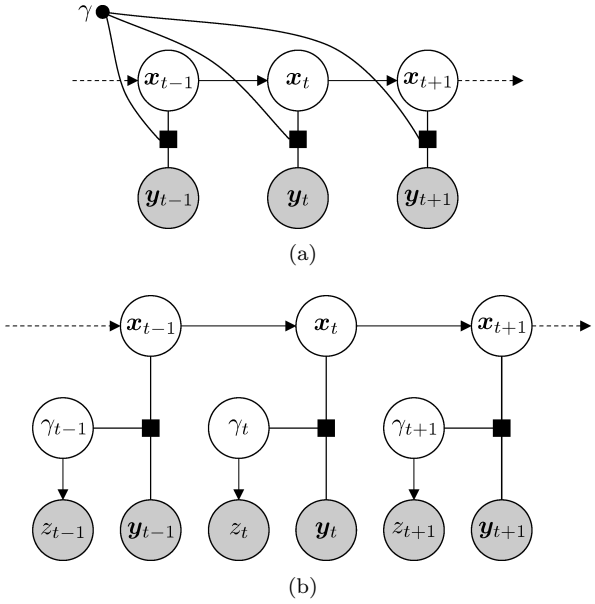



Figure 7: Graphical models of (a) DPF and (b) D-DPF.

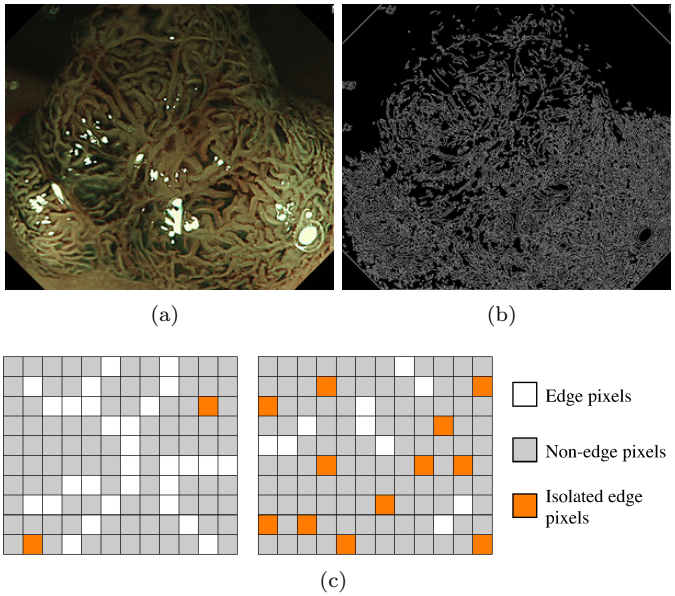



Figure 8: Concept of isolated pixel proposed by Oh et al. [7]. (a) Example of endoscopic image and (b) edges extracted by Canny edge detector. (c) Edges of focused (left) and defocused (right) frames.

is inferred from the additional defocus information. The graphical model of D-DPF in this section is shown in Figure 7(b). The factor nodes now represent $\boldsymbol{y}_t, \boldsymbol{x}_t$, and $\gamma_t$. We further assume that the hidden variable is generated by the defocus information $z_t$.

### *4.1. Update step*

We begin with a modified definition of the update step as follows:

$$p\left(\boldsymbol{x}_t \,|\, \boldsymbol{y}_{1:t}, \gamma_{1:t}, z_{1:t}\right) \propto p\left(\boldsymbol{y}_t, \gamma_t, z_t \,|\, \boldsymbol{x}_t\right) p\left(\boldsymbol{x}_t \,|\, \boldsymbol{y}_{1:t-1}, \gamma_{1:t-1}, z_{1:t-1}\right), \tag{11}$$

where $z_t$ is a scalar value representing the defocus information at time $t$.

We model the likelihood $p(\boldsymbol{y}_t, \gamma_t, z_t \,|\, \boldsymbol{x}_t)$ with a Dirichlet distribution with a peak at $\boldsymbol{y}_t$, whose broadness depends on $z_t$. According to the graphical model in Fig. 7(b), we propose to redefine the likelihood as follows:

$$\begin{aligned} p\left(\boldsymbol{y}_t, \gamma_t, z_t \mid \boldsymbol{x}_t\right) &= p\left(\boldsymbol{y}_t, \gamma_t \mid \boldsymbol{x}_t\right) p\left(z_t \mid \boldsymbol{x}_t, \boldsymbol{y}_t, \gamma_t\right) \\ &= p\left(\boldsymbol{y}_t, \gamma_t \mid \boldsymbol{x}_t\right) p\left(z_t \mid \gamma_t\right). \end{aligned} \tag{12}$$

Here, we use the fact that $z_t$ is conditionally independent of $\boldsymbol{y}_t$ and $\boldsymbol{x}_t$, given $\gamma_t$ based on the graphical model. The potential function $p(\boldsymbol{y}_t, \gamma_t \mid \boldsymbol{x}_t)$, corresponding to the factor node in the graphical model, has a form similar to Eq. (9); hence, we define it as follows:

$$p\left(\boldsymbol{y}_t, \gamma_t \mid \boldsymbol{x}_t\right) = \mathrm{Dir}_{\boldsymbol{x}_t}[\boldsymbol{\alpha}\left(\boldsymbol{y}_t, \gamma_t, 1\right)]. \tag{13}$$

### *4.2. Hidden variable $\gamma_t$*

We model $p(z_t \mid \gamma_t)$, the relation between the hidden variable $\gamma_t$, and the defocus information $z_t$. We wish to reduce the effect of classification failures using a frame-wise classifier at defocused frames. In that case, observation $\boldsymbol{y}_t$ is less reliable, and the likelihood is expected to have a broad peak with the result that particles far from the peak at $\boldsymbol{y}_t$ are assigned larger weights. Therefore, we control $\gamma_t$ to be smaller at defocused frames to have a broad likelihood. Herein, we use IPR as $z_t$, and model $p(z_t \mid \gamma_t)$ with the Rayleigh distribution.

#### *4.2.1. IPR*

*Isolated pixels* are edge pixels extracted by a Canny edge detector, whose eight-neighbors are not edge pixels, as shown in Figure 8. IPR is the ratio of the isolated pixels to all edge pixels and takes values in the range between 0 and 1. We observe connected edge pixels in a sharp and focused image, whereas many isolated pixels are observed in defocused frames. In other words, a focused frame has a lower IPR value, and a defocused frame has a higher IPR value. IPR can be used to classify frames as informative or non-informative.

In Oh et al.'s paper, IPR values are distributed in the range between 0 and 0.1. However, observations can differ in different endoscopic videos due to frame size, zooming, and optical magnification, or when different types of endoscopes are used. To estimate the distribution of IPR in our endoscopic videos, we computed a histogram of IPR extracted from 33 videos (see subsection 5.1 for details), as shown in Figure 9. We can see that IPR is distributed between 0 and 0.01. Using the IPR as $z_t$, we propose the model described in the next section.

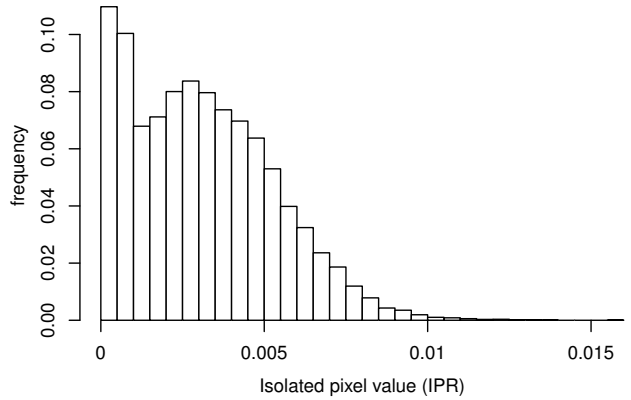



Figure 9: Histogram of IPRs computed from endoscopic videos.

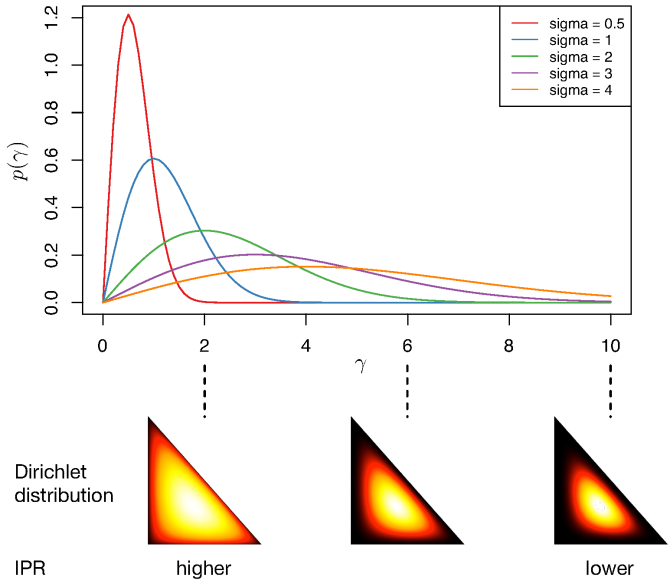



Figure 10: Relationship between IPR, Rayleigh distribution, and Dirichlet distribution. Top: examples of the probability density function of the Rayleigh distribution. Bottom: examples of Dirichlet distributions corresponding to different values in the Rayleigh distributions.

### *4.2.2. Rayleigh distribution*

We use the Rayleigh distribution [46] to represent the relationship between the hidden variable $\gamma_t$ and defocus information $z_t$. The Rayleigh distribution is defined by

$$\mathrm{Ray}_\gamma\left[\sigma\right] = \frac{\gamma}{\sigma^2}\exp\left(-\frac{\gamma^2}{2\sigma^2}\right), \tag{14}$$

where $\sigma > 0$ is a parameter. The top row of Figure 10 shows a few examples of the probability density function of the Rayleigh distribution. Smaller values of $\sigma$ cause the distribution to peak toward zero, whereas larger values of $\sigma$ broaden it.

As discussed above, we wish to have a broad peak of the Dirichlet distribution as a likelihood when the frame is defocused, and a lower value of $\gamma_t$ is preferred in that case. In terms of IPR, a defocused frame contains many isolated pixels, resulting in larger IPR values. In summary, at a defocused frame, IPR or $z_t$ is larger, and smaller values of $\gamma_t$ must be sampled during the sampling procedure of the particle filter, resulting in a broad peak of the likelihood. Consequently, we propose to use $z_t$ for controlling the parameter $\sigma$ of the Rayleigh distribution as follows:

$$p\left(z_t \,|\, \gamma_t\right) = \mathrm{Ray}_{\gamma_t}\left[\sigma\left(z_t\right)\right], \tag{15}$$

where $\sigma(z_t)$ is now a function of $z_t$. To achieve the desired behavior, we use a function of the form

$$\sigma\left(z_t\right) = a\exp\left(bz_t\right), \tag{16}$$

where $a$ and $b$ are parameters to be tuned. A plot of this function is shown in Figure 11. The reason for using exponential decay is the range of $z_t$. If a frame is in focus, then the IPR might be zero or some small positive value. However, it can be extremely large (as much as one) for a defocused frame. Therefore, we assume that the range of $z_t$ is $[0, 0.01]$, as mentioned above, but also allow larger values if they have little effect. The use of exponential decay allows larger values beyond the range above, but they would be effectively squeezed into an extremely narrow range on the vertical axis, as shown in Figure 11.

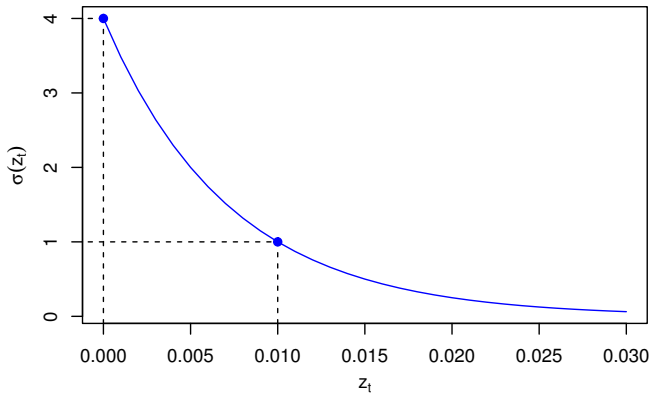



Figure 11: Proposed scaling function of $\sigma(z_t)$.

As a reasonable range for the vertical axis, $\sigma$, we choose from 1 to 4 for $\sigma$, which is in accordance with observations of typical values of $\gamma_t$. At the end of the previous section, we mentioned that we prefer $\gamma_t$ to take values of 10 or less. When we observe the horizontal axis $\gamma$ of Figure 10, the Rayleigh distribution with $\sigma = 4$ has support that almost covers the range $[0, 10]$. Therefore, in the current work we use the fixed (but flexible) range $[0, 0.01]$ for $z_t$, $[1, 4]$ for $\sigma$, and $[0, 10]$ for $\gamma_t$. To this end, we solved the following system of equations

$$\begin{cases} 4 = a\exp\left(b\cdot 0\right) \\ 1 = a\exp\left(b\cdot 0.01\right), \end{cases} \tag{17}$$

to obtain $a = 4$ and $b = \frac{1}{0.01}\ln\left(\frac{1}{4}\right) = -\frac{\ln 4}{0.01}$.

**Algorithm 1** Defocus-aware Dirichlet particle filter (D-DPF).

1: Sample $K$ particles $\{\boldsymbol{s}_{0|0}^{(i)}\}_{i=1}^{K}$ from $p(\boldsymbol{x}_0)$.
2: **for** time $t = 1 \dots T$ **do**
3:  **for** $i = 1 \dots K$ **do**
4:  Draw a sample $\boldsymbol{s}_{t|t-1}^{(i)} \sim \mathrm{Dir}_{\boldsymbol{x}_t}[\boldsymbol{\alpha}(\boldsymbol{s}_{t-1|t-1}^{(i)}, \theta, 0)]$.
5:  **end for**
6:  Compute $z_t$ from video frame at time $t$.
7:  Compute $\gamma_t \sim \mathrm{Ray}_{\gamma_t}[z_t]$.
8:  **for** $i = 1 \dots K$ **do**
9:  Compute a weight $\pi_t^{(i)} = \mathrm{Dir}_{x_t = s_{t|t-1}^{(i)}}[\boldsymbol{\alpha}(\boldsymbol{y}_t, \gamma_t, 1)]$.
10:  **end for**
11:  Sample $K$ times as $\{\boldsymbol{s}_{t|t}^{(i)}\}_{i=1}^{K}$ from $\{\boldsymbol{s}_{t|t-1}^{(i)}\}_{i=1}^{K}$ with replacement according to the weights $\pi_t^{(i)}$.
12:  Estimate $\boldsymbol{\alpha}$ from $\{\boldsymbol{s}_{t|t}^{(i)}\}_{i=1}^{K}$.
13:  Compute the mode $\hat{\boldsymbol{x}}_t$ of the Dirichlet distribution from $\boldsymbol{\alpha}$.
14: **end for**

### *4.3. Prediction step*

We use the same state transition as that discussed in Section 3 and define the prediction step as

$$p\left(\boldsymbol{x}_y \,|\, \boldsymbol{y}_{1:t-1}, \gamma_{1:t-1}, z_{1:t-1}\right) = \int p\left(\boldsymbol{x}_t \,|\, \boldsymbol{x}_{t-1}\right) p\left(\boldsymbol{x}_{t-1} \,|\, \boldsymbol{y}_{1:t-1}, \gamma_{1:t-1}, z_{1:t-1}\right) d\boldsymbol{x}_{t-1}, \tag{18}$$

where

$$p(\boldsymbol{x}_t \,|\, \boldsymbol{x}_{t-1}) = \mathrm{Dir}_{\boldsymbol{x}_t}[\boldsymbol{\alpha}\left(\boldsymbol{x}_{t-1}, \theta, 0\right)]. \tag{19}$$

### *4.4. Algorithm*

Algorithm 1 details the proposed D-DPF. At each time step $t$, the mode $\hat{\boldsymbol{x}}_t$ of the Dirichlet distribution is obtained to visualize a plot along with the input observation $\boldsymbol{y}_t$ in the experiments:

1. Sample $K$ particles according to an initial Dirichlet distribution $p(\boldsymbol{x}_0)$ with $\boldsymbol{\alpha} = (0.4, 0.3, 0.3)$ (line 1).
2. Sample prediction particles by performing a transition of particles at time $t-1$ according to the state transition probability (lines 2 to 5).
3. Compute the defocus information $z_t$ and sample $\gamma_t$ according to $z_t$ (lines 6 and 7).
4. Estimate weights $\pi_t^{(i)}$ for prediction of particles (lines 8 to 10).
5. Sample with replacement for $\boldsymbol{s}_{t|t}^{(n)}$ to be proportional to weight $\pi_t^{(i)}$ (line 11). Subsequently, compute the maximum likelihood estimate of $\boldsymbol{\alpha}$ of the Dirichlet distribution [47] from particles $\boldsymbol{s}_{t|t}^{(n)}$ (line 12). Then, each component of the mode $\hat{\boldsymbol{x}}_t$ of the Dirichlet distribution is separately computed using

$$x_i = \frac{\alpha_i - 1}{\sum_{i=1}^{N} \alpha_i - N}, \alpha_i > 1. \tag{20}$$

In the event that $\alpha_i$ is less than one, we set $\alpha_i = 0$ (line 13).

6. Return to step 2.

## 5. Experimental results

This section demonstrates the effectiveness of the proposed D-DPF smoothing. The following subsections describe the dataset of image patches and videos, and the classification results for blurred image patches and endoscopic video sequences.

### *5.1. Dataset and frame-wise classification*

We used a dataset of 1,671 NBI image patches of different sizes (type A: 504, type B: 847, type C3: 320) to train an SVM classifier for frame-wise classification. Each of the image patches was trimmed from an endoscopic video frame and labeled by endoscopists. These endoscopic video frames were captured and collected during endoscopic examinations, and each frame has the same label as the corresponding image patch. Details about the dataset, features used, and classification can be found in [6].

For evaluation, we have 33 NBI-endoscopic videos (type A: 5, type B: 27, type C3: 1) whose frame rate is 30 fps and size is full HD ($1,980 \times 1,080$ pixels), wherein the window size displaying the endoscopic video is $1,000 \times 870$ pixels. Each endoscopic video shows a single tumor, but there are many defocused frames. For each video frame, a $200 \times 200$ patch at the center of the window is classified by a pre-trained frame-wise classifier to obtain classification probabilities $\boldsymbol{y}_t$. Frame lengths of the videos range between 200 and 2,500, with more than 20,000 frames in total.

Labeling each video frame of these videos is, therefore, very expensive, as stated previously. Labeling still images in the aforementioned datasets of 1,671 images was possible because it took several years to collect that number of NBI images for various patients. In subsection 5.3, we instead use synthetic endoscopic video sequences, wherein frame labels are known, for analysis and evaluation of the proposed method. In subsection 5.4, we show some of the results for real videos to demonstrate the behavior of the proposed method.

The training NBI images have been used for clinical reports, while the endoscopic videos were collected for our experiments. All of these endoscopic images and videos were collected at the Hiroshima University Hospital, following the guidelines of the Hiroshima University ethics committee, and informed consent has been obtained from the patients and their families.

### *5.2. Classification results for blurred patches*

Before showing the results for the D-DPF, we demonstrate the performance deterioration of blurred image patch classification when using the classification method

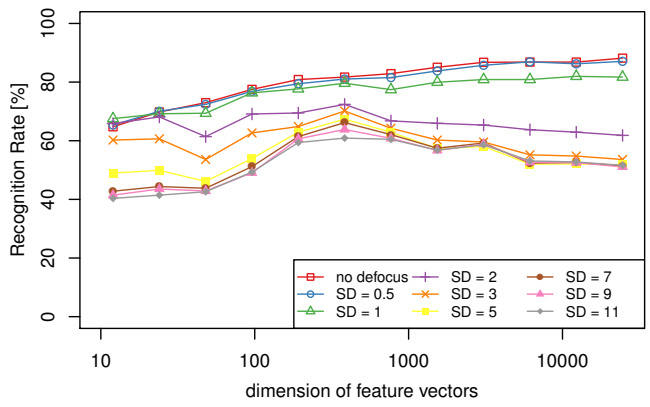



Figure 12: Classification performance on the NBI image patches with and without Gaussian blur. SD stands for $\sigma_{blur}$. The horizontal axis shows the dimension of the feature vectors (the number of visual words, see [6] for details).

proposed by Tamaki et al. [6]. For training, 160 NBI image patches for each class, 480 NBI image patches in total, were randomly selected from the 1,671 image patch dataset. The remainder of the dataset was used for evaluation by adding Gaussian blur with standard deviation $\sigma_{blur} = 0.5, 1, 2, 3, 5, 7, 9, 11$. The classification results are shown in Figure 12 for different dimensions of the feature vectors, as this is an important parameter for obtaining better classification performance. The performance on image patches without Gaussian blur is better than that with blur. When $\sigma_{blur} > 3$, the performance drops to approximately 50%. As shown in Figure 12, even small blur of $\sigma_{blur} = 2$ or 3 affects classification performance.

### *5.3. Results for synthetic video sequences*

Hereafter, we evaluate the performance of the proposed smoothing method. Therefore, in this subsection, we assess the performance on synthetic endoscopic video sequences.

We created the synthetic videos as follows. First, we selected three images from the dataset of 1,671 NBI images. These were not trimmed patches, but original video frames from which the patches were trimmed. Next, each of these images was repeated 200 times to create a 200-frame static video, resulting in a synthetic video of 800 frames corresponding to four different static scenes. Gaussian blur with $\sigma_{blur} = 5$ was then added into some parts of the videos. Next, we added noise in either of two ways. One was to add Gaussian noise with standard deviation $\sigma_{noise}$ to every frame, with classification probabilities then obtained using frame-wise classification. The other was to sample Dirichlet noise according to classification probabilities using

$$\mathrm{Dir}_{\boldsymbol{x}_t}[\boldsymbol{\alpha}(\hat{\boldsymbol{y}}_t, s, 1)], \tag{21}$$

where $\hat{\boldsymbol{y}}_t$ is an observation at an individual video frame and $s$ is a scale parameter. The sampled Dirichlet noise was used as an observation vector for each frame.

For training, we randomly selected 300 NBI image patches for each class, 900 NBI image patches in total, from the 1,671 NBI image patch dataset. Note that image patches corresponding to images used to create the synthetic video sequences were not used for training.

We should note that the conclusions obtained from the experimental results in this section are limited to observing how fast our method responds to the transitions between blurring and non-blurring frames. This is because the synthetic static videos with blur do not contain any other problematic issues such as abrupt motion or light condition changes. Results for real endoscopic videos are shown in the next section.

#### *5.3.1. Comparison of DPF, D-DPF and a Kalman filter*

First, we evaluate the difference between DPF from Section 3 and D-DPF from Section 4.

Figure 13 shows results for a synthetic video to which Gaussian noise has been added. Figure 13(a) shows the classification probabilities for each original (noise-free) frame. For this synthetic video, four NBI images were used, each of which lasts 200 frames. We can see three discontinuities at frames 200, 400, and 600. Gaussian blur of $\sigma_{blur} = 5$ is applied to 10 frames before and after the 100, 300, 500, and 700th frame (that is, between frames 90 and 110, and so on) as indicated by shading in Figure 13(b) to (e). Then, Gaussian noise with $\sigma_{noise} = 1$ was added to all frames to create a final synthetic video for processing. Classification probabilities for this video are shown in Figure 13(b).

As observed in Figure 13(b), between frames 200 and 400, the classification probabilities are highly unstable, and for the shaded frames (where blur is applied), the classification probability curves abruptly change. Figure 13(c) shows the IPR of each frame, and the values of IPR for the shaded (blurred) frames increase as expected. Results for DPF and D-DPF are shown in Figures 13(d) and (e), respectively. At approximately frames 100 and 500, where blur is applied, DPF is affected by a sudden change in classification results. In contrast, D-DPF is rather robust to the change due to the defocus information extracted from each frame.

Figure 13(f) shows the smoothing result obtained by a Kalman filter. Parameters were manually tuned; hence, the results for the Kalman filter look similar to those for DPF and D-DPF because an optimization with an EM algorithm could not find suitable parameters that would produce satisfactory smoothing results in this case. The Kalman filter was also affected by the sudden change of classification results at frames where blur was applied. Another defect of the Kalman filter was overshooting. Around frames 150, 450, and 650, results exceed the range between 0 and 1. Normalizing or clipping the results in the range of zero to one at each frame would lead to inconsistency with the results for the successive frames, and

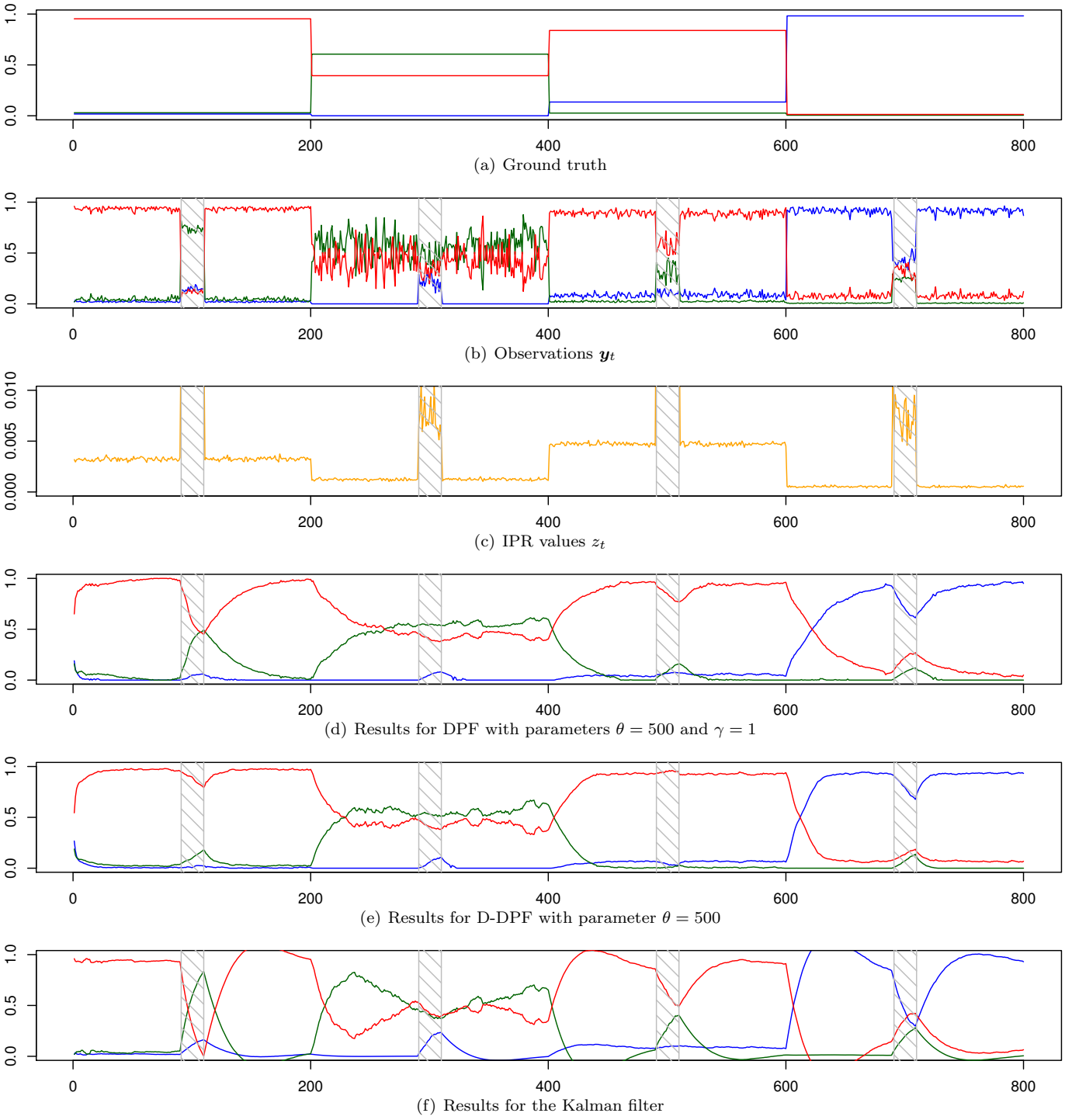



Figure 13: Smoothing results on a synthetic video with Gaussian noise of standard deviation $\sigma_{noise} = 1$. The horizontal axis shows frame number. The vertical axis is classification probabilities for the three classes of type A (blue), B (green), and C3 (red) (except (c)). From top to bottom, ground truth of classification probabilities, observations with no smoothing, IPR values, smoothing results for DPF, D-DPF with $K = 1,000$ particles, and Kalman filter. Shaded frames are blurred by Gaussian with $\sigma_{blur} = 5$.

the probabilistic framework would be lost.

Figure 14 shows the results obtained when Dirichlet noise with $s = 20$ has been added to the classification probabilities instead of adding Gaussian noise to the image frames. The procedure for creating the synthetic video was the same. In this experiment, the IPR values computed for the shaded (blurred) frames in Figure 14(c) are relatively small compared to those for Figure 13(c). Consequently, D-DPF in Figure 14(e) is affected much more by the observation. This experiment suggests that a care-

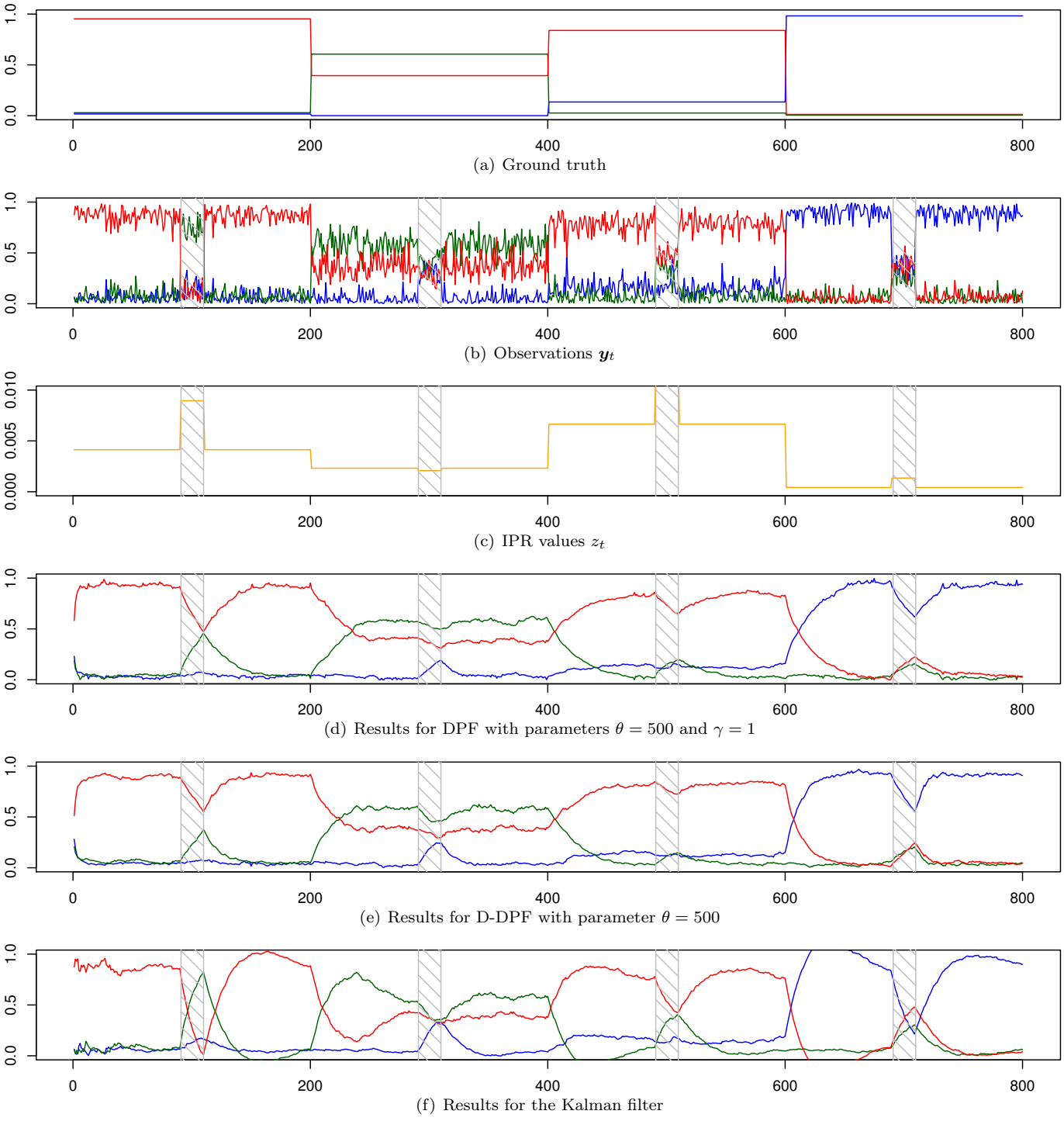



Figure 14: Smoothing results for a synthetic video using Dirichlet noise with parameter $s = 20$ (see Eq. 21). The horizontal axis shows the frame number. The vertical axis represents classification probabilities for the three classes of type A (blue), B (green), and C3 (red) (except (c)). From top to bottom, ground truth of classification probabilities, observations with no smoothing, IPR values, smoothing results for DPF, D-DPF with $K = 1,000$ particles, and a Kalman filter. Shaded frames are blurred Gaussian with $\sigma_{blur} = 5$.

fully selected model is necessary for the relation between $z_t$ and $\gamma_t$.

To obtain a quantitative evaluation, we compute the root-mean-square error (RMSE) between ground truth (Figures 13(a) and 14(a)) and the smoothing results (Figures 13(d)-(f) and 14(d)-(f)) over different amounts of Gaussian or Dirichlet noise. Figure 15(a) shows the RMSE for DPF, D-DPF, and the Kalman filter applied to synthetic videos with Gaussian noise for different values of $\sigma_{noise}$. Both DPF and D-DPF maintain low values of the

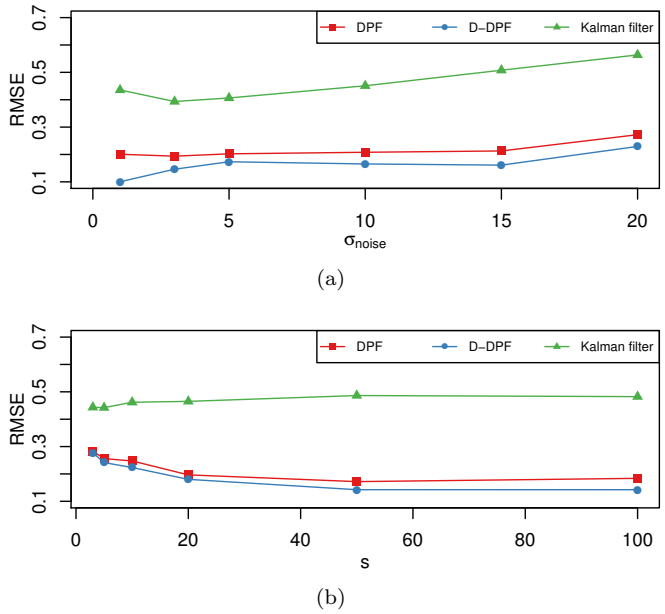



Figure 15: RMSE of the smoothing results for the synthetic videos shown in (a) Figure 13 and (b) Figure 14. The vertical axis shows RMSE. The horizontal axis is the value of $\sigma_{noise}$ for the Gaussian noise for (a) and the value of $s$ for the Dirichlet noise for (b) (see Eq. (21)).

RMSE. The RMSE is higher for the Kalman filter than for DPF and D-DPF for an entire range of $\sigma_{noise}$ values. Figure 15(b) shows the RMSE for synthetic videos with Dirichlet noise for different values of $s$. Note that larger values of $s$ generate smaller amounts of noise. Here again, D-DPF performs better than DPF and the Kalman filter.

### *5.3.2. Results for different $\theta$*

We compare the performance for different values of $\theta$ in the state transition. Figure 16 shows results for a synthetic video, which was created by the same procedure described above, except that two original images were used to create 200 frames. Then, Dirichlet noise with $s = 50$ was added to the classification probabilities. Figure 16(a) shows the classification probabilities for each original (noise-free) frame. For this synthetic video, two NBI images were used, each of which lasts 100 frames. Shading in Figures 16(b) through (e) indicates frames blurred with Gaussian with $\sigma_{blur} = 5$. We used $K = 1,000$ particles to generate the results in Figures 16(d) and (e).

At the discontinuities of frames 50, 100, and 150, smoothed probabilities are pulled to observations, and $\theta$ adjusts the speed of the convergence. When $\theta$ is small (e.g., 100), the state transition probability has a broad peak (the middle column of Figure 5), and successive states $\boldsymbol{x}_{t-1}$ and $\boldsymbol{x}_t$ are thus weakly linked; hence, the result rapidly converges to the observation, as in Figure 16(c). In contrast, when $\theta$ is relatively larger (500), the narrow peak of $p(\boldsymbol{x}_t \mid \boldsymbol{x}_{t-1})$ restricts $\boldsymbol{x}_t$ to be close to $\boldsymbol{x}_{t-1}$, resulting in a slow convergence such as that in Figure 16(e).

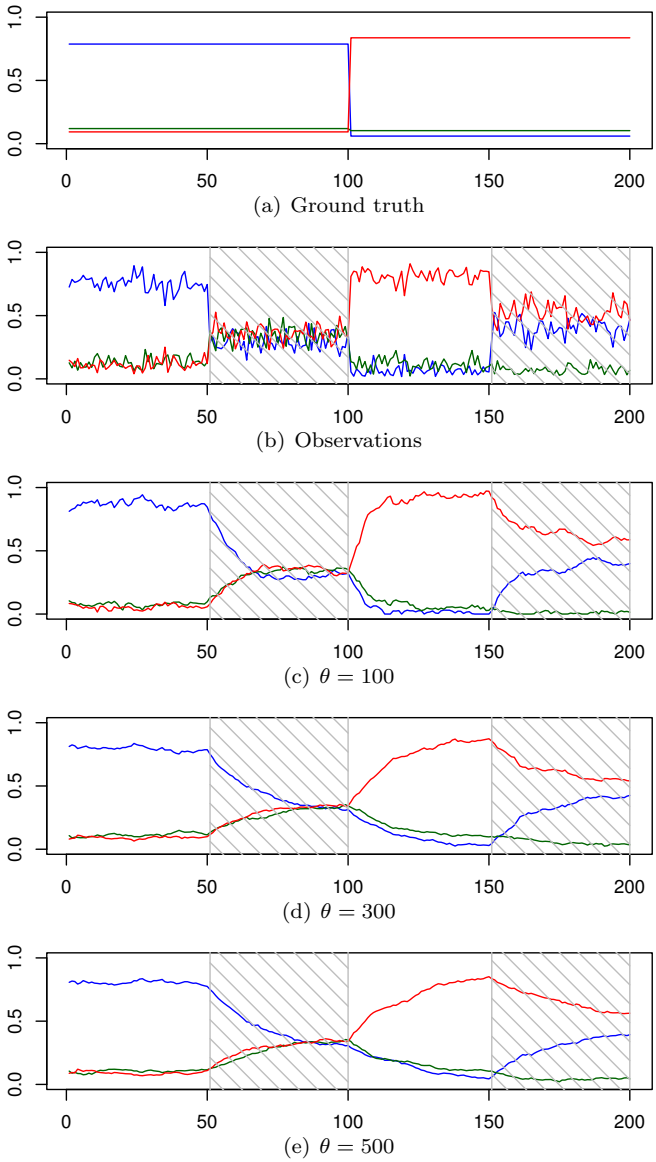
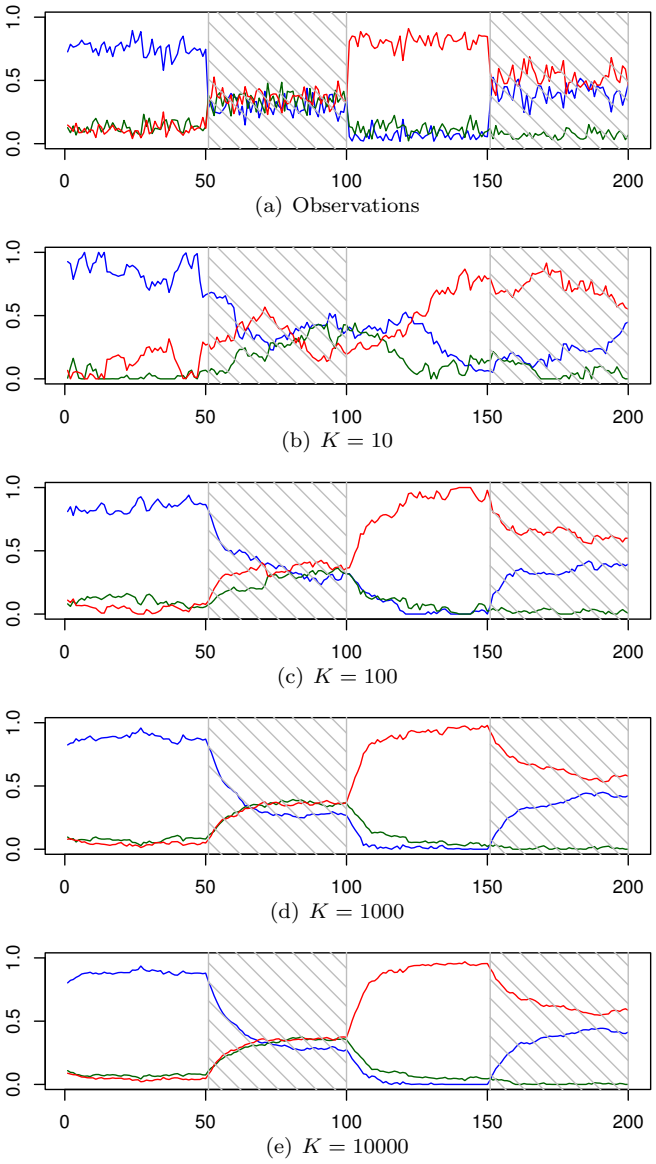



Figure 16: Smoothing results for a synthetic video with Dirichlet noise with parameter $s = 50$ (see Eq. 21) with three classes of type A (blue), B (green), and C3 (red). The horizontal axis shows the frame number. The vertical axis shows classification probabilities. From top to bottom: ground truth of classification probabilities, observation with no smoothing, smoothing results with $\theta = 100$, 300, and 500 by D-DPF with $K = 1,000$ particles. Shaded frames are blurred by Gaussian with $\sigma_{blur} = 5$.

As a simple extension, one might think of $\theta$ as another hidden variable that relates the defocus information and the state transition, as we did with $\gamma$ for the likelihood. However, we chose not to follow such a direction. If we loosely connected $\boldsymbol{x}_t$ to $\boldsymbol{x}_{t-1}$ as well as $\boldsymbol{y}_t$ when the frame is defocused, then $\boldsymbol{x}_t$ would not be under the control of either $\boldsymbol{x}_{t-1}$ or $\boldsymbol{y}_t$, and the result might be unpredictable. More sophisticated modeling of the relation between the defocus information and the state transition is left as future work.




(a) Observations

(b) $K = 10$

(c) $K = 100$

(d) $K = 1000$

(e) $K = 10000$

Figure 17: Smoothing results for a synthetic video with Dirichlet noise with parameter $s = 50$ (see Eq. 21) with three classes of type A (blue), B (green), and C3 (red). The horizontal axis shows the frame number. The vertical axis shows classification probabilities. From top to bottom: observation with no smoothing, smoothing results with the number of particles $K = 10, 100, 1,000$, and $10,000$. Shaded frames are blurred by Gaussian with $\sigma_{blur} = 5$.

#### 5.3.3. Number of particles

We evaluate the results in terms of the number of particles with the same dataset used in the last subsection because the optimal number of particles depends on each problem. Using a large number of particles generally provides good results, but there is a tradeoff due to increasing computational cost. We fix the parameter to $\theta = 100$ and change the number of particles to $K = 10, 100, 1,000$, and $10,000$. As shown in Figure 17, the smoothing effect is insufficient when using as few as $K = 10$ particles. In contrast, using many particles improves the accuracy of the

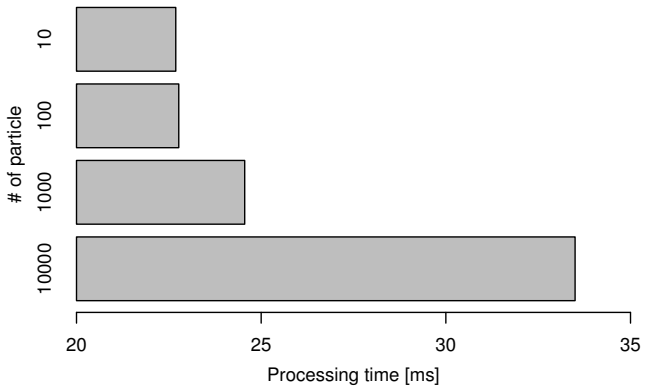



Figure 18: Computational cost of D-DPF per frame.

smoothing results. Evidently, $K = 100$ appears to be sufficient to achieve results comparable to the case wherein many more particles are used, e.g., $K = 1,000$ and $10,000$.

Figure 18 shows the computational cost per frame, which includes lines 3 to 13 in Algorithm 1. Even when we use $K = 10,000$ particles, it requires only 33 ms/frame, of which computing IPR takes 22 ms on average. Furthermore, frame-wise classification requires 50 ms. In total, it requires 88 ms $\cong$ 12 fps; this is a sufficient computation speed for a prototype system to be used in diagnosis support during actual endoscopic examinations. Further increase in speed can be achieved by additional fine-tuning of the system. Currently, our unoptimized implementation written in C++ uses a single thread on an Intel Core i5 (2.4 GHz) processor with 16 GB memory.

### 5.4. Results on real endoscopic videos

In this subsection, we demonstrate smoothing results for real endoscopic videos taken during actual endoscopic examinations. Demonstration videos of smoothed results are available as supplemental material.

For training, all of the 1,671 NBI image patches in the dataset were used. This dataset is unbalanced, but a preliminary experiment (not shown here) with a balanced dataset of 320 NBI image patches for each class showed results similar to those shown here.

Figure 19 shows observation and smoothing results for a video that captures a tumor labeled as type B. During the frames around frame numbers 150, 250, and 450, where observation and IPR values look nearly constant, endoscopists capture the screen to save images of the tumor, and the screen freezes. Due to defocus, type A is dominant between frames 30 and 120, and the observations are unstable particularly around frames 180, 290, 370, and 490. It is evident that the observations (classification probabilities) are highly unstable throughout the video, whereas the results from D-DPF are much smoother. The results obtained from DPF, shown in Figure 19(f), are also smooth, but slow to follow the observations. The result from D-

DPF with the same parameter shown in Figure 19(e) shows a quick follow; particularly, frames between 400 and 500 when observations are close to zero and one. Figure 19(g) shows the smoothing result obtained by a Kalman filter with the same parameters as those used for the results in Figures 13(f) and 14(f). We can see that the results are as slow to follow the observations as DPF. There is also overshooting as we have seen in the last section with the synthetic videos where observations suddenly change such as around frames 150, 250, 450, and 550.

Figures 20(a) to (c) show smoothing results for another video labeled as type A, wherein the frames around frame 200 are blurred and the observations are unstable. The results shown in Figure 20(c) are smooth and the probabilities for type A have the largest values for all frames as IPR values keep lower values. Figures 20(d) to (f) show smoothing results for yet another video labeled as type C3. The results shown in Figure 20(f) are smoother than the observations in Figure 20(d) as all frames are defocused slightly and IPR values are relatively high.

However, the results in Figure 20 (f) are type C3 only in frames between 120 and 190 because of the severe instability of the frame-wise classification results. In particular, the results for frames between 60 and 90, and between 190 and 270, are type B because it is dominant in the observations during these frames. This may be caused by defocus of frames, as well as other problematic issues that occur in real endoscopic videos such as illumination change, color bleeding, and abrupt camera motion.

## 6. Conclusion

We have proposed a novel method–D-DPF–to smooth the classification probabilities obtained from frame-wise endoscopic image classification by incorporating defocus information into a particle filter with a Dirichlet distribution. We assumed that the defocus information extracted from each frame influences classification probabilities, and we proposed linking the Dirichlet likelihood to the defocus information and the IPR proposed by Oh et al. [7], which is a ratio of the number of edge pixels isolated from neighbor edge pixels. Then we sampled parameter $\gamma_t$ in the likelihood from a Rayleigh distribution.

For endoscopists, unstable recognition results such as those in Figure 1 are difficult to use for diagnosis. The proposed smoothing method improves the visibility and understandability of the recognition results and facilitates the use of the results for diagnosis. Moreover, the proposed method has the potential to be used for training endoscopists who have less experience of endoscopic examinations.

D-DPF can be extended in several ways. One possible extension is to address other causes of instability. We have focused on defocus information herein, but other causes also exist. One example is color bleeding due to the following property of NBI endoscopes: different wavelengths of light are used to create a single frame by rotating a filter in front of the light sources. Thus, color bleeding (i.e., different color illuminations appear at the same time) occurs when the assumption that the scene is temporally static is violated owing to the large motion of the endoscope. Rapid movement of the endoscope also results in motion blur, another cause of instability. The proposed D-DPF might still be applicable in such situations if we could introduce metrics representing color bleeding or motion blur instead of defocus information. For other frame-wise classification results with four or five-classes, we can apply D-DPF by simply changing the dimension $N$ of the Dirichlet distribution. Gastrointestinal endoscopic videos are also in the application range of D-DPF. Given additional information along with the signals to be smoothed, effective smoothing results can be obtained.

How to visualize the results more effectively is another issue that deserves further attention. For every frame, we compute the mode of the Dirichlet distribution estimated in the update step as the smoothed classification probabilities. These probability values are displayed as in the videos in the supplementary material. Furthermore, the estimated label (the class having the largest probability from the estimated mode) is displayed as a colored rectangle shown at the patch used by the frame-wise recognition. Other possible means of visualization include displaying classification probability curves that are similar to an electrocardiogram or visualizing the estimated Dirichlet distribution shapes instead of probabilities and labels. In any case, further consideration is needed in terms of human-computer interaction.

In addition to the visualization issue, our future work includes embedding the proposed method into an actual working system for clinical evaluations. We also must explore alternative ways to represent defocus information (other than IPR) and other sampling strategies (apart from the Rayleigh distribution) for the likelihood parameter.

## Acknowledgements

This work was supported in part by JSPS KAKENHI grants numbers 14J00223, 26280015, and 24591026 as well as the Mazda Foundation grant number 13KK-210, and the Semiconductor Technology Academic Research Center (STARC).

[1] S. Tanaka, T. Kaltenbach, K. Chayama, R. Soetikno, High-magnification colonoscopy (with videos), Gastrointestinal Endoscopy 64 (2006) 604 – 613.

[2] A. Meining, T. Rösch, R. Kiesslich, M. Muders, F. Sax, W. Heldwein, Inter- and intra-observer variability of magnification chromoendoscopy for detecting specialized intestinal metaplasia at the gastroesophageal junction, Endoscopy 36 (2004) 160—164.

[3] B. Mayinger, Y. Oezturk, M. Stolte, G. Faller, J. Benninger, D. Schwab, et al., Evaluation of sensitivity and inter- and intra-observer variability in the detection of intestinal metaplasia and dysplasia in barrett's esophagus with enhanced magnification endoscopy, Scand J Gastroenterol 41 (2006) 349–56.

[4] S. Oba, S. Tanaka, S. Oka, H. Kanao, S. Yoshida, F. Shimamoto, et al., Characterization of colorectal tumors using narrow-band imaging magnification: combined diagnosis with both pit pattern and microvessel features, Scand J Gastroenterol 45 (2010) 1084–92.

[5] Y. Takemura, S. Yoshida, S. Tanaka, R. Kawase, K. Onji, S. Oka, et al., Computer-aided system for predicting the histology of colorectal tumors by using narrow-band imaging magnifying colonoscopy (with video), Gastrointest Endosc 75 (2012) 179–85.

[6] T. Tamaki, J. Yoshimuta, M. Kawakami, B. Raytchev, K. Kaneda, S. Yoshida, et al., Computer-aided colorectal tumor classification in NBI endoscopy using local features, Medical Image Analysis 17 (2013) 78 – 100.

[7] J. Oh, S. Hwang, J. Lee, W. Tavanapong, J. Wong, P. C. de Groen, Informative frame classification for endoscopy video, Medical Image Analysis 11 (2007) 110 – 127.

[8] T. Hirakawa, T. Tamaki, B. Raytchev, K. Kaneda, T. Koide, Y. Kominami, et al., Smoothing posterior probabilities with a particle filter of dirichlet distribution for stabilizing colorectal nbi endoscopy recognition, in: Image Processing (ICIP), 2013 20th IEEE International Conference on, 2013, pp. 621–625.

[9] H. Kanao, S. Tanaka, S. Oka, M. Hirata, S. Yoshida, K. Chayama, Narrow-band imaging magnification predicts the histology and invasion depth of colorectal tumors, Gastrointestinal Endoscopy 69 (2009) 631 – 636.

[10] Bowel cancer statistics, http://www.cancerresearchuk.org/cancer-info/cancerstats/types/bowel/ (Accessed: 15 January 2015), 2015.

[11] K. Gono, T. Obi, M. Yamaguchi, N. Ohyama, H. Machida, Y. Sano, et al., Appearance of enhanced tissue features in narrow-band endoscopic imaging, J Biomed Opt 9 (2004) 568–77.

[12] H. Machida, Y. Sano, Y. Hamamoto, M. Muto, T. Kozu, H. Tajiri, et al., Narrow-band imaging in the diagnosis of colorectal mucosal lesions: a pilot study, Endoscopy 36 (2004) 1094–8.

[13] Y. Sano, T. Horimatsu, K. I. Fu, A. Katagiri, M. Muto, H. Ishikawa, Magnifying observation of microvascular architecture of colorectal lesions using a narrow-band imaging system, Digestive Endoscopy 18 (2006) S44–S51.

[14] D. E. Maroulis, D. K. Iakovidis, S. A. Karkanis, D. A. Karras, Cold: a versatile detection system for colorectal lesions in endoscopy video-frames, Comput Methods Programs Biomed 70 (2003) 151–66.

[15] S. Karkanis, D. Iakovidis, D. Maroulis, D. Karras, M. Tzivras, Computer-aided tumor detection in endoscopic video using color wavelet features, Information Technology in Biomedicine, IEEE Transactions on 7 (2003) 141–152.

[16] D. K. Iakovidis, D. E. Maroulis, S. A. Karkanis, An intelligent system for automatic detection of gastrointestinal adenomas in video endoscopy, Computers in Biology and Medicine 36 (2006) 1084 – 1103.

[17] W. Li, U. Gustafsson, A. Yoursif, et al., Automatic colonic lesion detection and tracking in endoscopic videos, in: SPIE Medical Imaging, International Society for Optics and Photonics, 2011, pp. 79632L–79632L.

[18] S. Y. Park, D. Sargent, I. Spofford, K. G. Vosburgh, Y. A-Rahim, A colon video analysis framework for polyp detection, IEEE transactions on bio-medical engineering 59 (2012) 1408—1418.

[19] M. Biswas, D. Dey, Bi-dimensional statistical empirical mode decomposition-based video analysis for detecting colon polyps using composite similarity measure, in: L. C. Jain, S. Patnaik, N. Ichalkaranje (Eds.), Intelligent Computing, Communication and Devices, volume 309 of *Advances in Intelligent Systems and Computing*, Springer India, 2015, pp. 297–308.

[20] M. Häfner, C. Kendlbacher, W. Mann, W. Taferl, F. Wrba, A. Gangl, et al., Pit pattern classification of zoom-endoscopic colon images using histogram techniques, in: Signal Processing Symposium, 2006. NORSIG 2006. Proceedings of the 7th Nordic, 2006, pp. 58–61.

[21] M. Häfner, R. Kwitt, F. Wrba, A. Gangl, A. Vecsei, A. Uhl, One-against-one classification for zoom-endoscopy images, in: Advances in Medical, Signal and Information Processing, 2008. MEDSIP 2008. 4th IET International Conference on, 2008, pp. 1–4.

[22] M. Häfner, A. Gangl, R. Kwitt, A. Uhl, A. Vécsei, F. Wrba, Improving pit–pattern classification of endoscopy images by a combination of experts, in: G.-Z. Yang, D. Hawkes, D. Rueckert, A. Noble, C. Taylor (Eds.), Medical Image Computing and Computer-Assisted Intervention – MICCAI 2009, volume 5761 of *Lecture Notes in Computer Science*, Springer Berlin Heidelberg, 2009, pp. 247–254.

[23] M. Häfner, A. Gangl, M. Liedlgruber, A. Uhl, A. Vecsei, F. Wrba, Combining gaussian markov random fields with the discrete-wavelet transform for endoscopic image classification, in: Digital Signal Processing, 2009 16th International Conference on, 2009, pp. 1–6.

[24] M. Häfner, A. Gangl, M. Liedlgruber, A. Uhl, A. Vecsei, F. Wrba, Pit pattern classification using extended local binary patterns, in: Information Technology and Applications in Biomedicine, 2009. ITAB 2009. 9th International Conference on, 2009, pp. 1–4.

[25] M. Häfner, A. Gangl, M. Liedlgruber, A. Uhl, A. Vecsei, F. Wrba, Pit pattern classification using multichannel features and multiclassification, in: T. P. Exarchos, A. Papadopoulos, D. I. Fotiadis (Eds.), Handbook of Research on Advanced Techniques in Diagnostic Imaging and Biomedical Applications, IGI Global, Hershey, PA, USA, 2009, pp. 335–350.

[26] M. Häfner, R. Kwitt, A. Uhl, A. Gangl, F. Wrba, A. Vécsei, Feature extraction from multi-directional multi-resolution image transformations for the classification of zoom-endoscopy images, Pattern Analysis and Applications 12 (2009) 407–413.

[27] M. Häfner, R. Kwitt, A. Uhl, F. Wrba, A. Gangl, A. Vécsei, Computer-assisted pit-pattern classification in different wavelet domains for supporting dignity assessment of colonic polyps, Pattern Recognition 42 (2009) 1180 – 1191.

[28] M. Häfner, A. Gangl, M. Liedlgruber, A. Uhl, A. Vécsei, F. Wrba, Classification of endoscopic images using delaunay triangulation-based edge features, in: A. Campilho, M. Kamel (Eds.), Image Analysis and Recognition, volume 6112 of *Lecture Notes in Computer Science*, Springer Berlin Heidelberg, 2010, pp. 131–140.

[29] M. Häfner, A. Gangl, M. Liedlgruber, A. Uhl, A. Vecsei, F. Wrba, Endoscopic image classification using edge-based features, in: Pattern Recognition (ICPR), 2010 20th International Conference on, 2010, pp. 2724–2727.

[30] R. Kwitt, A. Uhl, Modeling the marginal distributions of complex wavelet coefficient magnitudes for the classification of zoom-endoscopy images, in: Computer Vision, 2007. ICCV 2007. IEEE 11th International Conference on, 2007, pp. 1–8.

[31] R. Kwitt, A. Uhl, Multi-directional multi-resolution transforms for zoom-endoscopy image classification, in: M. Kurzynski, E. Puchala, M. Wozniak, A. Zolnierek (Eds.), Computer Recognition Systems 2, Advances in Soft Computing, volume 45, Springer Berlin Heidelberg, 2007, pp. 35–43.

[32] R. Kwitt, A. Uhl, M. Häfner, A. Gangl, F. Wrba, A. Vécsei, Predicting the histology of colorectal lesions in a probabilistic framework, in: Computer Vision and Pattern Recognition Workshops (CVPRW), 2010 IEEE Computer Society Conference on, 2010, pp. 103–110.

[33] S. Gross, T. Stehle, A. Behrens, R. Auer, T. Aach, R. Winograd, et al., A comparison of blood vessel features and local binary patterns for colorectal polyp classification, Proc. SPIE 7260 (2009) 72602Q–72602Q–8.

[34] T. Stehle, R. Auer, S. Gross, A. Behrens, J. Wulff, T. Aach, et al., Classification of colon polyps in nbi endoscopy using vascularization features, Proc. SPIE 7260 (2009) 72602S–72602S–12.

[35] J. J. W. Tischendorf, S. Gross, R. Winograd, H. Hecker, R. Auer, A. Behrens, et al., Computer-aided classification of

colorectal polyps based on vascular patterns: a pilot study, Endoscopy 42 (2010) 203–7.

[36] S. Manivannan, R. Wang, E. Trucco, A. Hood, Automatic normal-abnormal video frame classification for colonoscopy, in: Biomedical Imaging (ISBI), 2013 IEEE 10th International Symposium on, 2013, pp. 644–647.

[37] S. Manivannan, R. Wang, M. Trujillo, J. Hoyos, E. Trucco, Video-specific svms for colonoscopy image classification, in: X. Luo, T. Reichl, D. Mirota, T. Soper (Eds.), Computer-Assisted and Robotic Endoscopy, Lecture Notes in Computer Science, Springer International Publishing, 2014, pp. 11–21.

[38] M. Arnold, A. Ghosh, G. Lacey, S. Patchett, H. Mulcahy, Indistinct frame detection in colonoscopy videos, in: Machine Vision and Image Processing Conference, 2009. IMVIP '09. 13th International, 2009, pp. 47–52.

[39] J. Liu, K. Subramanian, T. Yoo, A robust method to track colonoscopy videos with non-informative images, International Journal of Computer Assisted Radiology and Surgery 8 (2013) 575–592.

[40] R. Liu, Z. Li, J. Jia, Image partial blur detection and classification, in: Computer Vision and Pattern Recognition, 2008. CVPR 2008. IEEE Conference on, 2008, pp. 1–8.

[41] G. Kitagawa, Monte carlo filter and smoother for non-gaussian nonlinear state space models, Journal of Computational and Graphical Statistics 5 (1996) 1–25.

[42] N. Gordon, D. Salmond, A. Smith, Novel approach to nonlinear/non-gaussian bayesian state estimation, Radar and Signal Processing, IEE Proceedings F 140 (1993) 107–113.

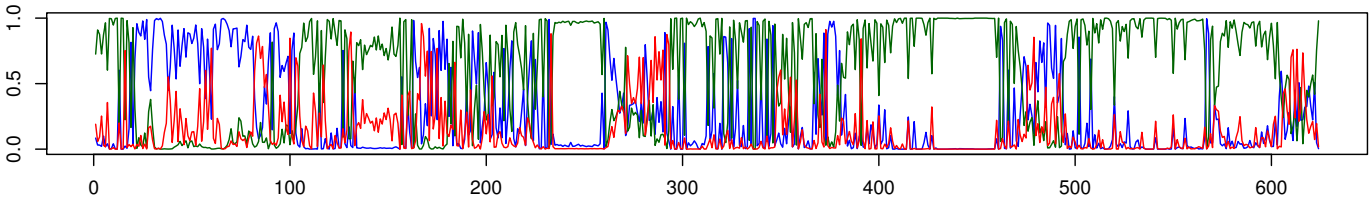
[43] J. V. Candy, Bayesian signal processing: classical, modern, and particle filtering methods, Adaptive and learning systems for signal processing, communications, and control, Wiley, Hoboken, N.J., 2009.

[44] N. Gordon, B. Ristic, S. Arulampalam, Beyond the kalman filter: Particle filters for tracking applications, Artech House, London (2004).

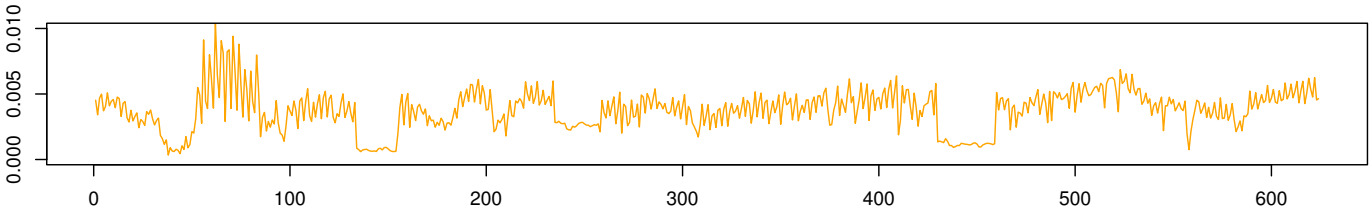
[45] C. M. Bishop, Pattern Recognition and Machine Learning (Information Science and Statistics), Springer-Verlag New York, Inc., Secaucus, NJ, USA, 2006.

[46] M. Evans, N. Hastings, B. Peacock, Statistical Distributions, Wiley Series in Probability and Statistics, Wiley, New York, 2000.

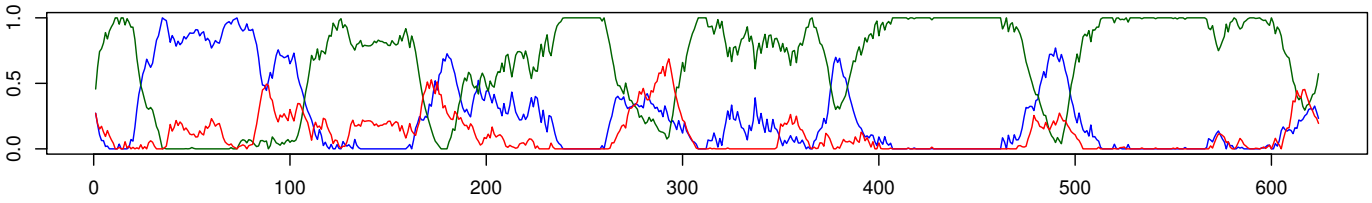
[47] A. Narayanan, Algorithm as 266: Maximum likelihood estimation of the parameters of the dirichlet distribution, Journal of the Royal Statistical Society. Series C (Applied Statistics) 40 (1991) pp. 365–374.



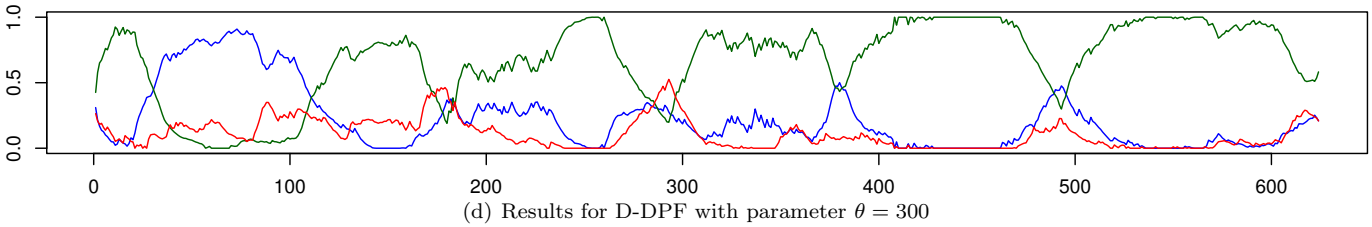

(a) Observations $\boldsymbol{y}_t$



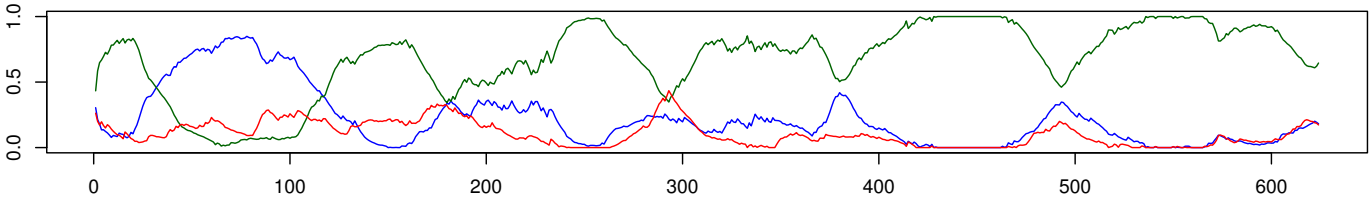

(b) IPR values $z_t$



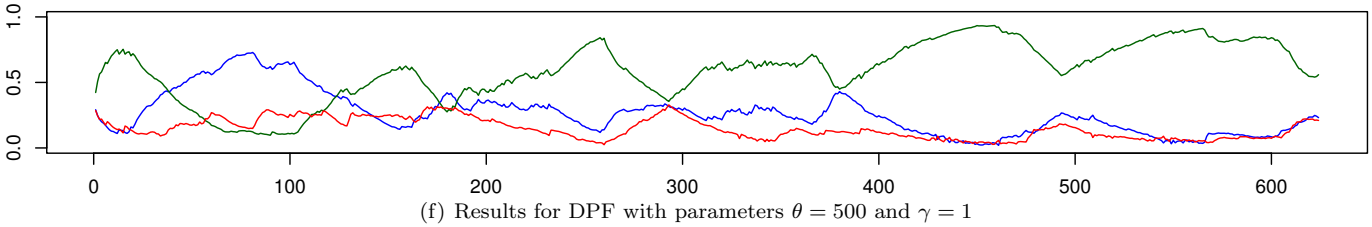

(c) Results for D-DPF with parameter $\theta = 100$



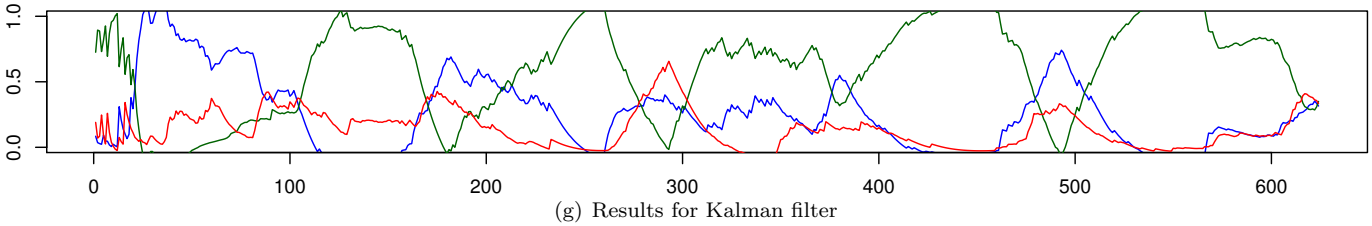

(d) Results for D-DPF with parameter $\theta = 300$




(e) Results for D-DPF with parameter $\theta = 500$

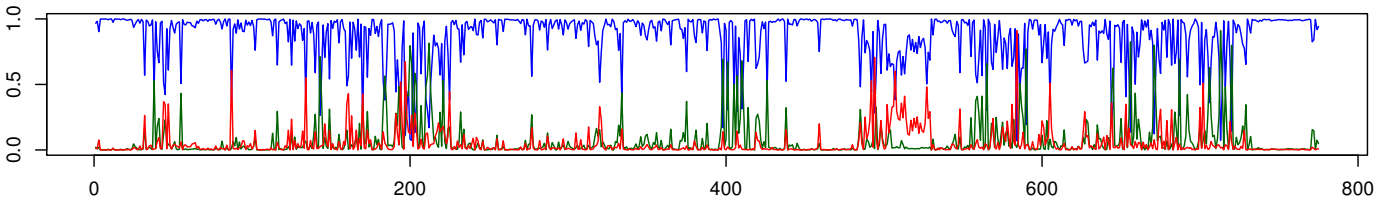



(f) Results for DPF with parameters $\theta = 500$ and $\gamma = 1$

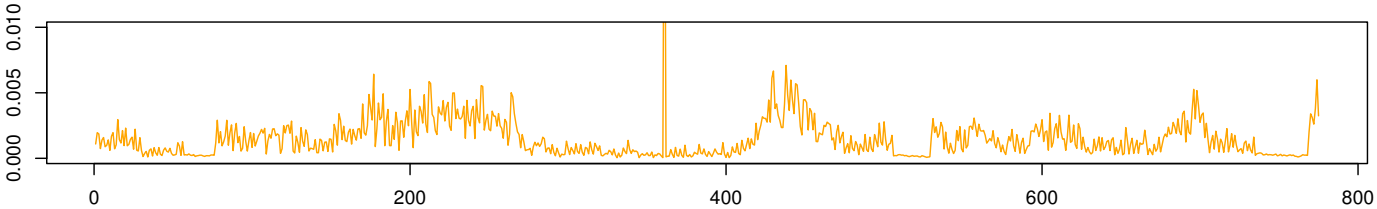



(g) Results for Kalman filter

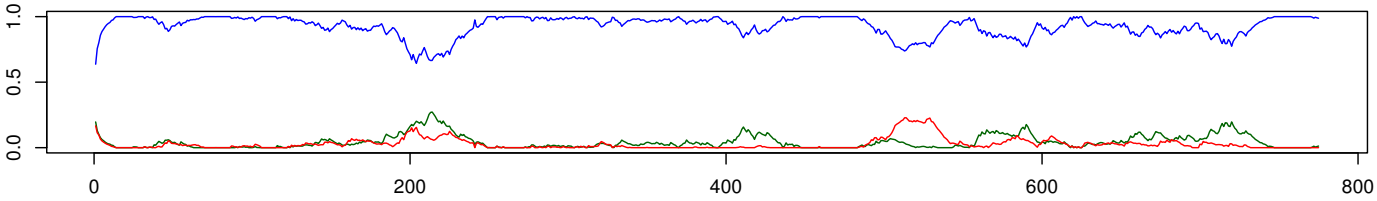
Figure 19: Smoothing results on a real endoscopic video of 629 frames labeled as type B. The horizontal axis shows the frame number. The vertical axis shows classification probabilities for the three classes of type A (blue), B (green), and C3 (red) except (b) and the IPR value for (b). From top to bottom, observations with no smoothing, IPR values, smoothing results for D-DPF with parameter $\theta = 100$, 300, and 500, smoothing results for DPF with parameter $\theta = 500$ and $\gamma = 1$, and smoothing results for the Kalman filter.



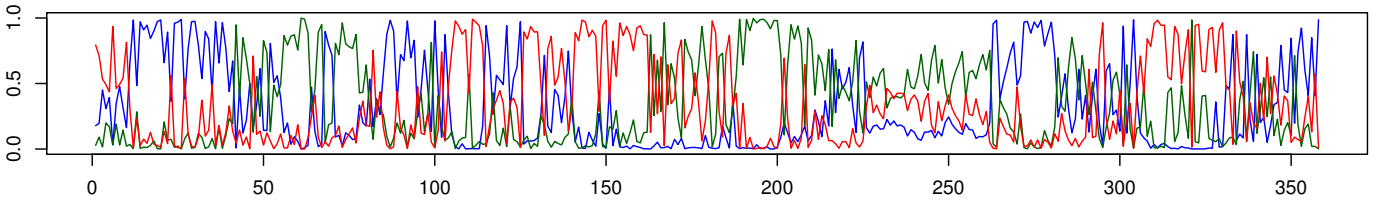

(a) Observations $\boldsymbol{y}_t$ of type A endoscopic video



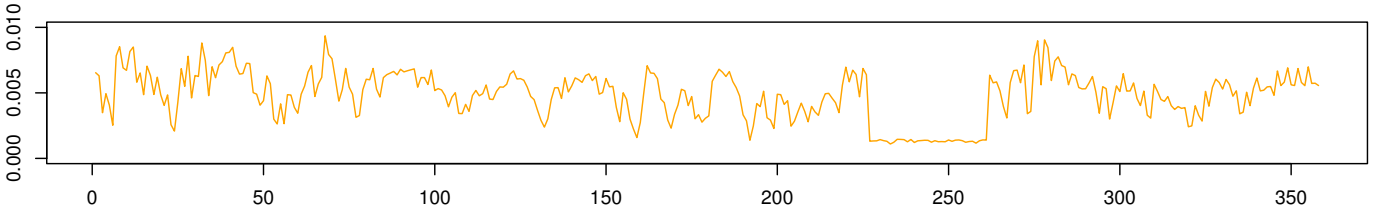

(b) IPR values $z_t$ for (a)



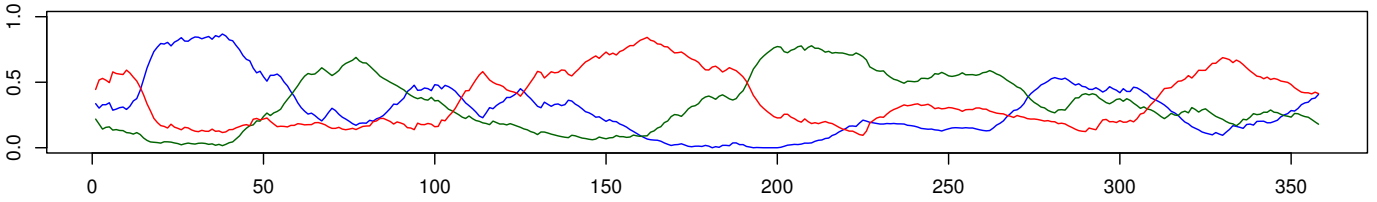

(c) Results of (a) for D-DPF with parameter $\theta = 500$




(d) Observations $\boldsymbol{y}_t$ of type C3 endoscopic video




(e) IPR values $z_t$ for (d)




(f) Results of (d) for D-DPF with parameter $\theta = 500$

Figure 20: Smoothing results on real endoscopic videos labeled as types A (775 frames) and C (358 frames). The horizontal axis shows the frame number. The vertical axis shows classification probabilities for the three classes of type A (blue), B (green), and C3 (red). From top to bottom, observations of a video labeled as type A, IPR values for (a), smoothing results for (a) by D-DPF, observations of a video labeled as type C3, IPR values for (d) and smoothing results for (d) by D-DPF. Parameter $\theta$ of D-DPF is set to 500.