Pattern Recognition 63 (2017) 612–624

Contents lists available at ScienceDirect

Pattern Recognition

journal homepage: www.elsevier.com/locate/pr

Heterogeneous data analysis: Online learning for medical-image-based diagnosis☆

Yuichi Motai[a,*,1], Nahian Alam Siddique[a,1], Hiroyuki Yoshida[b,2]

[a] Virginia Commonwealth University, USA
[b] Massachusetts General Hospital and Harvard Medical School, USA

ARTICLE INFO

Keywords:
Online learning
Computed tomographic colonography
Heterogeneous data analysis
Kernel feature analysis
Computer-aided detection
Principal composite kernel feature analysis

ABSTRACT

Heterogeneous Data Analysis (HDA) is proposed to address a learning problem of medical image databases of Computed Tomographic Colonography (CTC). The databases are generated from clinical CTC images using a Computer-aided Detection (CAD) system, the goal of which is to aid radiologists' interpretation of CTC images by providing highly accurate, machine-based detection of colonic polyps. We aim to achieve a high detection accuracy in CAD in a clinically realistic context, in which additional CTC cases of new patients are added regularly to an existing database. In this context, the CAD performance can be improved by exploiting the heterogeneity information that is brought into the database through the addition of diverse and disparate patient populations. In the HDA, several quantitative criteria of data compatibility are proposed for efficient management of these online images. After an initial supervised offline learning phase, the proposed online learning method decides whether the online data are heterogeneous or homogeneous. Our previously developed Principal Composite Kernel Feature Analysis (PC-KFA) is applied to the online data, managed with HDA, for iterative construction of a linear subspace of a high-dimensional feature space by maximizing the variance of the non-linearly transformed samples. The experimental results showed that significant improvements in the data compatibility were obtained when the online PC-KFA was used, based on an accuracy measure for long-term sequential online datasets. The computational time is reduced by more than 93% in online training compared with that of offline training.

1. Introduction

Colon cancer is the second leading cause of cancer deaths in the United States. An estimated 50,310 deaths were expected to occur in 2014 [60]. Colon cancers develops from small adenomatous polyps that arise on the inner colonic mucosa. Most polyps are benign when they appear; however, some develop into cancer over time. Thus, early detection of polyps through screening is a promising approach for preventing death from colon cancer. Currently, the detection of polyps is performed by mainly colonoscopy. However, in the screening context, Computed Tomographic Colonography (CTC) [13–16], also known as virtual colonoscopy, has been emerging as a promising non-invasive alternative approach to invasive colonoscopy.

However, CTC-based colon cancer screening often requires a lengthy (15–30 min) interpretation time by radiologists, and their diagnostic performance on polyps varies substantially according to their skill [5–7]. Perceptual errors by radiologists are a major source of false-negative (missed) polyps in CTC examinations [56]. Computer-aided Detection (CAD) for CTC refers to a computerized scheme that detects colorectal neoplasms automatically in CTC data and reports their locations to radiologists who make the final diagnostic decision. Observer studies have shown that the use of CAD in CTC yields a statistically significant improvement in detection sensitivity and also reduces inter-observer variance [62]. Fig. 1 shows an example of the user interface of a CAD system for interpretation of CTC images [9–11].

Our goal in this study was to improve the performance of CAD in a more realistic clinical context than that of observer studies, in which additional CTC cases of new patients are added to an existing database on a regular basis. Medical image datasets obtained at diagnostic

☆ This study was supported by Institutional Research Grant IRG-73-001-31 from the American Cancer Society, NSF CAREER Award 1054333, CTSA UL1TR000058 from the National Center for Advancing Translational Sciences, Center for Clinical and Translational Research Endowment Fund of Virginia Commonwealth University (VCU), Presidential Research Incentive Program at VCU, and partly supported by National Institutes of Health grants CA095279 and CA166816.

* Corresponding author.
[1] Electrical & Computer Engineering Department, Virginia Commonwealth University, USA.
[2] Department of Radiology, Massachusetts General Hospital and Harvard Medical School, USA.

http://dx.doi.org/10.1016/j.patcog.2016.09.035
Received 21 December 2015; Received in revised form 25 May 2016; Accepted 21 September 2016
Available online 29 September 2016

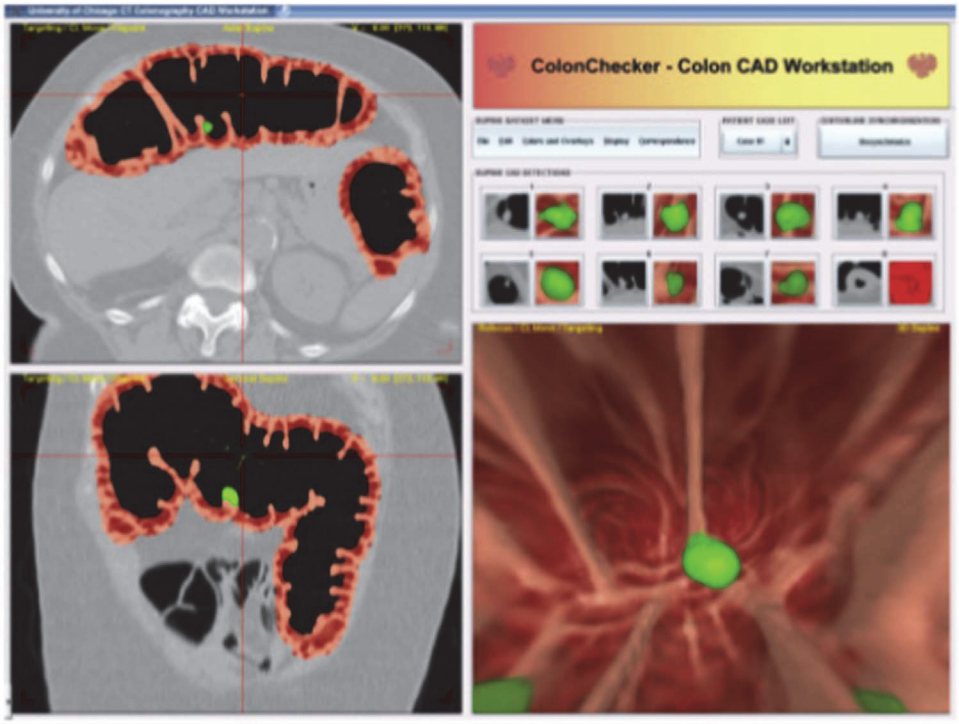

Fig. 1. Example of the user interface of a CAD system for interpretation of CTC images [9].

facilities accumulate over long periods as the number of patients increases. Adding these *online* datasets to update existing *offline* datasets is frequently required in many medical applications including CTC. In a large-scaled medical-image-based diagnosis such as the screening of colon cancer by CTC, one of the major challenges is how to process heterogeneous data that are generated when existing datasets are combined with new datasets. The proposed solution is Heterogeneous Data Analysis (HAD) for online CTC datasets with a larger number of patients, and use of online kernel learning for improvement of the performance of CAD. The concept of online learning has been shown to be effective for non-linear and online data analysis [18,19], and it is gaining popularity in the machine learning community [20].

Kernel analysis has also been shown to be effective for medical image analysis [21,22]. However, there has been little evidence as to whether it is effective for automated detection of colonic polyps in datasets with increasing size and heterogeneity. Recent success of kernel methods is capitalizing on broad pattern classification problems [1,21,23–25,63–65], especially in non-linear classification and clustering [26,27]. We previously developed a fast feature analysis technique called Accelerated Kernel Feature Analysis (AKFA) [28], and we extended it to Composite Kernel Analysis, called Principal Composite Kernel Feature Analysis (PC-KFA) [1]. These methods map the original, low-dimensional feature space into a higher-dimensional feature space. Such a high-dimensional feature space is expected to have a greater classification power than that of the original feature space, as suggested by the Vapnik-Chervonenkis theory in [29,30]. The PC-KFA method extracts texture-based features from the polyp candidates generated by a shape-based CAD system. In this paper, we propose a new online method, called *online PC-KFA*, which is specifically designed for HDA.

The main contributions of this paper are the development of a scale-free online learning framework that manages the diversity and complexity of large image databases, and its application to the improvement of the accuracy and speed of CAD systems in the detection of polyps in clinical CTC datasets. We address several issues such as data confidentiality and auditability as well as scalable storages to make larger data-driven CAD feasible.

The significance of the paper is that, to our knowledge, this is the first study that deals with large online CTC data by incorporating online learning. We develop a new solution for the problems of massive expansion in scale, diversity, and complexity of CTC databases. The data acquired over a long period of time can be highly diverse, and each dataset is unique in nature. Thus, obtaining a clear distinction between heterogeneous and homogeneous large online datasets is an important but challenging task. The HDA is an effective framework for solving this problem. The proposed online PC-KFA method can reevaluate and change the criteria established during the training phase for the algorithm to train the data correctly. This allows efficient differentiation of polyps from false positive detections, and thus improve detection performance of CAD for CTC.

The remainder of this paper is organized as follows. Section 2 provides an introduction and a brief review of the existing kernel-based feature extraction methods, kernel selection, and PC-KFA for the detection of polyps in the offline CTC data. Section 3 describes HDA as a new analytical measure of homogeneity or heterogeneity of the data. Section 4 discusses long-term online learning for large-size incoming data and their segmentation procedures. Section 5 presents the experimental results of the classification between polyps and false positives, followed by conclusions in Section 6.

2. Kernel basics: a brief review

We briefly review existing Kernel Principal Component Analysis (KPCA) [31–35] in Section 2.1. Kernel Selection is described in Section 2.2, our newly developed Kernel Adaptation of PC-KFA is described briefly in Section 2.3, and our online machine learning CAD methods for CTC is presented in Section 2.4.

2.1. Kpca

KPCA uses a Mercer kernel [36] to perform a linear principal component analysis of the transformed image. The eigenvalues λ_j and eigenvectors e_j are obtained by solving,

$$\lambda_j e_j = Se_j = \sum_{i=1}^{n} \Phi(x_i)\Phi(x_i)^T e_j = \sum_{i=1}^{n} \langle e_j, \Phi(x_i)\rangle \Phi(x_i), \tag{1}$$

where S is the scatter matrix, $\Phi(x_i)$ is the projection of x_i into higher dimensional space, in which dot product is given by the corresponding kernel function. If K is a $n \times n$ Gram matrix where the elements $k_{ij} = \langle \Phi(x_i), \Phi(x_j)\rangle$ and $a_j = [a_{j1}\ a_{j2}\ \cdots\ a_{jn}]^T$ are the eigenvectors associated with eigenvalues, λ_j, then the dual eigenvalue problem equivalent to the problem can be expressed as follows: $\lambda_j a_j = Ka_j$.

The KPCA process can then be summarized as follows:

(1) Calculate the Gram matrix, K, using kernels, which contains the inner products between pairs of image vectors.
(2) Use $\lambda_j a_j = Ka_j$ to get the coefficient vectors a_j for j=1,2,...,n.
(3) The projection of a test point $x \in \mathbb{R}^d$ along the j^{th} eigenvector is:

$$\langle e_j, \Phi(x)\rangle = \sum_{i=1}^{n} a_{ji}\langle \Phi(x_i), \Phi(x)\rangle = \sum_{i=1}^{n} a_{ji} k(x, x_i). \tag{2}$$

The above implicitly contains an eigenvalue problem of rank n; thus, the computational complexity of KPCA is $O(n^3)$. In addition, each resulting eigenvector is represented as a linear combination of n terms.

The success of the KPCA largely depends on the choice of the kernel used for constructing the Gram matrix [37–40]. Han also reported the utility of KPCA and its derivatives in detecting molecular patterns of cancer [58]. According to no free lunch theorem [41], there is no single best kernel function in general; rather, the performance of a kernel function depends on the applications as described in the following section.

2.2. Kernel Selection

For efficient feature analysis, extraction of the salient features of polyps is essential because of the large data size and the 3-D nature of the polyps. Moreover, the distribution of the image features of polyps is expected to be non-linear. The problem is how to select an ideal non-linear operator. Some of the commonly used kernels are referred to as Kernel 1 to Kernel 4 [42–45] as follows: 1) Linear kernel, 2) Gaussian Radial Basis Function (RBF) kernel, 3) Laplace RBF kernel, and 4) Sigmoid kernel. Use of multiple kernels, instead of one, for better detection of salient features is known to be effective [46]. For example, in Ref. [47] the use of composite kernels in extracting interesting visual features from images was successfully demonstrated.

We thus adopt a data-dependent kernel [28] to capture the relationship among the data in the classification task by using the composite form. This data-dependent composite kernel, k_r, r =1, 2, 3 and 4, can be formulated as:

$$k_r(x_i, x_j) = q_r(xi) q_r(xj) p_r(xi, xj) \tag{3}$$

where, $x \in \mathbb{R}^d$, $p_r(x_i, x_j)$ is one kernel among 4 chosen kernels, and $q(x_i)$ is a factor function, with the form: $q_r(x_i) = a_{r0} + \sum_{m=1}^{n} a_{rm} k_0(x_i, x_m)$, where, $k_0(xi, xm) = \exp(-||xi - xm||^2/2\sigma^2)$and a_{rm} are the *combination coefficient* or the *weighting coefficients*. These two terms will be used interchangeably throughout this article.

Let the kernel matrices corresponding to $k(xi, xj)$ and $p_r(xi, xj)$ be K_r and P_r, respectively. We can express the data-dependent kernel K_r as:

$$K_r = [q_r(xi)\ \ q_r(xj)\ \ p_r(xi, xj)]_{n\times n} \tag{4}$$

Defining Q_i as the diagonal matrix of elements $\{q_i(x_1), q_i(x_2), \ldots, q_i(x_{xn})\}$, we can express K_r as: $K_r = Q_r P_r Q_r$.

The criterion for selecting the best kernel function involves finding the kernel that produces the largest eigenvalue [31–35]. Then, the eigenvector corresponding to the maximum eigenvalue provides the optimum solution. Once we derive the eigenvectors, i.e., the combination coefficients of all 4 different kernels, we proceed to construct q_r (=$K_0 a_r(n)$) and Q_r to find the corresponding Gram matrices of these kernels. Once we have these Gram matrices ready, we can find the optimum kernels for the given dataset. To undertake this procedure, we have to arrange the eigenvalues that determined the combination coefficients for all kernel functions in a descending order.

Zheng et al. proposed a similar method for batch learning, in which the input data were divided into a few groups of similar sizes, and KPCA was applied to each group [19]. A set of eigenvectors was obtained for each group and the final set of features was obtained by application of KPCA to a subset of these eigenvectors. The application of the online concept to Principal Component Analysis is often referred to Incremental Principal Component Analysis [15,48–51], and it has been shown to be computationally effective in many image processing applications and pattern classification systems. Kernel based methods are also used effectively for the application of online learning to non-linear space [31–35].

2.3. Kernel adaptation of PC-KFA

Both offline and online data are used for a data-dependent composite kernel. This extended framework has been developed, called Principal Composite Kernel Feature Analysis (PC-KFA) [1]. Because we have computed the Gram matrices for p most dominant kernels for the composite data (offline plus online data), we now combine them to yield the best classification performance. As defined in Section 2.2, we can write the composite kernel for the new composite data:

$$K_{comp}^{n'}(\rho) = \sum_{i=1}^{p} \rho_i' Q_i^{n'} P_i^{n'} Q_i^{n'} \tag{5}$$

We determine the optimum composite coefficients that will yield the best composite kernel. We can determine the composite coefficient ρ' that will maximize the appropriate criterion. Once we know the eigenvectors, i.e. combination coefficients of the composite training data (offline data + online data), we can compute q'_r and hence Q'_r to find out the 4 g matrices corresponding to the 4 different kernels. Now the first p kernels corresponding to the first p eigenvalues (arranged in descending order) will be used in the construction of a composite kernel that will yield optimum classification accuracy. Before proceeding to the construction of the composite kernel, we have to determine whether our new data are homogeneous or heterogeneous and update our Gram matrices accordingly.

2.4. Online machine learning methods used for CAD of CTC

Due to a high demand of computer-aided diagnosis in medical images, both offline and online training algorithms are required for the improvement of the performance of CAD systems. This is particularly true in high-dimensional medical-image-based diagnosis, in which specialists' opinion for final diagnosis is required. Due to the increasing need of safe and easy-to-tolerate method for screening of colorectal cancer, virtual colonoscopy, *a.k.a.* CT colonography (CTC), is becoming a popular non-invasive diagnostic procedure. Properly trained machine learning methods can diagnose a suspicious CTC image and save cost [1,2,4]. The amount of new CTC data added to the repository every day is significantly increasing with advent of new technology [10,13,56]. Online learning is one of the most effective methods for incrementally improving the performance of a classifier by use of new samples available from the confidential repository of CAD cases collected in hospitals throughout the world [18,19,48,49,66]. Researchers in both medical and computing community are working on the improvement of online learning of CAD for CTC and other imaging modalities. The proposed algorithm aims to improve the aforementioned PC-KFA, which showed high performance in the detection of polyps in CTC

images with offline training [1], by adding an online training method for big heterogeneous data.

3. HDA: heterogeneous vs. homogeneous quantification

To make CTC a viable option for screening of larger patient populations as time passes, we propose HDA to handle online mass data of clinical colon screening. In Section 3.1, the proposed Class Separability is described for determining whether heterogeneous data are acquired. In Section 3.2, another criterion of data heterogeneous degree is described for PC-KFA.

3.1. Quantification of class separability for heterogeneous data

We use class separability as a measure to identify whether the data are either heterogeneous or homogeneous. If the data are homogeneous, separability is high. However, the heterogeneous data degrades the class separability. For restoring/improving the class separability, PC-KFA needs feature space adjustment. Let us introduce a variable ξ which is the ratio of the class separability of the composite online data and the offline data. It can be expressed as: $\xi = J'_*(\alpha'_r)/J_*(\alpha_r)$, where $J'_*(\alpha'_r)$ denotes the class separability yielded by the most dominant kernel for the composite data (i.e. new incoming data and the previous offline data). $J_*(\alpha_r)$ is the class separability yielded by the most dominant kernel for offline data. Thus separability can be rewritten as: $\xi = \lambda'_*/\lambda_*$, where λ'_* corresponds to the most dominant eigenvalue of composite data (both offline and online), and λ_* is the most dominant eigenvalue of the 4 different kernels for the offline data. If ξ is less than a threshold value η (e.g. one), then the incoming online data are heterogeneous; otherwise, they are homogeneous. In the case of homogeneous data (identified by ξ), we do not update the Gram matrix. Instead we discard all of the new data that is homogeneous. Hence, the updated Gram matrix can be given as, $K_r^{n'} = Q_r P_r Q_r$. Here $K_r^{n'}$ denotes the updated value of $K_r^n (= Q_r P_r Q_r)$.

An advantage of the separability measure is that it can reduce the size of gram matrix that determines the computational complexity. A disadvantage of the separability measure is that it may be prone to binary-classification problems. If the incoming data is homogeneous, there is not much useful information that can be used for improving the classification performance. However, the magnitude of the separability may be useful for measuring the confidence of the classification.

3.2. Heterogeneous degree

If the new sequence of incoming data is heterogeneous, we update our feature space. However, this update can be either incremental or non-incremental. We propose a criterion called "Heterogeneous Degree", $\xi' = (\bar{\alpha}' - \bar{\alpha})\lambda'/\lambda$, to determine if the data are highly heterogeneous or less heterogeneous. Let us denote the mean of combination coefficients $\bar{\alpha}$ for the most dominant kernel among the 4 kernels available. The class separability of the most dominant kernel for the "new" data denoted as "accent" is directly dependent on both the combination coefficient "$\bar{\alpha}'$" (from the optimum combination coefficient of 4 different kernels) as well as the maximum eigenvalue(s) λ'. Fig. 2 shows the relationship between separability and the heterogeneous degree.

If the heterogeneous degree ξ' is less than 1 (Case1), then the update is non-incremental. Hence, the dimensions of the Gram matrix remain unchanged. The input matrices P' and Q' are updated by replacing the row and column corresponding to the index. Fig. 3 illustrates the overall flow of online data training.

If the heterogeneous degree ξ' is greater than 1 and less than η (Case 2), this means that the difference between the previous eigenvectors and the new eigenvectors is large. Thus, it is very important to retain these highly heterogeneous new data to improve the class separability of the proposed algorithm. We simply retain them, and hence the size of the Gram matrix is incremented by the size of the new data. In this case, the input matrices P' and Q' are same as in Section 2.2.

Obtaining a higher-dimensional feature space of the training data with greater classification power depends on how effectively we are updating the Gram matrix. Because construction of Gram matrix is the crucial step for performing feature analysis, any redundant data (homogeneous data) in the Gram matrix are unnecessary and can be discarded. At the same time, data of a different nature (heterogeneous data) should be preserved for further training of the algorithm. The proposed method deals with the measures of heterogeneous degree and separability together for improvement of the performance. These two measures are correlated with kernel-alignment factor, which is a defining metric of underlying composite feature analysis.

Once we have our kernel Gram matrices for Cases 1 and 2, we can now determine the composite kernel that will give us the optimum classification accuracy when the existing offline data are incorporated with new online data. Because we have computed the Gram matrices for p most dominant kernels for the composite data (offline plus online data), we now combine them to yield the best classification performance. As defined in Section 2.2, we can write the composite kernel for the new composite data as shown in (5), $K_{comp}^{n'}(\rho) = \sum_{i=1}^{p} \rho_i' Q_i^{n'} P_i^{n'} Q_i^{n'}$.

Online PC-KFA incorporates HDA when the training properties of the data change dynamically. This update procedure could be much simpler in the presence of a unified kernel [52] for the existing class of data. Extracting the optimum combination of kernels for composing a complex kernel is challenging [53]. The real CTC dataset suffers from the non-linear and diverse distributions among actual cancer datasets used for CAD [23], especially when the size of the datasets increases. By extending KPCA and kernel selection to address this data obstacle, we introduce an adaptive kernel algorithm called PC-KFA [1] with a composite kernel function that is defined as the weighted sum of the set of different optimized kernel functions for the training datasets as follows: $K_{comp}(\rho) = \sum_{i=1}^{p} \rho_i Q_i P_i Q_i$ where the value of the composite coefficient ρ_i is a scalar value, and p is the number of kernels we intend to combine in the training datasets. Through this approach, the relative contribution of all of the kernels to the model can be varied over the input space when the new datasets are added. Instead of using K_r as the old kernel matrix, we will use $K_{comp}(\rho)$, which we call "kernel adaptation". According to [57], this composite kernel matrix $K_{comp}(\rho)$ satisfies Mercer's condition. Now the problem is how to determine this composite coefficient $\hat{\rho}(=[\rho_1, \rho_2, \ldots \rho_p])$ such that the classification performance is optimized. To this end, we use the concept of a "Kernel Factor" to determine the best $\hat{\rho}$ that gives the optimum performance for both offline and online datasets.

It is efficient to perform online learning if each pattern is presented in the limited storage space, thus PC-KFA with the use of HDA requires little to no additional memory for storing of the patterns in the data, for improving the data compatibility over a long period.

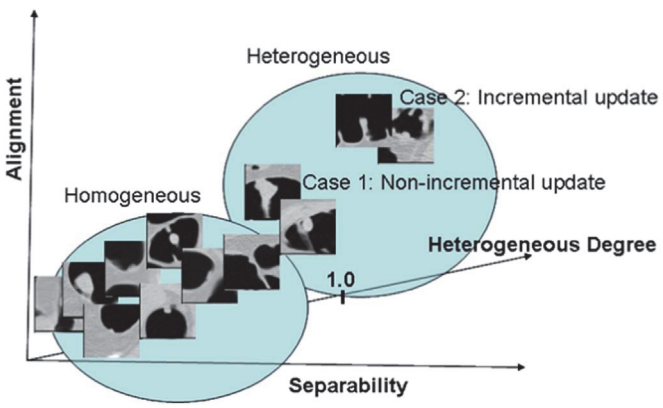

Fig. 2. Criteria concept for homogeneous and heterogeneous online data.

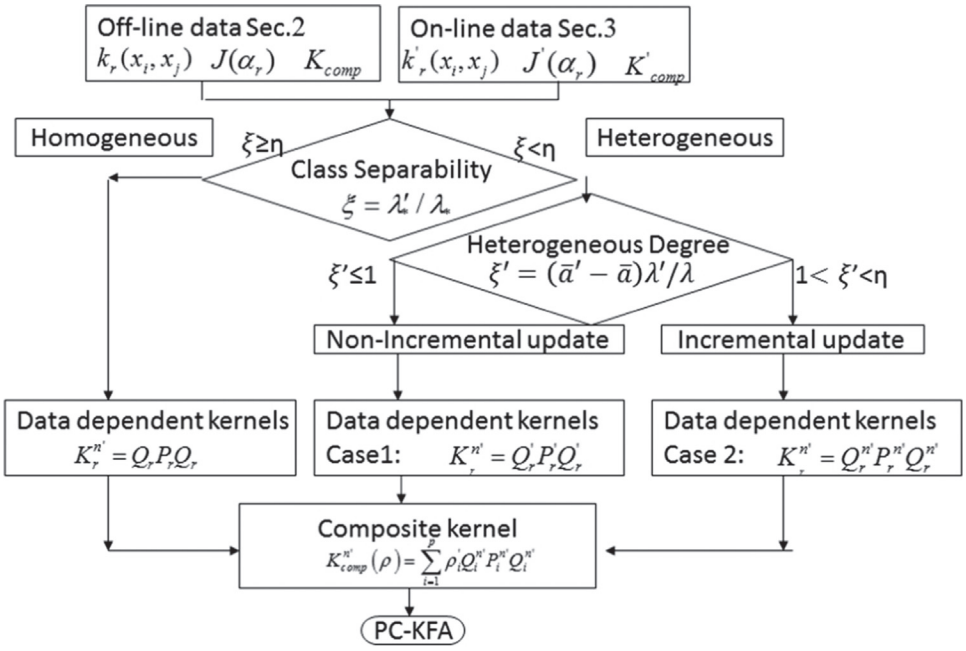

Fig. 3. Overall flow of online data training using HDA with PC-KFA.

4. Long-term sequential trajectories with re-evaluation

Data validation is described when CTC data are acquired "over a long period of time". When the proportion of online datasets is dominant for the CAD system, CTC data can sometimes become highly diverse. Section 4.1 describes the Alignment Factor for Auditing Large Online Data, Section 4.2 describes Validation by Decomposing Online Data into Small Subsets, and Section 4.3 describes Detecting Major Changes from Longitudinal Studies.

4.1. Alignment factor for auditing large online data

The "alignment" measure was introduced for measuring the adaptability of a kernel to the target data, and for providing a practical objective for kernel optimization. The "alignment" measure, called Alignment Factor (AF), is defined as a normalized Frobenius inner product between the kernel matrix and the target label matrix. The empirical alignment between kernel k_1 and kernel k_2 with respect to the scatter matrix of the training set S is given as: $Frob(k_1, k_2) = \langle K_1, K_2\rangle_F / \|K_1\|_F \|K_2\|_F$.

where, K_i is the kernel matrix for the training data using kernel function k_i.$\langle K_i, K_j\rangle_F$ is the Frobenius inner product between K_i and K_j, and $\|K_i\|_F = \sqrt{\langle K_i, K_i\rangle_F}$.

It has been shown that, if a chosen kernel is well aligned with the other datasets, it does not change anything. If there exists a separation of the data with a low bound on the generalization error [54], it would be better to add one more kernel so that we can optimize the kernel alignment based on training both offline and online dataset to improve the generalization performance on further test datasets. Let us consider the combination of kernel functions corresponding to [24] as: $k(\rho) = \sum_{i=1}^{p} \rho_i k_i$— where the kernels, k_i (i=1, 2, ..., p), are known in advance. Our purpose is to tune ρ to maximize the empirical alignment between $k(\rho)$ and the target vector y. Hence,

$$\hat{\rho} = \arg_\rho \max(Frob(\rho, k_i, k_j)) = \arg\max_\rho \left(\frac{\langle \sum_i \rho_i K_i, K_j \rangle}{\sqrt{\langle \sum_i \rho_i K_i, \left\langle \sum_j \rho_j K_j \right\rangle}} \right) = \arg\max_\rho \left(\frac{\rho^T V_{ij} \rho}{\rho^T U_{ij} \rho} \right) \quad (6)$$

where $u_i = \sqrt{\langle K_i, yy^T\rangle}$, $v_{ij} = \sqrt{\langle K_i, K_j\rangle}$, $U_{ij} = u_i u_j$, $V_{ij} = v_i v_j$,

$$\rho = \left(\sqrt{\rho_1}, \sqrt{\rho_2}, \ldots, \sqrt{\rho_p} \right)$$

Let the generalized Raleigh coefficient be given as:

$$J(\rho) = \rho^T V \rho / \rho^T U \rho \quad (7)$$

Therefore, we can obtain $\hat{\rho}$ by solving the generalized eigenvalue problem $V\rho = \delta U\rho$ where δ denotes the corresponding eigenvalue. Once, we find this optimum composite coefficient $\hat{\rho}$, which will be the eigenvector corresponding to maximum eigenvalueδ, we can compute the composite data-dependent kernel matrix $K_{comp}(\rho)$.

We use the AF method to determine the optimum composite coefficients that will yield the best composite kernel. AF can be calculated as in Eq. (6). We can determine the composite coefficient ρ' that will maximize the AF value as in Eq. (7) of the new generalized Raleigh coefficient. We can obtain the value of $\hat{\rho}'$, by solving the generalized eigenvalue problem $U^{n'}\rho' = \delta' V^{n'}\rho'$, where δ' denotes the eigenvalues. We choose a composite coefficient vector that is associated with the maximum eigenvalue. Once we determine the best composite coefficients, we can compute the composite kernel.

If the deviation of AF is huge, then the step size is reduced further and the algorithm is computed again. This process is repeated until we find an appropriate window size of the incoming data that allows the proper classification of the homogeneous and heterogeneous data and training results with reduced error. After the training of the Gram matrix that incorporates the dynamic features of the new online data is finished, the PC-KFA algorithm is applied to the kernel Gram matrix. This entire process is summarized in the flow chart in Fig. 4.

The long sequential online datasets to the other existing online datasets are tracked using AF for the kernel adaptability. We extend AF of the normalized Frobenius inner product with the time sequential index. If there is no compromise when we train the algorithm without further modifications, then there is no need to break down the incoming sequence of data into small windows or to change any parameters in the previous setting. Let us denote the time index ${}^{t+1}(K_{comp}^{n'})_t$ as the update of Gram matrix from the time index t to $t+1$, and let ${}^{t}(K_{comp}^{n'})_0$ be the updated the Gram matrix from time '0' (i.e. from the beginning) to the time index t. The AF can be given as ${}^{t+1}(Frob({}^{t+1}(K_{comp}^{n'})_t, {}^{t}(K_{comp}^{n'})_0)_t$ and ${}^{t}(Frob({}^{t+1}(K_{comp}^{n'})_t, {}^{t}(K_{comp}^{n'})_0)_0$. In or-

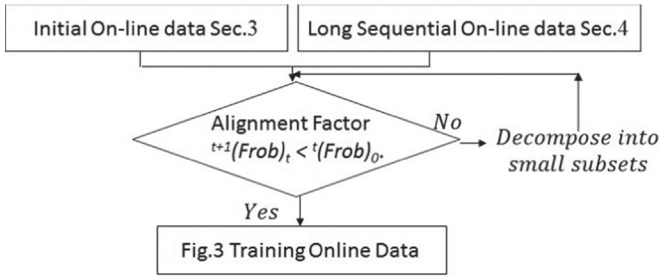

Fig. 4. Training of online datasets acquired over long-term sequences using Alignment Factor.

der to avoid unnecessary complexity in notations, we simplify the Gram matrix notation $^{t+1}(K^{n'}_{comp})_t$ into the term $^{t+1}(Frob)_t$. The proposed criterion is to audit the erroneous training of the data by comparing these AFs by checking the following conditions: $^{t+1}(Frob)_t < {}^{t}(Frob)_0$. Based on the comparison, the data are either decomposed or sent to the Fig. 3 process in Section 3. The next subsection shows how the online data is decomposed.

4.2. Validation by decomposing online data into small subsets

CTC data acquired "over a long period of time" can sometimes be highly diverse, i.e. the portion of online datasets is dominant for the CAD system. None of the existing methods are known to efficiently validate large online CTC data [55]. To facilitate online training of long-term sequential trajectories of CTC datasets, we validate whether retention is required for a portion of the training data due to the size of long sequential datasets. Thus, the challenge is to determine what data to retain. Hence, we present an auditing way of validating the incoming long sequential data as to either decompose or not, and automatically updating the feature space over time accordingly.

Previously, we divided the new online data into equal-sized sub-datasets, which we called *d*1, *d*2, *d*3 etc., where the size of each new sub-dataset was equal to 1. We extend this restriction for the long sequential datasets as heterogeneous or homogenous data by auditing the class separability of each dataset to find the appropriate window size for the incoming data and determine whether it is homogeneous or heterogeneous data by using the updated criteria. Fig. 5 illustrates the segmentation of the incoming online data into small equal subsets if the incoming online data are heterogeneous. If the data are homogeneous, most of the data are redundant; thus, we simply discard these homogeneous data, and we do not update any information from the new subset on to the previous Gram matrix. Thus, the kernel Gram matrix at time t+1 is the same as the previous one at time t. On the contrary, if the data are heterogeneous, we update the Gram matrix either incrementally or non-incrementally depending on the level of heterogeneous degree of the online data.

The modified criterion for validating the degree of heterogeneity is time-sequential class separability. The time-sequential class separability decomposes the online data into small subsets (data up to time t and d1 data) by using:

$$^{t+1}(\xi)_t = {}^{t+1}(\lambda'_*)_t / {}^{t}(\lambda_*)_0 \tag{8}$$

where, $^{t+1}(\lambda'_*)_t$ is the largest eigenvalue of the data (of the dominant kernel) received from time t to t+1, and $^{t}(\lambda_*)_0$ is the largest eigenvalue of the dominant kernel calculated from time *0* to time t. If $^{t+1}(\xi)_t$ is greater than a new threshold value η', then the incoming data are heterogeneous; otherwise, they are homogeneous. The threshold value η defines the boundary condition between homogenous data and heterogeneous data as well as type of updates of the Gram matrix. The magnitude of η is both data and feature dependent, i.e., too small η may prohibit any kinds of updates of the Gram matrix, and too high η may allow the Gram matrix to grow beyond acceptable size.

4.3. Detecting major changes from longitudinal studies

Major changes from longitudinal studies are analyzed by introducing another data compatibility measure, called Gap-in-data. The Gap-in-data measures indicate whether data incompatibility for online time sequential data sets may occur or not. If the Gap-in-data is high, these data should be detected as a major change and reevaluated. During the online training of long-term sequential trajectories, Heterogeneous Degree ξ' in Section 3 brought two cases based on threshold values such as η. We reevaluated this data measure ξ', by adjusting the threshold value η, to detect the major change shown as Case 3 in Fig. 6.

By experimentally adjusting the threshold value η, the heterogeneous online data may be excluded because they contain too large Gap-in-data to be compatible. In this case, kernel computation using such online data is excluded. For determining the threshold value η is relevant to Mean Square Error (MSE) values, defined as $MSE = 1/t \sum \lambda_i$ [1]. All MSE values for heterogeneous datasets are re-evaluated for detecting Gap-in-data. The threshold value η is determined empirically.

4.4. Data classification measure based on TP/FP accuracy

Accuracy is also used as a statistical measure of how well a binary classification test correctly identifies that our medical images have sufficient data quality. The accuracy is defined as the proportion of True Positives (TP) and True Negatives (TN) in the entire database. False Positive (FP) and False Negative (FN) are the counterpart of the binary classification. TP is the case when CAD detects cancer in CTC successfully, and TN is the case when CAD detects the absence of cancer in CTC correctly. FP is the case when CAD incorrectly detects cancer in CTC, and FN is the case when CAD does not detect cancer, even though it actually exists. These are used to measure the effectiveness with *True Positive Ratio (TPR)* and *False Positive Ratio (FPR)*. *TPR* is treated as CAD hit rate, and *FPR* is CAD fall-out:

$$TPR = \frac{TP}{TP + FN}, \quad FPR = \frac{FP}{FP + TN}.$$

The classification accuracy and Mean Square Reconstruction Error (MSRE) for the datasets are analyzed in Section 5. The classification accuracy was calculated as

$$Accuracy = \frac{TP + TN}{TP + TN + FN + FP}.$$

The MSRE in PC-KFA algorithm is calculated as the normalized difference between the original feature space and the kernel mapped feature space. Another data classification measure, the Receiver Operating Characteristic Curve (ROC), is used for comparing FPR versus TPR. The Area Under the Curve (AUC) is calculated as the region under the ROC curve. The F-score expresses the effectiveness of a binary classification test from the perspective of the class of interest, which is typically useful for class imbalanced datasets. As shown in Section 5.1, the numbers of samples from two classes are highly

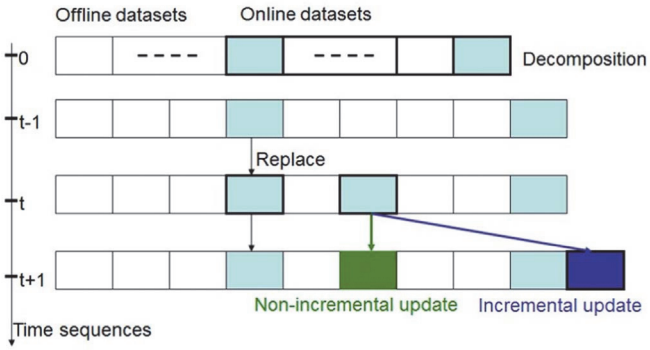

Fig. 5. Online decomposition for heterogeneous sequences.

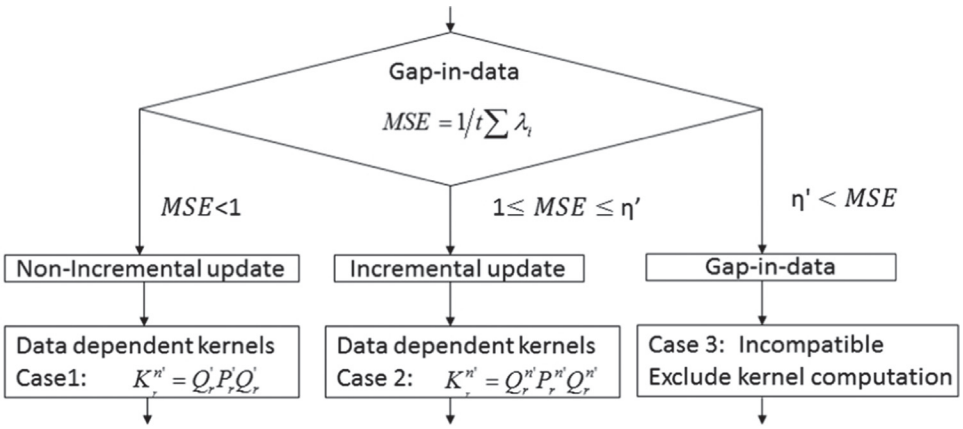

Fig. 6. Detection of major change using heterogeneous degree.

imbalanced. In this context, the F-score is defined as,

$$F\text{-}score = \frac{2*precision*recall}{precision + recall},$$

where $precision = \frac{TP}{TP+FP}$ and $recall = \frac{TP}{TP+FN}$. The F-score is considered as a more accurate evaluation of a binary classifier dealing with class imbalanced data. Section 5 shows the CTC data classification with use of all of these measures, including the computation time for PC-KFA using the proposed HDA.

5. Experimental results of data classification

The experimental results are organized as follows: In Section 5.1., cancer datasets from online medical images are described. In Section 5.2., the offline classification performance for PC-KFA is presented. In Section 5.3., data decomposition for the new online sequences is shown. In Section 5.4., a quantification of heterogeneous versus homogeneous data is presented. In Section 5.5., data validation of long-term sequence is described. Finally, the evaluation of computational time is presented in Section 5.6.

5.1. Cancer datasets from online medical images

The proposed HDA with PC-KFA was evaluated using CTC image datasets of colonic polyps comprised of TP and FP polyps detected by our existing CAD system [21,56,61]. We used CTC cases of 146 patients who underwent a colon-cleansing regimen which was the same as that of optical colonoscopy. These CTC cases were acquired by a total of eight different models of single- and multi-detector CT scanners by use of 1.25–5.0 mm collimations, a pitch of 1–2, reconstruction interval of 1.25–5.0 mm, and tube current of 50–200 mA. The patients were followed by conventional optical colonoscopy, which served as the gold standard for the presence of polyps. Among 464 CTC cases, 59 cases were abnormal (having at least one polyp ≥6 mm in size), and 405 cases were normal (having no colonic polyps). Each patient was scanned in both supine and prone positions; thus each CTC case consisted of two reconstructed CTC volumes, resulting in 928 CTC volumes with an effective voxel size of 0.5 mm. The volumes of interest (VOIs) representing each polyp candidate have been calculated as follows. First, the CAD scheme provided a segmented region for each candidate, and the center of the VOI was placed at the center of mass of the region. The size of the VOI was chosen so that the entire segmented region was covered. The resampling was carried out to generate VOIs with dimensions of 12×12×12 voxels for building Dataset1, which consisted of 29 true polyps and 101 FPs. For the rest of the datasets, the VOI was resampled to 16×16×16 voxels. The VOIs thus computed comprised Dataset2, which consisted of 54 TPs and 660 FPs, Dataset3, which consisted of 16 TPs and 925 FPs, and Dataset4, which consisted of 11 TPs and 2250 FPs.

Table 1 shows the distribution of TPs and FPs in the offline training and test sets of the colon polyp datasets. Instead of using cross validation, we randomly divided the entire dataset into training and test sets because they were highly imbalanced, i.e., the majority of datasets were FPs, whereas few were TPs. We used these training and test datasets for all of our experiments.

5.2. Offline classification performance for PC-KFA

We used the KPCA method described in Section 2 to create four different data-dependent kernels and selected the kernel that best fitted the offline CTC data listed in Table 1. We determined the optimum kernel depending on the eigenvalue that produces maximum separability. Table 2 indicates the eigenvalues λ and hyper-parameters of four kernels for each dataset.

The kernel with the maximum eigenvalue are bold faced for each offline dataset in Table 2. In Dataset1, for example, we combined RBF and Laplace to form a composite kernel. We observed that different composite kernels yielded the largest eigenvalue for different databases. The composite coefficients of the two most dominant kernels are listed in the last column. For the four datasets, the most dominant kernel was the RBF kernel, whereas the second most dominant kernel varied substantially and thus was undetermined.

The MSRE, accuracy and F-score for performance on offline test datasets are shown in Table 3. Fig. 7 shows ROC curves for FPR versus

Table 1
True Positive and False Positive Distributions in Offline Datasets.

Data	No. of Vectors in Training Set			$\frac{\#TP}{Total}$ (%)	No. of Vectors in Test Set			$\frac{\#TP}{Total}$ (%)
	TP	FP	Total		TP	FP	Total	
Dataset1	21	69	90	23.33	8	32	40	20.00
Dataset2	38	360	398	9.55	16	300	316	5.06
Dataset3	10	500	510	1.96	6	425	431	1.39
Dataset4	7	1050	1057	0.66	4	1200	1204	0.33

Table 2
Eigenvalues of 4 kernels for offline datasets.

Data	Linear	Poly	RBF	Laplace	Combination
Dataset1	λ=10.66	λ=10.25 *d*=.2, *Offset*=2	λ= **14.13** σ =4.12	λ=**12.41** σ =0.9	*RBF* 0.98 *Laplace* 0.14
Dataset2	λ=102.08	λ=**105.91** *d* =1, *Offset*=4	λ=**116.64** σ=5.29	λ=80.57 σ =3.5	*RBF* 0.72 *Linear* 0.25
Dataset3	λ=**57.65**	λ=51.35 *d* =1.4, *Offset*=1	λ=**74.55** σ =5.65	λ=30.23 σ =1.0	*RBF* 0.98 *Linear* 0.23
Dataset4	λ=72.41	λ=**83.53** *d* =0.8, *Offset*=2	λ=**124.13** σ =4	λ=56.35 σ =2.5	*RBF* 0.91 *Poly* 0.18

Table 3
MSRE, accuracy and F-score of offline data classification.

Data	MSRE of offline data with Composite Kernel (%)	Classification accuracy of offline data with Composite Kernel (%)	F-score of offline data with Composite Kernel (%)
Dataset1	1.0	90	71.38
Dataset2	9.64	94.62	58.27
Dataset3	6.25	98.61	15.56
Dataset4	14.03	99.67	66.94

TPR. From Table 3 and Fig. 7, we obtained a good classification accuracy and ROC performance by using PC-KFA over all the offline datasets. However, the F-scores are much lower than accuracy and hence further improvement of the classifier is recommended. This justifies the motivation for online learning as mentioned in Section 1. All of the 4 CTC image datasets of colonic polyps were well detected by using PC-KFA [1], and the largest Dataset4, for example, achieved 99.67% classification accuracy. In the next sections, we further evaluated the results of this method based on an online HDA that included heterogeneous large datasets of CTC images.

5.3. Data decomposition for the new online sequences

We followed Section 3 to tune the selection of appropriate kernels when new online data became available. Table 4 shows such a new online data stream called "Online Sequence ##". Each of the data sequences originates from the same original cumulated database. However, they differed in the fraction of the initial training sets. Dataset 1–4 corresponded to 0.5%, 5%, 10% and 20% of the main

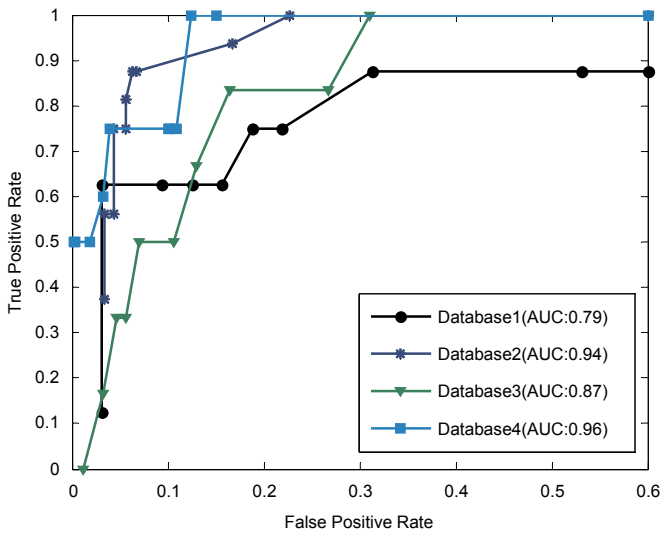

Fig. 7. The ROC curves and AUC values for offline data using PC-KFA.

Tablee 4
Size of online data at different time sequences.

Data	Online Sequence #1	Online Sequence #2	Online Sequence #3	Online Sequence #4
Dataset1	3 TP, 12FP	3 TP, 12FP	3 TP, 12FP	3 TP, 12FP
Dataset2	7 TP, 85 FP	7 TP,85 FP	7 TP, 85 FP	7 TP, 85 FP
Dataset3	2 TP, 87 FP	2 TP, 87 FP	2 TP, 87 FP	2 TP, 87 FP
Dataset4	2 TP,126 FP	2 TP,126 FP	2 TP,126 FP	2 TP,126 FP

database used for initially labeled training. Each dataset was used multiple times in a different order. The term "online data sequence" was defined in Section 4.2, as shown in Fig. 8. This figure explains the relationship of the dataset and data sequence corresponding to Table 4.

After we tentatively formed the input matrices for 4 different kernels, we used Eq. (6) to find the dominant kernels for the new data and the previous offline data. These results are summarized in Table 5.

Table 5 shows, using online data sequences, the dominant kernel with the bold-face λ as the largest eigenvalue. As seen in the online data sequences, the eigenvalues calculated were gradually shifting, but choice of the dominant kernels remained the same. Therefore PC-KFA consistently maintained the choice of dominant (and second dominant) kernels for updating data-dependent kernel matrices for the computation of the composite kernel matrix. Under the screening of large patient populations, due to the kernel representation, the detection of colonic polyps was expected its stabilities by associating existing data with newly acquired online sequences. We will evaluate more HDA characteristics of long-term online data sequences in the next subsection 5.5.

5.4. HDA: quantification of heterogeneous versus homogeneous data

After determining the two dominant kernels, the next step was to update these data-dependent kernel matrices for the computation of the composite kernel matrix. We imported the criterion in Section 3.2 to classify these new data into the 2 categories, "Homogeneous" and "Heterogeneous."

Table 6 shows that the majority of online data sequences was homogeneous: the Gram matrix in this sequence was not updated to save computation time. If the database was homogenous, we set the data-dependent kernels as "No update". However the HD determines Case 1 as "Non-Incremental Update," or Case 2 as "Incremental Update," for all the heterogeneous data sequences in Section 3.2. Once we have the data-dependent kernel Gram matrices for 4 different kernels, we can proceed to calculate the data-dependent composite kernels for the new online data sequences based on the method presented in Section 2.3.

Table 7 shows the composite kernel coefficients for the two most dominant kernels. The metric MSE corresponds to the MSRE of the

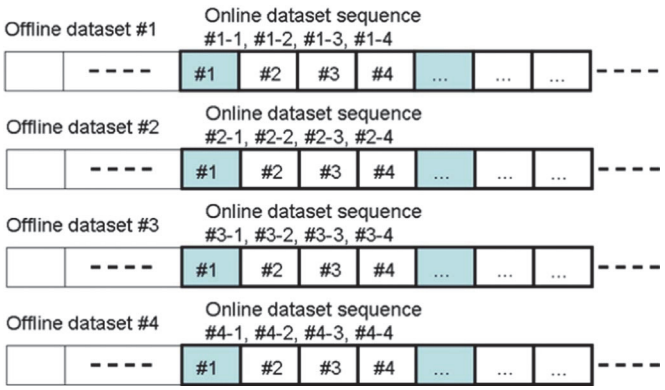

Fig. 8. The online dataset sequences of Table 4.

Table 5
Eigenvalues of 4 different Kernels (Linear, Poly, RBF, Laplace) with λ for Online Data Sequences.

Data		Online Sequence #1	Online Sequence #2	Online Sequence #3	Online Sequence #4
Dataset1	Linear	9.99**	9.89	10.16	9.73
	Poly	9.31**	8.96	9.52	10.31
	RBF	**11.34****	**12.47**	**18.09**	**24.69**
	Laplace	8.41**	6.7	12.85	13.27
Dataset2	Linear	129.44	84.97**	95.63**	109.02
	Poly	106.44	85.01**	96.27**	109.61
	RBF	**134.59**	**101.21****	**122.81****	**133.40**
	Laplace	90.76	38.49**	84.12**	115.78
Dataset3	Linear	75.13	73.88	62.59**	72.45
	Poly	36.38	69.25	66.48**	83.39
	RBF	**80.16**	**86.41**	**73.66****	**109.29**
	Laplace	44.66	13.87	49.46**	50.25
Dataset4	Linear	82.72	52.99**	69.43	98.64
	Poly	92.33	75.83**	79.85	112.39
	RBF	**132.47**	**103.25****	**144.59**	**157.27**
	Laplace	38.67	43.26**	32.59	62.85

** Indicates the eigenvalues after updating of the matrix according to the method presented in Section 4.

Table 7
Classification accuracy and MSE with composite kernels of online data sequences.

Data	Online Sequence #1	Online Sequence #2	Online Sequence #3	Online Sequence #4
Dataset1	MSE:0.43 Accuracy:92.5 RBF:ρ_1=0.99 Linear: ρ_2=0.17	*sab*	*sab*	*sab*
Dataset2	*sab*	**MSE:15.06** Accuracy:93.99 RBF:ρ_1=0.93 Linear:ρ_2=0.16	**MSE:14.50** Accuracy:94.94 RBF:ρ_1=0.99 Linear:ρ_2=0.14	*sab*
Dataset3	*sab*	*sab*	MSE:7.68 Accuracy:96.52 RBF:ρ_1=0.73 Poly: ρ_2=0.32	*sab*
Dataset4	*sab*	**MSE:15.63** Accuracy:97.65 RBF:ρ1=0.90 Poly:ρ2=0.28	*sab*	*sab*

**"*sab*" indicates "Same as before"

online data sequences, which were calculated in Section 5. A smaller MSRE signifies higher data reconstruction ability of PC-KFA. Table 7 also shows that the classification accuracy (how accurately the data were determined) for the online data sequence was, on average, 95.12% with variance 2.03%. This was comparable to the offline data performance that yielded the classification accuracy of 95.72% with variance 4.39%. The results for all of the online data sequences were re-evaluated using MSE as described in Section 4.3 for determining whether Gap-in-data exists in the online sequences shown in a bold font in Table 7. The large MSE values were shown in Dataset2 online sequences #2, #3, and in Dataset4 online sequence #2, indicating that auditing may have detected major changes. The next subsection shows the re-evaluation results for a long-term sequential trajectory.

5.5. Data validation of long-term sequence

In this subsection, long-term sequential data were used for auditing and validating online datasets. We performed online HDA with PC-KFA for long-term sequential data that were much larger than those of the online sequences analyzed in previous subsections. Fig. 9 shows that data distribution of long-term sequential trajectories as an evaluation of long-term online learning.

For the long-term sequential trajectory, the experimental dataset was comprised of all of the available images that we accumulated. A preset percentage of images was selected randomly from the dataset and used for the offline training stage by use of PC-KFA. The rest of the data were sequentially fed into the online HDA algorithm for evaluation of the long-term sequential trajectory. The different ratios of training data between online and offline were prepared as shown in Fig. 9 from 0 to 9; for example, '5' means that online data were 5 times larger than offline datasets. Note that we had three different cases of offline data; thus, "Large Online Data" was large online data compared to relatively small offline data. We analyzed how these ratios affect the online learning performance for the AF, AUC, and classification accuracy.

The AF was calculated using several long-term online sequences. If there was no increase, we divided this online sequence of data into small subsets by using the threshold value of online HDA with PC-KFA.

The results shown in Fig. 10 indicate that the AF was increased when more data were used for training of the online HDA with PC-KFA. As the ratio of the online to offline size increased, the AF increased for all three small, medium, and large online data. Albeit more diverse data were fed into the form of online sequences from larger patients, the proposed online HDA with PC-KFA adapted itself, as shown in the increase of the AF. Fig. 10 shows that the AF consistently increased for long-term sequences of "large/medium/small" online data using online HDA with PC-KFA even though heterogeneous characteristics were affected as in the cases shown in Tables 7, 8.

The online HDA with PC-KFA was evaluated based on the AUC for the online long-term sequences. Fig. 11 shows that online PC-KFA yielded an AUC performance similar to that of Fig. 7 for classification in all three long-term sequences. The ability to track changes by using long-term online sequences was also verified by the results shown in Fig. 12.

Fig. 12 shows that the proposed online HDA with PC-KFA, which handles very large online data over long-term sequences, performed with a classification accuracy similar to that of the offline counterpart shown in Table 3 in Section 5.2. After a finite number of sequences, the classification performance of the online data sets approached to that of the offline training data. This indicates that training of the subsequent

Table 6
Homogeneous and Heterogeneous Categories of Online Sequences with the data updates.

Data	Online Sequence #1	Online Sequence #2	Online Sequence #3	Online Sequence #4
Dataset1	**Heterogeneous** Non-incremental	Homogeneous	Homogeneous	Homogeneous
Dataset2	Homogeneous	**Heterogeneous** Non-incremental	**Heterogeneous** Non-incremental	Homogeneous
Dataset3	Homogeneous	Homogeneous	**Heterogeneous** Non-incremental	Homogeneous
Dataset4	Homogeneous	**Heterogeneous** Non-incremental	Homogeneous	Homogeneous

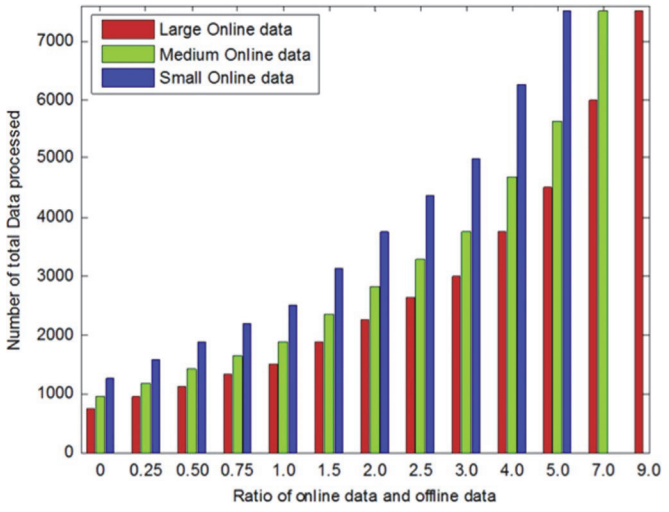

Fig. 9. Long-term data sequence used for evaluation of online learning. The horizontal axis denotes the ratio of number of online data to the number of offline data. The vertical axis denotes the number of total data processed for training using online HDA with PC-KFA. The three long sequences are labeled as 'Large Online Data', 'Medium Online Data', and 'Small Online Data', which were used corresponding to offline training dataset sizes of 750, 937, and 1250, respectively.

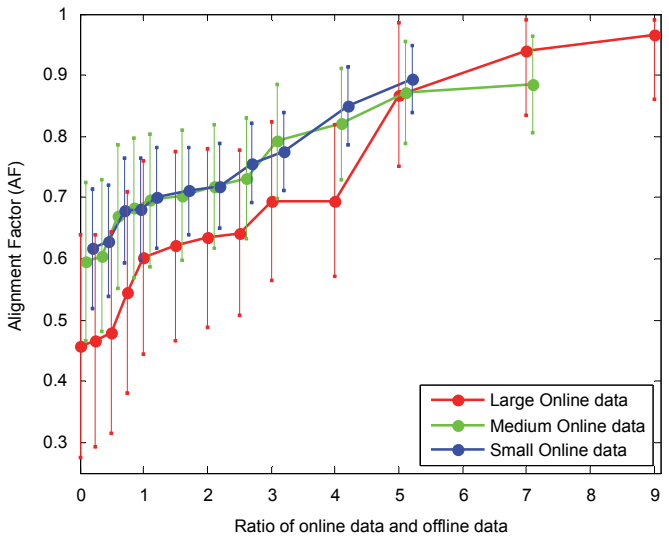

Fig. 10. Alignment Factors (AF) for long-term sequences. The horizontal axis denotes the progress of the online HDA (PC-KFA) with time (ratio of online to offline training). The solid lines denote the mean of the observed AF value, and the dashed lines show the range of observed AF.

online data sequence was advantageous over statistical offline learning. Due to the reduced size of the kernel space, even though the data size increased, the proposed online HDA with PC-KFA achieved a consistent performance in classification accuracy.

The data auditing by validating the long-term sequence was evaluated by means of AF, AUC, and accuracy (shown in Figs. 10–12) to show the degree of uniformity between offline and online datasets. The

Table 8
Online data sequence computation time.

Online data	Mean offline training time (millisecond/sample)	Mean online training time (millisecond/sample)
Large	884.34	85.95
Medium	936.65	59.99
Small	976.16	52.51

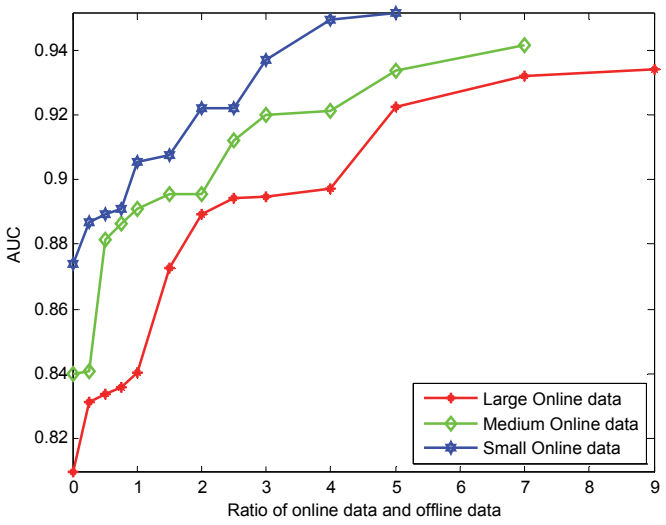

Fig. 11. AUCs of ROC curves for three long-term sequences. The horizontal axis denotes the progress of the online HDA (PC-KFA) with time (ratio of online to offline training).

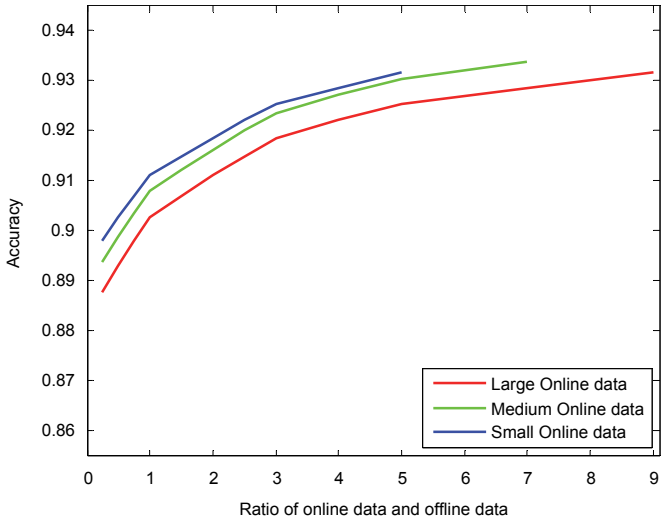

Fig. 12. Accuracy versus ratio of online data to offline data.

ratio of online to offline datasets are shown by three representative colored plots as small, medium, and large online data. In general, small online data performed with a higher accuracy and consistency in data-contexts, compared to large online data. In datasets pooled from more online data locales, if the ratio of online data is increased, these auditing measures AF, AUC, and accuracy were all increased as if these online and offline datasets were converted into a single merged dataset.

5.6. Evaluation of computation time

Finally, we analyzed the time required for the processing of HDA by using online PC-KFA. All of the experiments were performed by MATLAB R2010a using the Statistical Pattern Recognition Toolbox for the Gram matrix calculation and kernel projection. For processing of the large volume of data, an Intel® Core i7 with 3.40 GHz clock speed was used along with a workstation containing 16 GB system memory. The run time was determined by using the *cputime* command. Because we do not consider all available online data after the initial offline training, the proposed HDA is expected to yield some savings regarding computation time.

Fig. 13 shows the total computation time in millisecond (sum of the times for both offline and online) for the three long-term online sequences of different sizes. The mean training time was computed

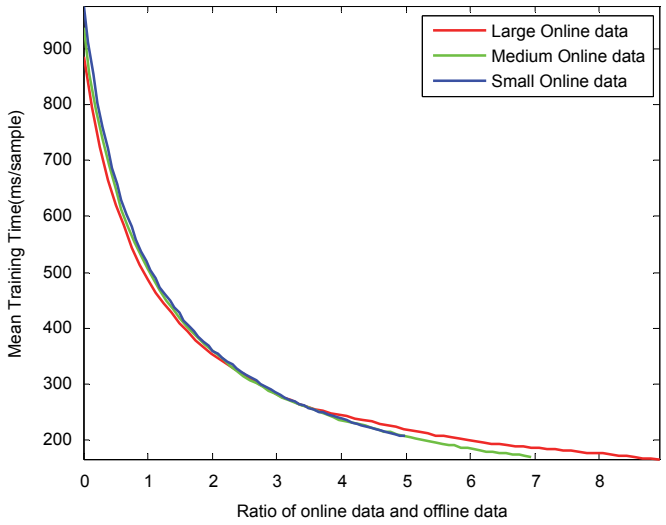

Fig. 13. Computational time for online HDA with offline PC-KFA.

as the total training time divided by the number of processed samples.

Table 8 shows the individual means of offline and online training from Fig. 13. Table 8 demonstrates that the computation time for online training was much shorter than that for offline training; on average, a spectacular 93% reduction of computation time per sample was achieved. Fig. 13 also shows that the calculation of online training was computationally efficient as the ratio of online to offline training increased. Therefore, HDA using online PC-KFA was better-suited to handle long-term sequences in a real-time manner. Our computation speed for a larger CTC database was promising in making CTC an acceptable technique for larger screening datasets.

5.7. Comparison between offline and online training

The comparison between offline and online data was evaluated in Fig. 14. As in the case of three long term sequences discussed in Section 5.5, the data ratio between online and offline varied from 0 to 9, in which 0 means offline datasets, and 9 means the majority of data is

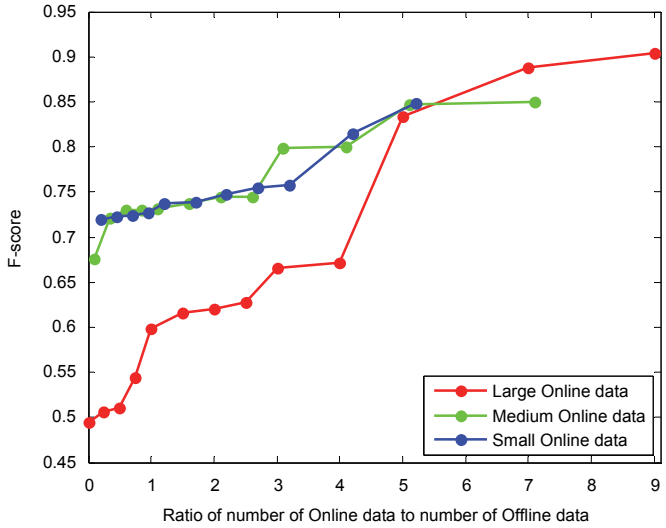

Fig. 14. F-score between online versus offline data to evaluate the merit of online learning. The horizontal axis shows the ratio of the number of online data to the number of offline data. The vertical axis denotes the F-score to measure the classification performance. The three long sequences are labeled as 'Large Online Data', 'Medium Online Data' and 'Small Online Data', which corresponded to offline training datasets of sizes of 750, 937, and 1250, respectively. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

online, i.e. the online trained dataset contains 9 times of the offline trained data. The F-score was used for the evaluation criterion. Overall, large online data (Red in Fig. 14) consistently demonstrates higher improvement in performance compared to small online data (Blue in Fig. 14). The large online data-set had a very small offline data set, hence demonstrated very low F-score (~50%) for offline training. But with increasing online training, this dataset (large online dataset, marked red in Fig. 14) achieved the highest F-score, thus validating the effectiveness of online training.

Table 9 shows the average F-scores shown in Fig. 14 for comparing offline training and online training. Table 9 also shows the improvement of F-scores over the different ratio of online datasets. Lager online data improved F-scores consistently, which demonstrated the merit of the proposed online training based on heterogeneous data. This table validates that, even though offline data had a very low F-score (49.5% in average), the large online data significantly improved the metric with the improvement of 82%, compared to offline training.

The experimental results showed that the proposed online learning continues to increase the classification performance metrics at various magnitude. Our experiments showed that the amount of increment varied randomly. If the data chunks fed to the algorithm is deterministically controlled (i.e., preset sequence of data samples), the increment becomes predictable; this supports our claim regarding the data-dependent performance. Table 10 shows a comparison between estimated improvements achieved by state-of-the-art online learning methods [18,49,66]. In these experiments, the improvements showed non-decreasing relation with relative size of the online-trained data and offline-trained data. Thus, in very high-dimensional and difficult-to-learn data, online learning was shown to be more beneficial because the new data come with new scenarios that are never experienced by the learning machine. The results shown in Table 10 support this observation because the CTC CAD data used in our experiments had a high dimension (in the order of 10^6 features).

Ozawa et al. [49] tested their Chunk IPCA on several 2D image datasets and text data, both of which are much lower in dimension. The maximum achievable performance is likely to be lower, as shown in Table 10. Kim et al.'s [18] and Langone et al.'s [66] examined the IPCA and IKSC on a number of OCR images and short PM_{10} datasets, respectively. The results shown in Table 10 support our claim that a significant improvement can be obtained by the PC-KFA that underlines the proposed online training for big heterogeneous data.

Table 9
F-score for the three online trajectories and comparison of offline vs. online training.

Online data size	F-Score (%)		
	Offline	300% Online	Maximum Improvement
Small online data sequence	71.9	75.4	18
Medium online data sequence	67.6	79.9	25
Large online data sequence	49.5	66.5	82

Table 10
Comparison with other relevant methods.

Method	Data Type	Online-Offline ratio	Mean Improvement (%)
Proposed Online KPCA	CTC CAD	1	0.9
Proposed Online KPCA	CTC CAD	3	2.1
Chunk IPCA[49]	2D Image/ Text	1	0.2
IKPCA[18]	OCR	1	0.12
IKSC[66]	PM_{10}	1	0.05

6. Conclusions

This paper addresses the problems of adding online datasets to existing offline datasets. To quantify the data compatibility, we introduced measures specifically designed for online medical image studies. We proposed an HDA to handle the long-term heterogeneous trajectories of online data based on PC-KFA. We applied data dependent composite kernels to clinical datasets of colonic polyps by maximizing of a measure of class separability in the empirical feature space of the datasets. The composite combination vector (i.e., weight vector) for the most dominant kernels was determined by maximizing the AF. We used the properties of heterogeneous degree to dynamically adjust the changes in the heterogeneous data during the online training. The advantages of the online method were to 1) achieve a fast and efficient feature extraction for the detection of polyps on CTC images, and 2) improve the performance of CAD when applied to large datasets that are continuously expanded with additional CTC cases. Experimental results showed that the online PC-KFA provided a classification performance of online training data to that of the offline training data. HDA applied to long-term data sequences in a model-based CAD scheme yielded a high detection performance of polyps. Future work to overcome the current limitation includes 1) improvement of the results such as type-I errors, and 2) addition of more datasets from heterogeneous data sources. Such an online framework has the potential of making CTC a viable option for screening of a large patient population, resulting in early detection of colon cancer, and ultimately leading to reduced mortality due to colon cancer.

Acknowledgments

The work reported here would not be possible without the help of many of the past and present members of the laboratory, in particular, D. Mar, S. Myla and L. Winter.

References

[1] Y. Motai, H. Yoshida, Principal composite kernel feature analysis: data-dependent kernel approach, IEEE Trans. Knowl. Data Eng. 25 (8) (2013) 1863–1875.
[2] D.S. Elizabeth, H.K. Nehemiah, C.S.R. Raj, A. Kannan, A novel segmentation approach for improving diagnostic accuracy of CAD systems for detecting lung cancer from chest computed tomography images, ACM J. Data Inf. Qual. 3 (2) (2012) 16 (Article 4).
[4] S. Sachdeva, S. Bhalla, Semantic interoperability in standardized electronic health record databases, J. Data Inf. Qual. 3 (1) (2012) 37 (Article 1).
[5] K.D. Bodily, J.G. Fletcher, T. Engelby, M. Percival, J.A. Christensen, B. Young, A.J. Krych, D.C.V. Kooi, D. Rodysill, J.L. Fidler, C.D. Johnson, Nonradiologists as second readers for intraluminal findings at CT colonography, Acad. Radiol. 12 (1) (2005) 67–73.
[6] J.G. Fletcher, F. Booya, C.D. Johnson, D. Ahlquist, CT colonography: unraveling the twists and turns, Curr. Opin. Gastroenterol. 21 (2005) 90–98.
[7] D. Hock, R. Ouhadi, R. Materne, A. Aouchria, I. Mancini, T. Broussaud, P. Magotteaux, A. Nchimi, Virtual dissection CT colonography: evaluation of learning curves and reading times with and without computer-aided detection, Radiology 248 (3) (2008) 860–868.
[9] H. Yoshida, J. Näppi, Three-dimensional computer-aided diagnosis scheme for detection of colonic polyps, IEEE Trans. Med. Imaging 20 (2001) 1261–1274.
[10] J. Näppi, H. Yoshida, Fully automated three-dimensional detection of polyps in fecal-tagging CT colonography, Acad. Radiol. 14 (3) (2007) 287–300.
[11] H. Yoshida, Y. Masutani, P. MacEneaney, D.T. Rubin, A.H. Dachman, Computerized detection of colonic polyps at CT colonography on the basis of volumetric features: pilot study, Radiology 222 (2002) 327–336.
[13] T.A. Chowdhury, P.F. Whelan, O. Ghita, A fully automatic CAD-CTC system based on curvature analysis for standard and low-dose CT data, IEEE Trans. Biomed. Eng. 55 (3) (2008) 888–901.
[14] J.W. Suh, C.L. Wyatt, Registration under topological change for CT colonography, IEEE Trans. Biomed. Eng. 58 (5) (2011) 1403–1411.
[15] L. Lu, D. Zhang, L. Li, J. Zhao, Fully automated colon segmentation for the computation of complete colon centerline in virtual colonoscopy, IEEE Trans. Biomed. Eng. 59 (4) (2012) 996–1004.
[16] A.L. Baert, P. Lefere, S. Gryspeerdt, Virtual Colonoscopy: A Practical Guide & Business Media, Springer Science, Mauer, Germany, 2009.
[18] B.J. Kim, I.K. Kim, K.B. Kim, Feature extraction and classification system for nonlinear and online data, Proc. Adv. Knowl. Discov. Data Min. 3056 (2004) 171–180.
[19] W. Zheng, C. Zou, L. Zhao, An improved algorithm for kernel principal component analysis, Neural Process. Lett. 22 (1) (2005) 49–56.
[20] A. Quan-Haase, Trends in online learning communities, SIGGROUP Bull. 25 (1) (2005) 2–6.
[21] M. Awad, Y. Motai, H. Yoshida, A clinical decision support framework for incremental polyps classification in virtual colonoscopy, special issue on machine learning for medical imaging, Algorithms 3 (2010) 1–20.
[22] L. Winter, Y. Motai, A. Docef, On-line versus off-line accelerated kernel feature analysis: application to computer-aided detection of polyps in CT colonography, Signal Process. 90 (8) (2010) 2456–2467.
[23] W. Cai, J.-G. Lee, M.E. Zalis, H. Yoshida, Mosaic decomposition: an electronic cleansing method for inhomogeneously tagged regions in noncathartic CT colonography, IEEE Trans. Med. Imaging 30 (3) (2011) 559–574.
[24] H. Xiong, M.N.S. Swamy, M.O. Ahmad, Optimizing the data- dependent kernel in the empirical feature space, IEEE Trans. Neural Netw. 16 (2) (2005) 460–474.
[25] J. Ye, S. Ji, J. Chen, Multi-class discriminant kernel learning via convex programming, J. Mach. Learn. Res. 9 (2008) 719–758.
[26] S. Althloothi, M.H. Mahoor, X. Zhang, R.M. Voyles, Human activity recognition using multi-features and multiple kernel learning, Pattern Recognit. 47 (5) (2014) 1800–1812.
[27] M. Gönen, E. Alpaydın, Localized algorithms for multiple kernel learning, Pattern Recognit. 46 (3) (2013) 795–807 (3).
[28] X.JiangY.MotaiR.R.SnappX.Zhu, Accelerated kernel feature analysis, in: Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 2006, pp. 109–116.
[29] V.N. Vapnik, The nature of statistical learning theory, 2nd ed., Springer, New York, 2000.
[30] F. Orabona, C. Castellini, B. Caputo, L. Jie, G. Sandini, On-line independent support vector machines (4)Pattern Recognit. 43 (4) (2010) 1402–1412.
[31] B.J. Kim, J.Y. Shim, C.H. Hwang, I.K. Kim, J.H. Song, Incremental feature extraction based on empirical kernel map, Found. Intell. Syst. 2871 (2003) 440–444.
[32] B.J.KimI.K.Kim, Incremental non-linear PCA for classification, in: Proceedings of the Knowledge Discovery in Databases, 3202, 2004, pp. 291–300.
[33] T.-J. Chin, D. Suter, Incremental kernel principal component analysis, IEEE Trans. Image Process. 16 (6) (2007) 1662–1674.
[34] L. Hoegaerts, L.D. Lathauwer, I. Goethals, J.A.K. Suykens, J. Vandewalle, B.D. Moor, Efficiently updating and tracking the dominant kernel principal components, Neural Netw. 20 (2) (2007) 220–229.
[35] H. Hoffmann, Kernel PCA for novelty detection, Pattern Recognit. 40 (3) (2007) 863–874.
[36] D.S. Paik, C.F. Beaulieu, G.D. Rubin, B. Acar, R.B. Jeffrey Jr., J. Yee, J. Dey, S. Napel, Surface normal overlap: a computer-aided detection algorithm with application to colonic polyps and lung nodules in helical CT, IEEE Trans. Med. Imaging 23 (2004) 661–675.
[37] T.DamoulasM.A.Girolami, Probabilistic multi-class multi-kernel learning: on protein fold recognition and remote homology detection, in: Proceedings of Bioinformatics, 2008, pp. 1264–1270.
[38] S. Amari, S. Wu, Improving support vector machine classifiers by modifying kernel functions, Neural Netw. 12 (6) (1999) 783–789.
[39] B.F. de Souza, A.P.L.F. de Carvalho, Gene selection based on multi-class support vector machines and genetic algorithms, Mol. Res. 4 (3) (2005) 599–607.
[40] Q. Chen, Q. Sun, D. Xia, Homogeneity similarity based image denoising, Pattern Recognit. 43 (12) (2010) 4089–4100.
[41] R.O. Duda, D.G. Stork, P.E. Hart, Pattern Classification, 2nd ed., John Wiley & Sons Inc, Hoboken, NJ, USA, 2001.
[42] B. Schölkopf, A.J. Smola, Learning with Kernels: Support Vector Machines, Regularization, Optimization, and Beyond. Adaptive Computation and Machine Learning, MIT press, Cambridge, MA, USA, 2002.
[43] H.FröhlichO.ChapelleB.Scholkopf, Feature selection for support vector machines by means of genetic algorithm. in: Proceedings of the 15th IEEE InternationalConference on Tools with Artificial Intelligence, 2003, pp. 142–148.
[44] X.W.Chen, Gene selection for cancer classification using bootstrapped genetic algorithms and support vector machines, in: Proceedings of the IEEE International Conference of Computational Systems, Bioinformatics, 2003, pp. 504–505
[45] C.ParkS.-B.Cho, Genetic search for optimal ensemble of feature-classifier pairs in DNA gene expression profiles, in: Proceedings of the International Joint Conference on Neural Networks, 3, 2003, pp. 1702–1707.
[46] M.SzafranskiY.GrandvaletA.Rakotomamonjy, Composite kernel learning, in: Proceedings of the 25th International Conference on Machine learning (ICML '08). ACM, New York, NY, USA, 2008, pp. 1040–1047.
[47] Y.YuanF.WuY.ZhuangJ.Shao, Image annotation by composite kernel learning with group structure, in: Proceedings of the 19th ACM InternationalConference on Multimedia (MM '11). ACM, New York, NY, USA, 2011, pp. 1497–1500.
[48] J. Kivinen, A.J. Smola, R.C. Williamson, Online learning with kernels, Trans. Signal Process. 52 (8) (2004) 2165–2176.
[49] S. Ozawa, S. Pang, N. Kasabov, Incremental learning of chunk data for online pattern classification systems, IEEE Trans. Neural Netw. 19 (6) (2008) 1061–1074.
[50] Y. Li, On incremental and robust subspace learning, Pattern Recognit. 37 (7) (2004) 1509–1518.
[51] Y. Kim, Incremental principal component analysis for image processing, Opt. Lett. 32 (1) (2007) 32–34.
[52] C.StevenH.HoiM.R.LyuE.Y.Chang, Learning the unified kernel machines for classification, in: Proceedings of the 12th ACM SIGKDD InternationalConference on Knowledge discovery and data mining (KDD '06). ACM, New York, NY, USA, 2006, pp. 187–196.
[53] Y.XuF.ShenW.PingJ.Zhao, TAKES: a fast method to select features in the kernel space, in: Proceedings of the 20th ACM InternationalConference on Information

and knowledge management (CIKM '11), ACM, New York, NY, USA, 2011, pp. 683–692.
[54] N. Cristianini, J. Kandola, A. Elisseeff, J. Shawe-Taylor, On kernel target alignment, Proc. Neural Inf. Process. Syst. (2001) 367–373.
[55] A.E. Kaufman, S. Lakare, K. Kreeger, I. Bitter, Virtual colonoscopy, Commun. ACM 48 (2) (2005) 37–41.
[56] H. Yoshida, J. Näppi, CAD in CT colonography without and with oral contrast agents: progress and challenges, Comput. Med. Imaging Graph. 31 (2007) 267–284.
[57] T.BriggsT.Oates, Discovering domain-specific composite kernels, in: Proceedings of the 20th National Conference on Artificial Intelligence, 2, 2005 pp. 732–738.
[58] X. Han, Nonnegative principal component analysis for cancer molecular pattern discovery, IEEE/ACM Trans. Comput. Biol. Bioinform. 7 (3) (2010) 537–549.
[60] Cancer Facts & Figures, American Cancer Society, 2014.
[61] J. Näppi, H. Yoshida, Fully automated three-dimensional detection of polyps in fecal-tagging CT colonography, Acad. Radiol. 14 (2007) 593–606.
[62] A.H. Dachman, N.A. Obuchowski, J.W. Hoffmeister, et al., Effect of computer-aided detection for CT colonography in a multireader, multicase trial, Radiology 256 (2010) 827–835.
[63] X. Zhu, H.I. Suk, D. Shen, A novel matrix-similarity based loss function for joint regression and classification in AD diagnosis, NeuroImage 100 (2014) 91–105.
[64] X. Zhu, Z. Huang, H.T. Shen, J. Cheng, C. Xu, Dimensionality reduction by mixed kernel canonical correlation analysis, Pattern Recognit. 45 (8) (2012) 3003–3016.
[65] X. Zhu, Z. Huang, Y. Yang, H.T. Shen, C. Xu, J. Luo, Self-taught dimensionality reduction on the high-dimensional small-sized data, Pattern Recognit. 46 (1) (2013) 215–229.
[66] R. Langone, O.M. Agudelo, B.D. Moor, J.A.K. Suykens, Incremental kernel spectral clustering for online learning of non-stationary data, Neurocomputing 139 (2014) 246–260.

Yuichi Motai received the B.Eng. degree in instrumentation engineering from Keio University, Tokyo,Japan, in 1991, the M.Eng. degree in applied systems science from Kyoto University, Kyoto, Japan, in1993, and the Ph.D. degree in electrical and computer engineering from Purdue University, WestLafayette, IN, USA, in 2002. He is currently an Associate Professor of Electrical and ComputerEngineering at Virginia Commonwealth University, Richmond, VA, USA. His research interests includethe broad area of sensory intelligence; particularly in medical imaging, pattern recognition, computervision, and sensory-based robotics.

Nahian Alam Siddique received his B.S. degree in Electrical and Electronic Engineering fromBangladesh University of Engineering and Technology, Dhaka, Bangladesh, in 2011. Presently he ispursuing graduate studies at the department of Electrical and Computer Engineering in VirginiaCommonwealth University, Richmond, VA, USA. His research interests include specific areas ofsensory intelligence—particularly in medical imaging, pattern recognition and computer vision.

Hiroyuki Yoshida received his B.S. and M.S. degrees in physics and a Ph.D. degree in informationscience from the University of Tokyo, Japan. He previously held an Assistant Professorship in theDepartment of Radiology at the University of Chicago. He was a tenured Associate Professor when heleft the university and joined the Massachusetts General Hospital (MGH) and Harvard Medical School(HMS) in 2005, where he is currently the Director of 3D Imaging Research in the Department ofRadiology, MGH and an Associate Professor of Radiology at HMS. His research interest is in computeraided diagnosis, in particular the detection of polyps in CT colonography, for which he receivedseveral awards at the Annual Meetings of Radiological Society of North America and the InternationalSociety for Optical Engineering.