Adaptive Batch Mode Active Learning

Shayok Chakraborty, *Member, IEEE*, Vineeth Balasubramanian, *Member, IEEE*,
and Sethuraman Panchanathan, *Fellow, IEEE*

***Abstract*— Active learning techniques have gained popularity to reduce human effort in labeling data instances for inducing a classifier. When faced with large amounts of unlabeled data, such algorithms automatically identify the exemplar and representative instances to be selected for manual annotation. More recently, there have been attempts toward a batch mode form of active learning, where a batch of data points is simultaneously selected from an unlabeled set. Real-world applications require adaptive approaches for batch selection in active learning, depending on the complexity of the data stream in question. However, the existing work in this field has primarily focused on static or heuristic batch size selection. In this paper, we propose two novel optimization-based frameworks for adaptive batch mode active learning (BMAL), where the batch size as well as the selection criteria are combined in a single formulation. We exploit gradient-descent-based optimization strategies as well as properties of submodular functions to derive the adaptive BMAL algorithms. The solution procedures have the same computational complexity as existing state-of-the-art static BMAL techniques. Our empirical results on the widely used VidTIMIT and the mobile biometric (MOBIO) data sets portray the efficacy of the proposed frameworks and also certify the potential of these approaches in being used for real-world biometric recognition applications.**

***Index Terms*— Batch mode active learning (BMAL), biometric recognition, numerical optimization, submodular functions.**

I. Introduction

THE rapid escalation of technology and the widespread emergence of modern technological equipments have resulted in the generation of humongous amounts of digital data (in the form of images, videos, and text). This has expanded the possibilities of solving real-world problems using computational learning frameworks. However, while gathering large amounts of digital data is cheap and easy, annotating them with class labels (to train a classifier) is an expensive process in terms of time, labor, and human expertise. This has paved the way for research in the field of *active learning*. Active learning algorithms automatically select the salient and promising instances from large quantities of unlabeled data. This tremendously reduces the human annotation effort as only a few examples, which are identified by the algorithm, need to be labeled manually.

Conventional active learning techniques select only a single data instance at a time for manual labeling and retrain the classifier after every individual query. This results in frequent model retraining; also, it utilizes only a single labeling oracle at a time. With the advent of technologies like the Amazon Mechanical Turk [1], it is now possible to leverage the intelligence of multiple human users simultaneously in labeling data instances to train a classification model. To this end, *batch mode active learning* (BMAL) techniques have been proposed in recent years. Such algorithms attempt to select a batch of unlabeled data points simultaneously from an unlabeled set instead of a single instance at a time. Sample applications of such a scheme include content-based image retrieval [2], medical image classification [3], and text classification [4]. BMAL algorithms are of paramount importance in applications involving video data. Modern video cameras have a high frame rate, and consequently, the captured data have high redundancy. Selecting batches of relevant frames from a superfluous frame sequence in captured videos is therefore a significant and valuable challenge.

An ideal BMAL system can be conceptualized as consisting of two main steps: 1) deciding the batch size (the number of image frames to be queried from a given unlabeled video stream) and 2) selecting the most appropriate images from the unlabeled video once the batch size has been determined. Both these steps are critical in ensuring maximum generalization capability of the learner with minimum human labeling effort, which is the primary objective in any active learning application. However, the existing few efforts on BMAL focus only on the second step of identifying a criteria for selecting informative batches of data samples and require the batch size to be specified in advance by the user [5], [6]. In a real-world application, deciding on the batch size (the number of relevant instances in a data stream) in advance and without any knowledge of the data stream being analyzed may not lead to a good generalization accuracy. The batch size should depend on the quality and complexity of the samples in the unlabeled stream and also on the level of confidence of the current classifier on the unlabeled data instances. In other words, there is a strong need for dynamic batch selection in BMAL algorithms.

Manuscript received April 6, 2013; revised June 11, 2014; accepted August 27, 2014. Date of publication September 29, 2014; date of current version July 15, 2015.

S. Chakraborty was with the Center for Cognitive Ubiquitous Computing (CUbiC), Arizona State University, Tempe, AZ 85281, USA. He is now with the Department of Electrical and Computer Engineering, Carnegie Mellon University, Pittsburgh, PA 15213 USA (e-mail: shayokc@andrew.cmu.edu).

V. Balasubramanian was with the Center for Cognitive Ubiquitous Computing (CUbiC), Arizona State University, Tempe, AZ 85281, USA. He is now with the Department of Computer Science and Engineering, Indian Institute of Technology Hyderabad, Telangana 502205, India (e-mail: vineethnb@iith.ac.in).

S. Panchanathan is with Center for Cognitive Ubiquitous Computing (CUbiC), Arizona State University, Tempe, AZ 85281 USA (e-mail: panch@asu.edu).

This paper has supplementary downloadable material available at http://ieeexplore.ieee.org., provided by the author. Supplemental File size: 700KB

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TNNLS.2014.2356470

In this paper, we propose two novel BMAL algorithms that adaptively select samples for manual annotation based on the complexity of the data stream being analyzed and the cost of labeling each unlabeled data sample. We develop one formulation for dynamic batch selection that directly optimizes the performance of the updated learner (the learner trained on the current training set together with the newly selected batch). The batch selection problem is solved using the stochastic gradient-descent (SGD) algorithm to simultaneously decide the batch size and identify the specific points that need to be queried for manual annotation, through a single framework. We also derive a second formulation for dynamic batch selection based on the uncertainty of the current learner. We exploit the properties of submodular functions and propose an efficient solution strategy for adaptive batch selection through a single optimization framework. We validate the proposed methods on challenging real-world data sets for face-based biometric recognition, which is used as the exemplar application in this paper. Although validated on biometric data, the proposed frameworks are generic and can be used in any application where it is required to select a batch of representative entities simultaneously from redundant/repetitive data samples.

The rest of this paper is organized as follows: we present a survey of the existing BMAL techniques in Section II, detail the mathematical formulations of our algorithms in Section III, present the results of our experiments in Section IV, and conclude with discussion in Section V.

II. Literature Survey

Active learning is a well-studied problem in machine learning literature [7], [8]. Most of the existing active learning algorithms have focused on selecting a single informative unlabeled instance to query each time, an approach called *pool-based active learning*. Such techniques can be broadly categorized into four groups:

1) Support Vector Machines (SVM)-based approaches, which decide the next point to be queried based on its distance from the hyperplane in the feature space [9];
2) statistical approaches, which query data instances such that some statistical property of the future learner (e.g., the learner variance) is optimized [10], [11];
3) query by committee, which chooses points to be queried based on the level of disagreement among an ensemble of classifiers [12], [13];
4) information theoretic approaches, which exploit the discriminative partition information contained in the unlabeled data and queries the instance that provides the maximum conditional mutual information about the labels of the unlabeled instances, given the labeled data, in an optimistic way [14].

To avoid frequent classifier retraining and to utilize the presence of parallel labeling oracles, BMAL schemes, which select multiple unlabeled points simultaneously for manual annotation, have been proposed in recent years. Existing approaches for BMAL have largely been based on extending pool-based active learning methods to select multiple instances simultaneously. They use greedy heuristics and select the top k informative instances (k being the required batch size) from the unlabeled set for manual annotation. Brinker [15] extended the version space concept proposed in [9] to query a diverse batch of points using SVMs, where diversity was measured as the angle induced by the hyperplane of the currently selected point to the hyperplanes of the already selected points. Schohn and Cohn [16] proposed to query a batch of points based on their distance from the separating hyperplane of a linear SVM. Xu *et al.* [17] proposed an SVM-based BMAL strategy that combined representativeness and diversity measures for batch selection.

However, extending the pool-based setting to the batch setting by considering the top k instances does not account for other factors such as information overlap among the selected points in a batch. More recently, this has led to newer efforts that are specifically intended to select batches of points using appropriate optimization strategies. Hoi *et al.* [2], [4] used the Fisher information matrix as a measure of model uncertainty and proposed to query the set of points that maximally reduced the Fisher information. Hoi *et al.* [18] proposed a BMAL scheme based on SVMs where a kernel function was first learned from a mixture of labeled and unlabeled samples, which was then used to identify the informative and diverse examples through a min–max framework. They also exploited submodular optimization for BMAL in the context of image retrieval [19]. Guo and Schuurmans [5] proposed a discriminative strategy that selected a batch of points that maximized the log-likelihoods of the selected points with respect to their optimistically assigned class labels and minimized the entropy of the unselected points in the unlabeled pool. Very recently, Guo [6] proposed a BMAL scheme that maximized the mutual information between the labeled and unlabeled sets and was independent of the classification model used. The methods described in [5] and [6] have been shown to be the best performing BMAL schemes till date.

All the aforementioned techniques of BMAL, including [5] and [6], concentrate only on the design of a selection criterion assuming that the batch size is chosen by the user in advance. In most real-world problems, this is not a practical assumption, as explained in Section I. We would expect the number of relevant samples to be large when the active learner is exposed to an unlabeled data stream of high complexity (for example, one which contains data samples very different from the current training set) and the number to be low for unlabeled data streams that are similar in composition to the current training set. Thus, there is a strong need for the active learner to adapt to different contexts and dynamically decide the batch size as well as the specific instances to be queried. In this paper, we present two novel optimization-based strategies to adaptively compute the batch size and decide the specific instances for manual annotation through a single framework. We now present the mathematical formulations of our approaches.

III. Dynamic BMAL: Mathematical Formulation

In this section, we present the details of the proposed dynamic BMAL formulations, which aim to simultaneously identify the batch size and the batch of samples itself with the same computational complexity as existing static BMAL approaches. To this end, we first pose the dynamic BMAL problem as selecting a set of instances that maximizes the performance of the future learner (the learner trained on the current training set and the newly selected batch) and solve the problem using SGD. We further propose a second framework for dynamic BMAL using submodular optimization, where the time complexity is significantly reduced by avoiding the use of the future learner (i.e., we select the batch of samples based on the learner trained on the current training set alone). We also show later in this section on how these two methods can be easily adapted to static BMAL (where a batch size is prespecified), thus making this contribution a generalizable BMAL framework.

A. Dynamic BMAL via SGD

Consider a BMAL setting that has a current labeled set L_t and a current classifier w^t trained on L_t. The classifier is exposed to an unlabeled set U_t at time t. The objective is to select a batch B from the unlabeled stream in such a way that the classifier w^{t+1}, at time $t+1$, trained on $L_t \cup B$ has maximum generalization capability (we refer to w^{t+1} as the future model or future classifier). With unlabeled data being available, semisupervised learning methods have been proposed that train models by minimizing the uncertainty of the labels for the unlabeled instances [20]. That is, to achieve a classifier with good generalization performance, one can minimize the entropy of the missing labels for the unlabeled data. In our active learning framework, we attempt to minimize the entropy of the updated learner on the remaining $|U_t - B|$ samples after batch selection. Let C denote the total number of classes. The entropy of the conditional distribution $P(y|x_j, w^{t+1})$ is given by

$$S(y|x_j, w^{t+1}) = -\sum_{y \in C} P(y|x_j, w^{t+1}) \log P(y|x_j, w^{t+1}). \quad (1)$$

Furthermore, to maximize the contribution of the selected unlabeled samples, diversity-based selection criteria have been proposed [21], which ensure that the selected samples are less similar with the already available labeled data. In our formulation, we quantify the diversity, ρ_j, of an unlabeled sample x_j as its mean kernelized distance from all the labeled points in the training set

$$\rho_j = \frac{1}{n_l} \sum_{i=1}^{n_l} \phi(x_i, x_j) \quad (2)$$

where n_l is the number of samples in the training set and ϕ denotes the kernel function. Such a distance measure is widely used in metrics like the maximum mean discrepancy to quantify the difference between two probability distributions [22], [23]. The two aforementioned criteria can be combined by defining a score function as follows:

$$f(B) = \sum_{j \in B} \rho_j - \lambda_1 \sum_{j \in U_t - B} S(y|x_j, w^{t+1}). \quad (3)$$

The first term denotes the sum of the average kernelized distances of each selected unlabeled point from the labeled set (to ensure selection from data densities with low representation in the original training set), while the second term quantifies the sum of the entropies of the updated learner on each remaining point in the unlabeled stream (which is expected to be low, if the selection is appropriate). λ_1 is a tradeoff parameter governing the relative importance of the two terms.

The problem therefore reduces to selecting a batch B of unlabeled points, which produces the maximum score $f(B)$. Let the batch size (the number of samples to be selected for annotation) be denoted by m, which is an unknown. Since there is no restriction on the batch size m, the obvious intuitive solution to this problem is to select all the samples in the unlabeled set. Then, the entropy term becomes zero, and the distance term attains its maximum value. Therefore, $f(B)$ will also attain its maximum score. However, querying all the samples for their class labels is not an elegant solution and defeats the basic purpose of active learning. To prevent this, we modify the score function by enforcing a penalty on the batch size as follows:

$$\tilde{f}(B) = \sum_{j \in B} \rho_j - \lambda_1 \sum_{j \in U_t - B} S(y|x_j, w^{t+1}) - \lambda_2 m. \quad (4)$$

The third term essentially reflects the cost associated with labeling the data samples, as the value of the objective function decreases with every single sample that needs to be labeled. Defining the score function in this way ensures that any and every sample is not queried for its class label; only samples for which the distance and entropy terms outweigh the labeling cost term get selected. The coefficient λ_2 is the cost parameter and denotes the cost associated with labeling one unlabeled data sample. This parameter can be set based on the given application. For instance, manually labeling a face image is less tedious as compared with labeling a voicemail message as urgent/nonurgent (as the human oracle has to listen to the entire message for accurate annotation). Thus, λ_2 will have a smaller value in the case of a face recognition application, as compared with a voicemail recognition system. In our experiments, we assume λ_2 to be one and also study the effect of this parameter on the batch size and the accuracy of recognition.

As per (4), we need to select a batch B of unlabeled points so as to maximize $\tilde{f}(B)$. Since brute force search methods are prohibitive, we employ numerical optimization techniques to solve this problem. We define a binary vector M of size $|U_t|$ where each entry denotes whether the corresponding point is to be queried for its class label. We rewrite the objective function in (4) into an equivalent function in terms of the defined vector M

$$\max_{M,m} \sum_{j \in U_t} \rho_j M_j - \lambda_1 \sum_{j \in U_t} (1 - M_j) S(y|x_j, w^{t+1}) - \lambda_2 m \quad (5)$$

s.t.

$$M_j \in \{0, 1\} \quad \forall j. \quad (6)$$

In this formulation, note that if an entry of M is 1, the corresponding image will be selected for annotation, and if it is 0, the image will not be selected. The number of images to be selected is therefore equal to the number of nonzero entries in the vector M, or the zero-norm of M. Hence

$$m = ||M||_0 \approx ||M||_1 = \sum_j M_j. \tag{7}$$

Here, we have replaced the zero norm of M by its tightest convex approximation, which is the one-norm of M (inspired by the work in [24]). In addition, from (6), the one-norm is simply the sum of the elements of the vector M. Substituting m in terms of M, the formulation becomes

$$\max_M \sum_{j \in U_t} \rho_j M_j - \lambda_1 \sum_{j \in U_t} (1 - M_j) S(y|x_j, w^{t+1}) - \lambda_2 \sum_j M_j \tag{8}$$

s.t.

$$M_j \in \{0, 1\} \quad \forall j.$$

The above optimization is an integer programming problem and is NP-hard. We therefore relax the constraint to make it a continuous optimization problem

$$\max_M \sum_{j \in U_t} \rho_j M_j - \lambda_1 \sum_{j \in U_t} (1 - M_j) S(y|x_j, w^{t+1}) - \lambda_2 \sum_j M_j \tag{9}$$

s.t.

$$0 \le M_j \le 1 \quad \forall j.$$

1) Solving the Optimization Problem: We define an objective function $f(M)$ as (from 9)

$$\begin{aligned} f(M) = &\sum_{j \in U_t} \rho_j M_j - \lambda_1 \sum_{j \in U_t} (1 - M_j) S(y|x_j, w^{t+1}) \\ &- \lambda_2 \sum_j M_j. \end{aligned} \tag{10}$$

To solve the optimization problem, we use the quasi-Newton method [25]. The first derivative of the function and the Hessian matrix of second derivatives need to be computed as part of the solution procedure. Assuming that w^{t+1} remains constant with small iterative updates of M, the first-order derivative vector is obtained by taking the partial of the objective with respect to M

$$\nabla f(M_j) = \rho_j + \lambda_1 S(y|x_j, w^{t+1}) - \lambda_2. \tag{11}$$

The Hessian starts as an identity matrix and is updated according to the Broyden–Fletcher–Goldfarb–Shanno (BFGS) method [25]. The final value of M is used to govern the number of points and the specific points to be selected for the given data stream (by greedily setting the top m entries in M as 1 to recover the integer solution, where $m = \sum_j M_j$). Hence, solving a single optimization problem helps in dynamically deciding the batch size as well as selecting the specific points for manual annotation. It is to be noted that the objective function is defined in terms of the future classifier w^{t+1}, which is unknown. To compute the entropy term using w^{t+1} in the quasi-Newton iterations, we therefore need to estimate the class labels of the currently selected batch of unlabeled samples so as to intelligently approximate w^{t+1}. We used the semisupervised graph-based label propagation method, graph transduction via alternating minimization (GTAM), proposed in [26], to derive the labels of the selected unlabeled samples in each quasi-Newton iteration. This method is efficient in terms of accuracy and computational overhead [26]. We validate the efficiency of this method in our empirical evaluations (please refer to the supplemental file for the details of this algorithm). The pseudocode of the complete dynamic BMAL algorithm is outlined in Algorithm 1.

Algorithm 1 Dynamic BMAL via SGD

```
Require: Training set $L_t$, Unlabeled set $U_t$, parameters $\lambda_1$ and $\lambda_2$,
    initial random guess for $M$, a stopping threshold $\alpha$

Initialize the Hessian matrix $H$ as the identity matrix $I$
Evaluate the objective function $f(M)$ (Equation 10) and the
    derivative vector $\nabla f(M)$ (Equation 11)
repeat
    Solve the QP problem as required by Quasi-Newton:
        QP($H$,$\nabla f(M)$,$M$) and let the solution be $M^*$
    Compute the step size $s$ from the Armijo Goldstein Equations.
    Update $M$ as $M_{new} = M + s(M^* - M)$
    Evaluate the new objective $f(M_{new})$ and the new derivative
        vector $\nabla f(M_{new})$ using $M_{new}$
    Calculate the difference in objective value: $diff = abs(f(M)$
        $- f(M_{new}))$
    Update the Hessian $H$ using the BFGS Equations
    Update the objective value: $f(M) = f(M_{new})$
    Update the derivative vector: $\nabla f(M) = \nabla f(M_{new})$
    Update the vector $M$: $M = M_{new}$
until $diff \le \alpha$
Compute batch size $m = \sum M$ (Equation 7)
Greedily set the top $m$ entries in $M$ as 1 to recover the integer
    solution.
Select $m$ points accordingly
```

We also note that the specific terms in the objective function can be modified based on the particular application in question. For instance, one may want to design an objective function that selects samples by minimizing the uncertainty on the unselected examples and by maximizing the representativeness between the selected and the unselected samples in the unlabeled set. The same strategy based on a penalty on the batch size can be used in the objective function containing the relevant terms.

The proposed dynamic batch selection framework has the computational complexity of $O(n^2)$ (where n is the number of unlabeled data samples), which is the same as the state-of-the-art static BMAL techniques [5], [6], where the batch size needs to be prespecified (this complexity is due to the quasi-Newton method, which has quadratic complexity [25]). Thus, with the same computational complexity as state-of-the-art static BMAL schemes, we solve for both the size and the samples in a batch that needs to be queried from a given unlabeled data stream.

In our experiments, we performed a single run of the quasi-Newton method. We started with a random initial guess and iteratively updated the solution until convergence. Performing multiple runs may help in finding better quality local optima; however, it will also increase the computation time. Thus, in applications where computation time is not a major concern, one can perform multiple runs and select the best solution.

B. Dynamic BMAL via Submodular Optimization

Submodularity has been used for active learning in [19] in the context of image retrieval; however, the research was focused on static BMAL with a prespecified batch size. We propose a novel dynamic BMAL scheme based on submodular optimization. Similar to the previous problem, we are given a training set L_t and an unlabeled set U_t for adaptive batch selection. In this method, the uncertainty of an unlabeled sample is computed as the entropy of the current model w^t on this sample (instead of the updated model w^{t+1}, as in the previous formulation). However, since the goal in active learning is to select a batch of unlabeled samples that are maximally informative for the updated model w^{t+1}, we need to consider a redundancy-based criterion (which quantifies the similarity between a pair of samples) if we design the batch selection condition based on the current model w^t. This is because, if two points separately furnish valuable information, but they furnish the same/overlapping information, then both of them together may not be maximally informative for w^{t+1}. The redundancy criterion is important in this formulation, as the objective is to select a batch of useful samples for w^{t+1} using only the current model w^t. This was not necessary in the previous formulation as the performance was directly optimized with respect to the future model w^{t+1}. In this paper, redundancy was quantified as the minimum kernelized distance of an unlabeled point from the already selected batch (other measures of distance or similarity may be used based on the application in question). A greater value of the minimum distance denotes a more promising point from the redundancy perspective. We would like to select a batch of points where each point furnishes useful, but distinctly unique information. For this purpose, we formulate an objective function denoting the score of a set of points B as follows:

$$S(B) = \sum_{x_i \in B} [\rho_i + \lambda_1 E(x_i) + \lambda_2 D(x_i)] \tag{12}$$

where ρ_i is the average kernelized distance of the unlabeled point x_i from the training set, as defined in Section III-A, $E(x_i)$ is the entropy of x_i based on the current model w^t

$$E(x_i) = -\sum_{y \in C} P(y|x_i, w^t) \log P(y|x_i, w^t)$$

and

$$D(x_i) = \min_{x_j \in B: j \neq i} \phi(x_i, x_j)$$

which quantifies the similarity of an unlabeled point from the already selected set (ϕ denotes the kernelized distance). Thus, while ρ_i ensures selection of samples diverse from the training set, $D(x_i)$ avoids selection of duplicate samples in the batch. The tradeoff parameters λ_1 and λ_2 control the relative importance of the distance and entropy terms. Since the goal is to select a batch of points with high aggregate uncertainty scores and high distance among them, the objective is to select a set of points that maximizes the score $S(B)$, as defined in (12). This score function is monotonically nondecreasing (will be proved later) and since there is no restriction on the batch size, the obvious solution is to select all points in the unlabeled set for manual annotation. Similar to the previous formulation, we therefore impose a penalty on the batch size and modify the score function as follows:

$$S^{\text{new}}(B) = \sum_{x_i \in B} [\rho_i + \lambda_1 E(x_i) + \lambda_2 D(x_i)] - \lambda_3 |B|. \tag{13}$$

The last term in (13) represents the cardinality of the set B and increases as more points are queried in the batch. λ_3 is the cost parameter, as discussed in the case of the first BMAL method in Section III-A. The optimal batch selection criterion can thus be expressed as

$$\max_{B \subseteq U_t} S^{\text{new}}(B). \tag{14}$$

Due to the exponential nature of the search space, exhaustive search techniques are not feasible. In the following sections, we derive an efficient strategy to solve the above optimization problem.

1) Submodularity of the Objective Function: The definition of submodularity of a function is as follows.

Definition 1: Let Z be a finite set and let X and Y be two subsets of Z such that $X \subseteq Y \subseteq Z$. Consider an element $x \in Z \backslash Y$. A function $f : 2^Z \rightarrow \Re$ is submodular if

$$f(X \cup \{x\}) - f(X) \geq f(Y \cup \{x\}) - f(Y).$$

That is, a function is submodular if adding an element to a set increases the functional value by at least as much as adding the same element to its superset (also called the diminishing returns property [27], [28]).

Lemma 1: The score function $S(B)$, as defined in (12), is a submodular set function.

Proof: Let B_1 and B_2 be two sets formed by selecting unlabeled points from U_t, such that $B_1 \subseteq B_2 \subseteq U_t$ and consider an unselected instance $x \in U_t \backslash B_2$. The increment in the value of the objective function achieved by appending x to the set B_1 is given by

$$S(B_1 \cup \{x\}) - S(B_1) = \rho_x + \lambda_1 E(x) + \lambda_2 \min_{x_j \in B_1} \phi(x, x_j).$$

Similarly, the increment obtained by appending x to the set B_2 is

$$S(B_2 \cup \{x\}) - S(B_2) = \rho_x + \lambda_1 E(x) + \lambda_2 \min_{x_j \in B_2} \phi(x, x_j).$$

Since $B_1 \subseteq B_2$, the minimum distance of a point x from the other points will always be greater for the set B_1 as there may exist some point x_j in the superset B_2, which is closer to x than any element in its subset B_1. Hence

$$\min_{x_j \in B_1} \phi(x, x_j) \geq \min_{x_j \in B_2} \phi(x, x_j).$$

Thus, we have

$$S(B_1 \cup \{x\}) - S(B_1) \geq S(B_2 \cup \{x\}) - S(B_2).$$

This completes the proof of the lemma. □

Algorithm 2 Dynamic BMAL via Submodular Optimization

```
Require: Training set $L_t$ and Unlabeled set $U_t$, parameters $\lambda_1$, $\lambda_2$ and $\lambda_3$

1: Train a classifier $w^t$ on the training set $L_t$
2: $B = \{\phi\}$
3: for $i = 1 \rightarrow |U_t|$ do
4:     for all $x \in U_t \backslash B$ do
5:         $B_{temp} = B \cup \{x\}$
6:         Compute $S(B_{temp})$ as in Equation (12)
7:     end for
8:     Select the point $x_{max}$ producing the largest gain in the objective function (Equation 12)
9:     $B = B \cup \{x_{max}\}$
10:    $U_t = U_t \backslash \{x_{max}\}$
11:    Evaluate the current score $S(B)$
12:    $S^{new}B(i) = S(B) - \lambda_3 * |B|$
13: end for
14: Batch Size $m = argmax(S^{new}(B))$
15: Point Set $P = B(1 : m)$
16: return $m$ and $P$
```

Lemma 2: The score function $S(B)$ is a monotonically nondecreasing function.

Proof: Let B_1 denote the currently selected set of points and consider an element $x \in U_t \backslash B_1$, U_t being the unlabeled pool. If x is added to the current set, the value of the objective function changes by $\rho_x + \lambda_1 E(x) + \lambda_2 \min_{x_j \in B_1} \phi(x, x_j)$. Both the entropy and distance are nonnegative quantities, and hence

$$S(B_1 \cup \{x\}) \geq S(B_1).$$

This completes the proof. □

2) Greedy Solution to the Optimization Problem: The problem of maximizing a submodular function is NP-hard. However, Nemhauser *et al.* [27] established that for a function S, which is submodular and nondecreasing, with $S(\Phi) = 0$ (Φ being the null set), a greedy approach can provide an efficient solution with near-optimal results [from the definition of S in (12), it is obvious that $S(\Phi) = 0$]. In our case, the suggested greedy approach incrementally selects points from the unlabeled set by maximizing the gain in the objective function in each iteration. It presents an incremental ordering of the samples based on their degree of usefulness. A single run of the algorithm over the unlabeled set therefore provides an ordered set of the unlabeled samples based on their information content. We then compute the final objective value $S^{\text{new}}(B)$ for every possible batch size by subtracting the cost term $\lambda_3 * |B|$ from the corresponding score $S(B)$. The maximal value of $S^{\text{new}}(B)$ represents the desired batch size $|B|$ and the desired set of points in the set B. The pseudocode is presented in Algorithm 2.

Similar to the previous formulation, solving a single optimization problem yields the size and the samples to be selected for batch query. The time complexity is $O(n^2)$ (obtained from lines 3 and 4 in the algorithm), similar to the state-of-the-art static BMAL algorithms, where n is the number of unlabeled instances.

C. Using the Proposed Frameworks for Static BMAL

It is to be noted that the proposed frameworks can be used for BMAL in cases where the batch size is specified. If the batch size is fixed, there is no need to balance the computation cost against the classification performance. Thus, the penalty terms from the objective functions are dropped and a constraint is imposed on the batch size. For example, for the gradient-descent-based method, the following problem is solved for static BMAL with batch size m:

$$\max_M \sum_{j \in U_t} \rho_j M_j - \lambda_1 \sum_{j \in U_t} (1 - M_j) S(y|x_j, w^{t+1})$$

show that

$$0 \leq M_j \leq 1 \quad \forall j \quad \text{and} \quad \sum_{j=1}^{|U_t|} M_j = m.$$

An analogous strategy is applied for static BMAL using the submodular optimization framework

$$\max_{B \subseteq U_t : |B| = m} S(B).$$

To achieve this, the outer loop in Algorithm 2 is run from 1 to the desired batch size m and the set B returns the optimum set of points after the loop ends on line 13.

IV. Experiments and Results

We conducted extensive experiments to study the efficacy of the proposed dynamic BMAL algorithms. Due to its wide usage and the need for BMAL in face recognition from video streams, we focus on face-based biometric recognition as the exemplar application in this paper. The cost parameters (λ_2 for the SGD algorithm and λ_3 for the submodularity-based framework) were selected to be one in our initial set of experiments, and we study the effect of these parameters later in this section empirically. The other weight parameters were selected to be 1 using cross validation. Gaussian mixture models were used as the classifier in our experiments because of their success in face recognition [29]. The parameters of each Gaussian were trained using the expectation–maximization algorithm [30]. A Gaussian kernel with parameter one was used to compute the kernelized distances. For the quasi-Newton solution, a stopping threshold of 10^{-4} was used and a threshold of 200 was set on the number of iterations. Our experiments, however, revealed that the latter threshold was never met and the algorithm terminated in less than 10 iterations for most experiments, based on the objective value threshold. The algorithms were implemented in MATLAB on a quad-core Intel processor with 2.66-GHz CPU and 8-GB RAM.

Our experiments are structured as follows.

1) Experiment 1 studies the overall objective of this paper, i.e., it studies the accuracy obtained on an independent test set using the proposed dynamic BMAL algorithms, as compared against the use of the existing static

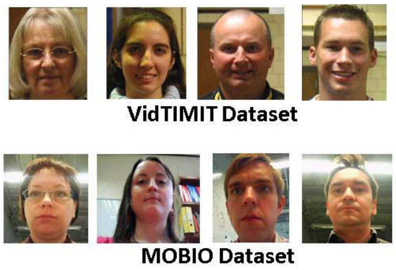


Fig. 1. Sample images from VidTIMIT and MOBIO data sets.

BMAL algorithms (with a prespecified batch size). Considering that static BMAL algorithms can perform differently with different prespecified batch sizes, we also compare the performances by changing the batch size in static BMAL methods in Experiment 1 (described further in Section 6.4 in the supplemental file).

2) Experiment 2 focuses on the *static* component of the proposed methods, by studying the effectiveness of the criteria chosen in the objective functions, given a specific batch size.
3) Experiment 3 systematically investigates the *dynamic* nature of the proposed BMAL algorithms, by studying the batch sizes identified using these methods on video streams with varying degrees of unknown data.
4) Experiment 4 studies the effect of the cost parameters involved in this paper.
5) Experiment 5 studies the quality of the approximations obtained using the proposed optimization strategies, as against the optimal solutions obtained using an exhaustive search for both the methods.
6) Finally, the SGD-based method (first of the two proposed methods) relies on the GTAM to obtain the future learner, w^{t+1}. Hence, Experiment 6 studies the effectiveness of the GTAM algorithm in assigning labels to unlabeled data, to ensure that the use of this approach is justified.

A. Data Sets and Feature Extraction

We used two challenging biometric data sets for our experiments: 1) the VidTIMIT data set [31], which contains video recordings of subjects reciting short sentences under unconstrained natural conditions and 2) the mobile biometric (MOBIO) data set [32], which was recently created for the MOBIO challenge to test state-of-the-art face and speech recognition algorithms. It contains recordings of subjects under challenging real-world conditions, captured using a hand-held device. Sample images from these data sets are shown in Fig. 1. The face images in the video frames were automatically detected using the Viola–Jones algorithm [33] and cropped to 128 by 128. The discrete cosine transform feature was used in all our experiments (for details about the feature extraction process, please refer [34]).

B. Experiment 1: Dynamic Versus Static BMAL

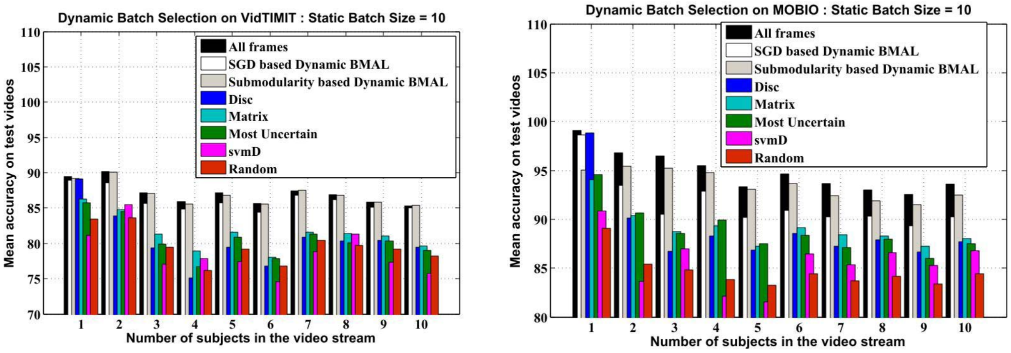
As mentioned earlier, the objective of this experiment is to study the performance of the proposed dynamic BMAL methods, as against the existing static BMAL methods with a prespecified batch size, on the task of face recognition. We selected 25 subjects at random from each data set. A classifier was induced with 250 training images (10 images from each subject). Unlabeled video streams (each containing 100 frames) were then presented to the learner. To vary the complexity of the task, the number of subjects in each unlabeled stream was varied between 1 and 10 (selected randomly from the set of 25). For each stream, the size and samples in a batch were selected simultaneously using the proposed methods. The classifier was updated with the images selected using the dynamic or static BMAL method, and tested on independent test videos [containing the same subject(s) as in the unlabeled videos].

The accuracy of the proposed techniques was compared against the case when all the frames in the unlabeled video were used for learning (this is assumed to be an estimate for the best achievable performance, as there is no better way to quantify the same for a given video stream), and also against the following static BMAL algorithms.

1) *Disc*, a discriminative BMAL strategy, proposed in [5].
2) *Matrix* that queries a batch of data samples by maximizing the mutual information between the labeled and unlabeled sets [6].
3) *Most Uncertain*, where the top k uncertain points were queried from the unlabeled video, k being the batch size.
4) *svmD* that incorporates diversity in active learning using SVMs, as proposed in [15].
5) *Random*, where a batch of points is queried at random. The Disc and the Matrix approaches have been shown to be the state-of-the-art BMAL techniques [6].

When the batch size is fixed at 10 for the static BMAL methods, the results are shown in Fig. 2 (averaged over 10 trials). The x-axis denotes the number of subjects in the video stream and the y-axis denotes the accuracy on test videos containing the corresponding number of subjects. We observe that, in both data sets, the accuracy obtained with the proposed methods matches the best achievable accuracy (when all images are used for training) more closely than any of the static BMAL algorithms.

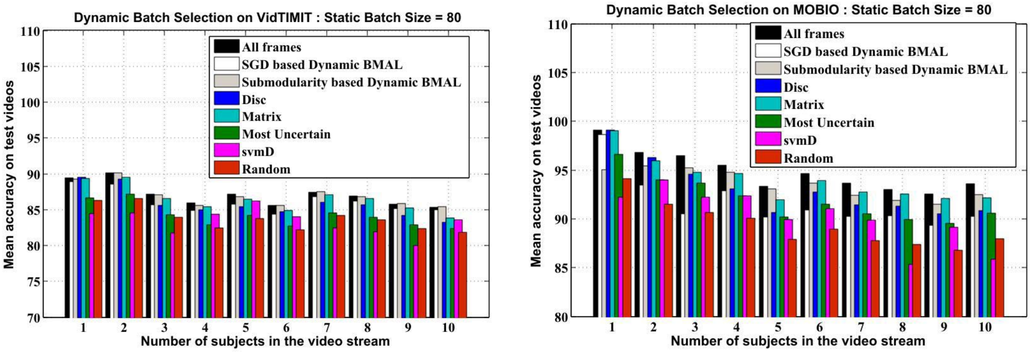
In general, we expect that if we select a greater number of images from an unlabeled set, the updated learner will perform better on a test set containing the same subjects. Thus, if we select a higher value of batch size in a static BMAL learner, its performance is expected to improve. This is studied in Fig. 3, where the static batch size was taken as 80 instead of 10. We see that the static BMAL schemes perform much better than before and the best static BMAL techniques marginally outweigh dynamic batch selection in terms of classification accuracy. However, to achieve this performance, static BMAL methods required a significantly greater number of images to be labeled than dynamic selection. Table I shows the mean predicted batch size (PBS) and mean percentage reduction in the number of images that had to be labeled using SGD optimization-based dynamic selection against static selection with batch size 80. Evidently, the static framework required a much greater number of images to be labeled to marginally outweigh dynamic selection. The same conclusion



(a)

(b)

Fig. 2. Dynamic versus static BMAL on (a) VidTIMIT and (b) MOBIO data sets (static batch size = 10). Best viewed in color.



(a)

(b)

Fig. 3. Dynamic versus static BMAL on (a) VidTIMIT and (b) MOBIO data sets (static batch size = 80). Best viewed in color.

TABLE I

MEAN PBS AND PERCENT LABELING COST REDUCTION (LCR) USING SGD-BASED DYNAMIC SELECTION AGAINST STATIC SELECTION WITH BATCH SIZE 80 ON A VIDEO STREAM WITH 100 FRAMES

Subjects	1	2	3	4	5	6	7	8	9	10
VidTIMIT (PBS)	18.7±1.06	12.4±0.96	17.9±1.12	25.6±1.17	16.5±0.82	24.2±0.98	22.3±1.09	22.5±0.96	19.9±1.13	24.5±1.03
VidTIMIT (LCR)	61.3±1.06%	67.6±0.96%	62.1±1.12%	54.4±1.17%	63.5±0.82%	55.8±0.98%	57.7±1.09%	57.5±0.96%	60.1±1.13%	55.5±1.03%
MOBIO (PBS)	15.5±1.17	12.2±1.05	12.7±1.09	15.1±0.92	10.9±1.11	10.9±1.16	10.4±0.89	9.9±0.97	11.7±1.14	11.6±0.91
MOBIO (LCR)	64.5±1.17%	67.8±1.05%	67.3±1.09%	64.9±0.92%	69.1±1.11%	69.1±1.16%	69.6±0.89%	70.1±0.97%	68.3±1.14%	68.4±0.91%

TABLE II

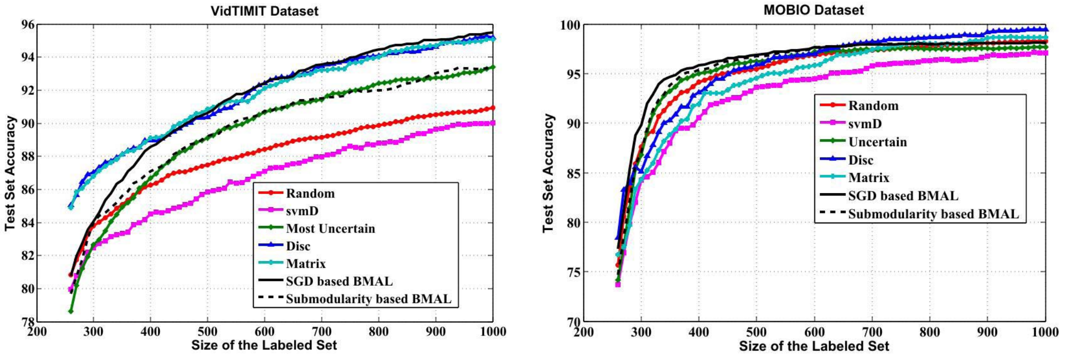
MEAN PBS AND PERCENT LCR USING SUBMODULARITY-BASED DYNAMIC SELECTION AGAINST STATIC SELECTION WITH BATCH SIZE 80 ON A VIDEO STREAM WITH 100 FRAMES

Subjects	1	2	3	4	5	6	7	8	9	10
VidTIMIT (PBS)	29.8±1.22	54.2±1.03	56.8±0.9	60.9±1.12	51.8±1.1	59.2±1.28	61.7±1.18	56.4±1.02	54.1±0.88	50.9±0.79
VidTIMIT (LCR)	50.2±1.22%	25.8±1.03%	23.2±0.9%	19.1±1.12%	28.2±1.1%	20.8±1.28%	18.3±1.18%	23.6±1.02%	25.9±0.88%	29.1±0.79%
MOBIO (PBS)	19.3±1.07	17.9±0.96	21.4±1.19	21.2±1.27	21.1±1.14	20.9±0.97	22.0±1.26	18.4±1.08	21.8±1.28	21.4±0.84
MOBIO (LCR)	60.7±1.07%	62.1±0.96%	58.6±1.19%	58.8±1.27%	58.9±1.14%	59.1±0.97%	58.0±1.26%	61.6±1.08%	58.2±1.28%	58.6±0.84%

is reflected in Table II that contains the analogous values for the submodular optimization-based dynamic BMAL framework. We infer that using a prespecified batch size, the static batch selection strategies can sometimes query too few points leading to poor generalization power of the updated learner, while in some cases, it can entail considerable labeling cost to attain a marginal improvement in accuracy. The proposed dynamic methods, on the other hand, strike a balance between the uncertainty of the learner on the images in the unlabeled video and the cost of labeling the images, and thus provide a more concrete basis to decide the size and samples in the batch.

C. Experiment 2: Performance of the Proposed Batch Selection Criteria for Given Batch Size

The purpose of this experiment was to analyze the effectiveness of the batch selection criteria of BMAL algorithms



(a)

(b)

Fig. 4. BMAL on (a) VidTIMIT and (b) MOBIO data sets. Best viewed in color.

given a batch size (static component) to study their usefulness in real-world settings. The static versions of the proposed algorithms were used, as described in Section III-C. Similar to Experiment 1, the proposed approaches were compared against the two state-of-the-art BMAL techniques: 1) Disc and 2) Matrix, and the three heuristic techniques: 1) Most Uncertain, 2) svmD, and 3) Random. A classifier was induced with 250 training images (10 from each of 25 randomly chosen subjects). Unlabeled video streams (each containing about 250 frames) were then presented to the classifier sequentially. The images in the video streams were randomly chosen from all 25 subjects and did not have any particular proportion of subjects in them. A batch of 10 images was queried from each video stream (that is, the batch size was fixed at 10 for each unlabeled video). After each batch selection, the selected images were appended to the training set, the classifier updated, and then tested on an independent test video containing about 5000 images spanning all the 25 subjects. We studied the accuracies on the test set with increasing sizes of the training set. The results (averaged over five runs) are shown in Fig. 4, where the x-axis denotes the size of the labeled set and the y-axis denotes the accuracy on the test set.

It is evident that the proposed SGD and submodularity-based techniques perform much better than svmD and Random sampling. The Most Uncertain method shows the best performance among the heuristic techniques. The proposed algorithms perform comparably with Disc and Matrix, which are the state-of-the-art static BMAL schemes (they marginally outperform Matrix on the MOBIO data set). We infer that our choice of the objective function performs comparably with the existing state-of-the-art methods, even in static settings. We also note that the SGD-based scheme performs better than the submodular BMAL technique for both data sets (in the static setting explored in this experiment). This can be attributed to the fact that the SGD-based strategy selects unlabeled points for manual annotation by optimizing the performance with respect to the future learner (the learner trained on the current training set together with the newly selected batch); it, therefore, is more effective in choosing the set of points that furnish maximal information. The submodular technique, on the other hand, uses the uncertainty of the current model together with a redundancy-based batch selection criterion and does not involve a look-ahead strategy using the future learner.

TABLE III

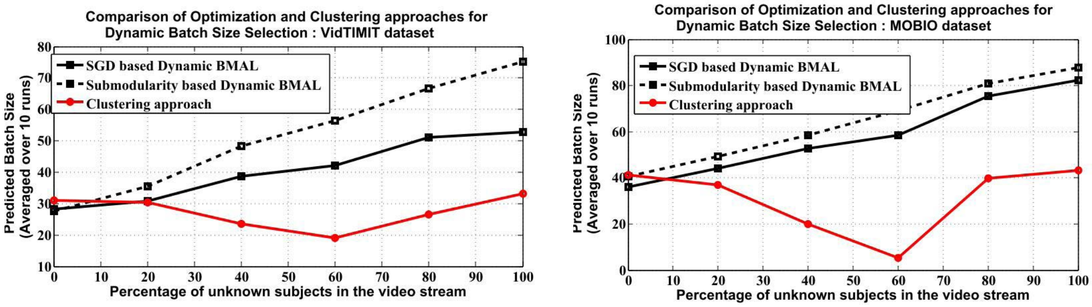
AVERAGE TIME TAKEN (IN SECONDS) TO QUERY A BATCH OF 10 IMAGES FROM AN UNLABELED VIDEO WITH 250 IMAGES

	VidTIMIT	MOBIO
SGD Dynamic BMAL	71.02 ± 2.35	82.45 ± 3.77
Submodular Dynamic BMAL	3.03 ± 0.92	5.27 ± 0.73
Disc	112.78 ± 3.08	122.45 ± 3.62
Matrix	38.22 ± 2.21	36.79 ± 3.38
svmD	1.05 ± 0.22	1.17 ± 0.79
Most Uncertain	1.23 ± 0.56	1.92 ± 0.74
Random	0.01 ± 0.0	0.01 ± 0.0

Table III reports the computation time comparison of the algorithms. We note that for both the SGD and the submodularity-based algorithms, the complexity is $O(n^2)$; however, this complexity merely depicts the pattern of growth in the running time of the algorithms with increasing size of the data set, that is, the runtime of both the algorithms grow quadratically with the size of the unlabeled set n. The actual runtime of the submodularity-based method is much lesser than the SGD method, as evident from Table III. This is because the SGD-based BMAL strategy involves classifier retraining in each iteration (due to the involvement of the future learner). The submodular framework, on the other hand, is solved using a greedy algorithm (and is devoid of model retraining) and involves much lesser computational overhead, as depicted in the runtime values. Thus, depending on the requirements of a particular application, an appropriate scheme can be adopted. While the heuristic techniques (svmD, Random, and Most Uncertain) depict promising running time values, their active learning performances are worse than those of the proposed algorithms (Fig. 4).

D. Experiment 3: Performance of Dynamic BMAL With Varying Complexities of Video Streams

In real-world settings, video streams can have varying levels of complexities, in terms of the presence of unknown subjects (not present in the training set), unknown expressions,



(a)

(b)

Fig. 5. Study of the proposed dynamic batch selection frameworks with varying complexities of a video stream. Experiment with unknown subjects from (a) VidTIMIT and (b) MOBIO data sets.

TABLE IV
TEST SET ACCURACIES USING THE PROPOSED AND CLUSTERING-BASED DYNAMIC BMAL ON THE VidTIMIT DATA SET WITH INCREASING PROPORTIONS OF NEW IDENTITIES

Proportion of new identities	0%	20%	40%	60%	80%	100%
SGD Dynamic BMAL	96.5 ± 0.28%	89 ± 0.34%	82 ± 0.26%	75 ± 0.33%	87.1 ± 0.17%	81.3 ± 0.31%
Submodular Dynamic BMAL	95.1 ± 0.22%	92.2 ± 0.35%	91.9 ± 0.37%	89 ± 0.24%	88.9 ± 0.18%	89.5 ± 0.3%
Clustering approach	95.9 ± 0.36%	85.6 ± 0.41%	81.4 ± 0.29%	70.6 ± 0.33%	79.7 ± 0.4%	74.6 ± 0.32%

TABLE V
TEST SET ACCURACIES USING THE PROPOSED AND CLUSTERING-BASED DYNAMIC BMAL ON THE MOBIO DATA SET WITH INCREASING PROPORTIONS OF NEW IDENTITIES

Proportion of new identities	0%	20%	40%	60%	80%	100%
SGD Dynamic BMAL	86 ± 0.2%	73.5 ± 0.22%	75.7 ± 0.15%	78.3 ± 0.31%	83.3 ± 0.19%	87.4 ± 0.07%
Submodular Dynamic BMAL	87.1 ± 0.23%	79.9 ± 0.29%	81.6 ± 0.38%	85.3 ± 0.25%	86.7 ± 0.37%	90.5 ± 0.18%
Clustering approach	72.23 ± 0.48%	68.17 ± 0.42%	53.28 ± 0.33%	57.54 ± 0.32%	55.89 ± 0.45%	56.19 ± 0.24%

head poses, and changing illumination among others. This experiment studies the *dynamic* component of the proposed BMAL methods, by observing the computed batch sizes with varying complexities of a video stream. In [35], we had proposed a heuristic clustering methodology for dynamic batch selection. This algorithm segregates the images in the unlabeled pool into separate clusters using the DBSCAN clustering algorithm, and then uses a heuristic score based on the Silhouette coefficient of each cluster to decide the batch size. Since no other dynamic batch selection strategies have been proposed till date (for comparison), we compared our approaches against this heuristic scheme.

Twenty-five subjects from each data set were selected and divided into two groups: 1) a known group containing 20 subjects and 2) an unknown group containing the remaining five subjects. A classifier was induced, as before, with 10 training images of each of the known subjects. Unlabeled video streams were then presented to the learner, with the proportion of unknown subjects in the unlabeled video gradually increased from 0% (where all the subjects in the unlabeled video were from the training set) to 100% (where none of the subjects in the unlabeled video were present in the training set) in steps of 20%. Thus, the classifier was exposed to video streams of varying levels of new information. However, the learner was not given any information about the composition of the video streams. In addition, the size of each video stream was kept the same (approximately 100 frames) to facilitate fair comparison.

The results (averaged over 10 trials) are shown in Fig. 5. The x-axis denotes the percentage of atypical images in the unlabeled pool, and the y-axis denotes the batch size predicted using both the proposed and clustering-based strategies. We note that in both the experiments, as the proportion of salient images in the unlabeled stream increases, the uncertainty term outweighs the annotation cost term in the objective functions and the proposed algorithms decide on a larger batch size. This matches our intuition because, with growing percentages of atypical images in the video stream, the confidence of the learner on those images decreases, and thus it needs to query more images to attain good generalization capability. The clustering-based scheme, on the other hand, does not consider the training set and fails to reflect the uncertainty of the classifier. The batch size, therefore, does not bear any specific trend to the percentage of atypical unlabeled images. We infer that the proposed optimization-based techniques provide a more sound basis to adaptively decide the batch size by considering the data typicalness with respect to the training set together with the labeling cost.

Besides the PBS, it is equally important to analyze the accuracy obtained on test sets with similar compositions as the unlabeled videos. In the case of the clustering technique, the gradient-descent-based approach (Section III-C) was used

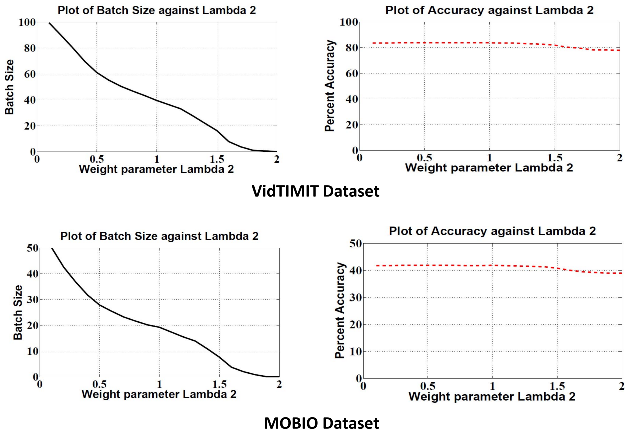


Fig. 6. Effect of the cost parameter in the SGD-based dynamic BMAL.

for batch query once the batch size was determined [35]. Tables IV and V show the accuracies obtained on test videos from the VidTIMIT and MOBIO data sets using the optimization and clustering-based strategies. While the optimization-based techniques consistently deliver high-accuracy values on test videos, the accuracy obtained from the clustering scheme is erratic and inconsistent with varying proportions of new identities in the unlabeled stream. This is more accentuated in the MOBIO data set. We also note that the submodularity-based technique depicts better accuracy than the SGD-based method for both the data sets. However, a comparison of the two dynamic BMAL techniques will not be fair here, as their selected batch sizes are different (evident from Fig. 5), unlike the previous experiment, where the batch size was kept constant to facilitate fair comparison. The important thing to note in this experiment is the fact that for both the proposed algorithms, the PBS appropriately reflects the complexity of the data.

E. Experiment 4: Effect of Cost Parameter

In the experiments described above, the cost parameter (λ_2 for the SGD-based method and λ_3 for the submodularity-based method) was set as 1. Here, we study the effect of this parameter on batch size and accuracy. As in the previous experiment, the training set consisted of 250 images and the test set had 5000 images spanning all subjects. An unlabeled video stream (with 250 frames) was then presented for dynamic batch selection, and the selected images were appended to the training set (note that in this case, we are not interested in studying the growth in accuracy with increasing size of the training set; hence, we focus on the accuracy obtained after a single round of dynamic batch selection from an unlabeled video).

Fig. 6 shows the results (averaged over 20 different unlabeled video streams) of the SGD-based algorithm, where the weight parameter λ_2 was varied between 0.1 and 2 (for $\lambda_2 > 2$, the learner did not select any image in the batch). We note that an increase in the cost parameter value leads to a reduction of the PBS and also the generalization accuracy. This corroborates our intuition as an increase in the labeling cost per sample restricts the number of unlabeled samples that can be purchased for labeling, which also degrades the accuracy on the same test set. Our observation revealed that the difference in accuracy for $\lambda_2 = 0.1$ and $\lambda_2 = 2$ was about 7%. A similar result was obtained for the cost parameter, λ_3, in the submodularity-based algorithm (the results not presented due to space constraints).

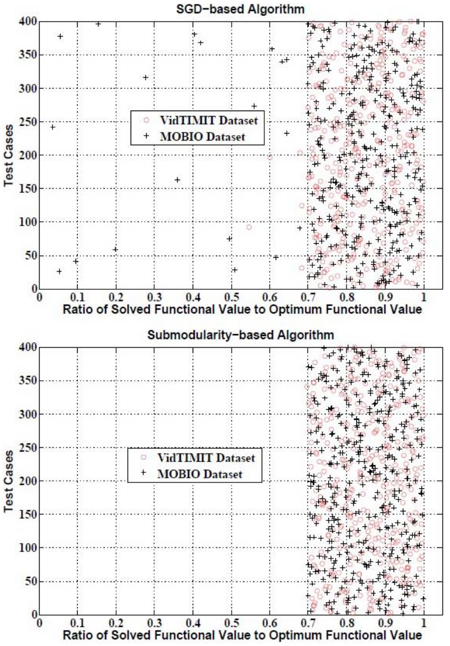


Fig. 7. Validation of solution quality for the SGD and submodularity-based dynamic BMAL.

F. Experiment 5: Quality of Optimization Solutions

To solve the SGD-based optimization problem, the integer constraints in (8) were relaxed into continuous constraints in (9). Similarly, for the submodularity-based approach, a greedy algorithm was used to solve the dynamic batch selection problem in (14). Both these strategies lead to suboptimal solutions, and it is important to study the quality of the solutions obtained from the relaxations. To this end, 400 random unlabeled video streams were taken from the VidTIMIT and the MOBIO data sets and the relaxed batch selection algorithms were applied for dynamic batch selection. In addition, an exhaustive search was performed to find the best solution for a given unlabeled stream by brute force. The ratio $(f(\widehat{x})/f(x^*))$ was computed for the 400 random samples, where $\widehat{x}$ is the solution obtained after relaxation, x^* is the optimal solution obtained by a brute-force search, and f is the objective function to be maximized [(8) for the SGD-based approach and (14) for the submodularity algorithm].

The results are shown in Fig. 7, and depict the fact that the aforementioned ratio is very close to 1 (greater than 0.8 for most of the test cases). Thus, the functional value attained by solving the relaxation is very close to the optimal functional value. The results lead to the conclusion that both the relaxations produce high-quality solutions of the corresponding optimization problems. However, we also note

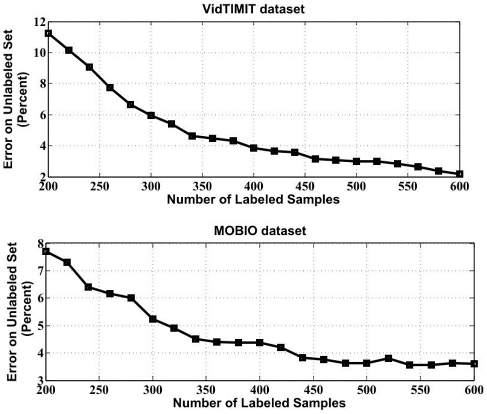


Fig. 8. Validation of the efficacy of GTAM.

that for the MOBIO data set, the SGD algorithm sometimes yielded poor solutions (where the ratio is less than 0.3); this is mostly because of bad starting points of the gradient-descent algorithm, which led to bad local optima.

G. Experiment 6: Effectiveness of GTAM Algorithm for Label Prediction

In this experiment, we validate the efficacy of the graph-based transductive algorithm (GTAM) [26] in assigning labels to the current batch of unlabeled samples to estimate the future classifier w^{t+1} in the iterations of the SGD algorithm. The performance of GTAM was studied on a test set of 1000 images (from the VidTIMIT and MOBIO data sets) with different sizes of the training set ranging from 200 to 600. The results are reported in Fig. 8, which plots the test error against different training set sizes. We note that with only 200 labeled samples, the GTAM algorithm produces a generalization error of about 10% and it reduces further with increasing sizes of the labeled set. This corroborates our choice of the GTAM algorithm for assigning labels to unlabeled samples, thus providing a good approximation of the future classifier w^{t+1} in the quasi-Newton iterations of the SGD-based dynamic BMAL algorithm.

V. Conclusion

In this paper, we proposed two novel approaches of dynamic BMAL, which adaptively select the batch size and the specific data samples for manual annotation based on the complexity of a data stream and the cost of annotation of each unlabeled data sample. Unlike the previously proposed BMAL methods, which need the batch size as an input, our framework incorporates the labeling cost in the batch selection criterion and computes the batch size automatically. The batch size and selection criteria are integrated into a single optimization formulation, whose solution yields the desired batch size and the specific samples for query. The frameworks were validated on the face recognition application using two challenging biometric data sets. Our results corroborated the effectiveness of the approaches against static BMAL in terms of dynamically identifying the batch size for a given data stream based on its complexity level and the labeling cost of the images. The proposed algorithms also depicted comparable performance against the state-of-the-art static BMAL techniques, when the batch size was prespecified. We further note that for a given batch size, the gradient-descent-based scheme has a better label complexity than the submodularity approach, but the latter outweighs the former in terms of computation time. Thus, based on the requirements of a given application, an appropriate technique can be selected. Moreover, the algorithms are flexible and the specific terms in the objective function can be modified based on the requirements of a particular application. We also empirically established that our algorithms yield high-quality solutions of the relaxations of the corresponding NP-hard problems. In general, the proposed methods work well when it is not easy to identify a batch size in an application setting, or when there is variation expected within a single video, resulting in the need for dynamic batch size selection.

The proposed frameworks can also be used in problems where multiple sources of information are available, such as both face and speech data of an individual or multiple image features extracted from a given face image. Learning from multiple sources can be superior to learning from a single source, if the sources are used appropriately [36]. Let U_{t1} and U_{t2} denote the unlabeled data streams from two sources of information. The objective functions can then be modified by adding relevant terms from the two sources, together with a penalty on batch size. In the case of the SGD-based method, the following criterion can be used for dynamic BMAL:

$$\max_M \sum_{j\in U_{t1}} \rho_j M_j - \sum_{j\in U_{t1}} (1-M_j)S(y|x_j, w^{t+1}) + \sum_{j\in U_{t2}} \rho_j M_j - \sum_{j\in U_{t2}} (1-M_j)S(y|x_j, w^{t+1}) - \sum_j M_j.$$

This can be solved as before using the quasi-Newton method. Furthermore, let x_{1i} and x_{2i} denote the feature representations from the two sources of information, and let E_1, D_1, and E_2, D_2 be the entropy and the distance functions for the two sources respectively, as defined in Section III-B. The submodular technique can be adapted for dynamic batch selection from two sources using the following score function:

$$S^{\text{new}}(B) = \sum_{x_{1i}\in B} \{\rho(x_{1i}) + E(x_{1i}) + D(x_{1i})\} + \sum_{x_{2i}\in B} \{\rho(x_{2i}) + E(x_{2i}) + D(x_{2i})\} - |B|.$$

This can be solved in an analogous way as Algorithm 2, where the submodular and nondecreasing score function is obtained by removing the penalty term $|B|$ from $S^{\text{new}}(B)$.

Moreover, if contextual information is available (e.g., location of a subject, at home or in office), the same approach can be used to construct a prior probability vector depicting the chances of seeing particular acquaintances in a given context. The entropy term can then be computed on the posterior probabilities obtained by multiplying the likelihood values returned by the classifier with the context aware prior. Thus, subjects not expected in a given context (e.g., a home acquaintance in an office setting) will have low priors, and consequently, the corresponding posteriors will not contribute much in the entropy calculation. The frameworks can therefore be extended to context-aware adaptive batch selection.

Furthermore, real-world problems often have variable labeling costs, where the cost of annotating each unlabeled sample is different. For instance, consider a voicemail classification application, where the objective is to classify voice messages as urgent/nonurgent. To label a data sample in such an application, the human oracle has to listen to the entire message. Thus, it is natural for shorter messages to have a lower labeling cost as compared with longer messages. To address such a problem, the labeling cost terms (λ_2 for the SGD method and λ_3 for the submodular framework) can be extended to vectors, of dimension same as the number of unlabeled instances, where each entry denotes the labeling cost of the corresponding unlabeled data sample. The same algorithms can then be used to solve for the batch size and the unlabeled samples to be annotated. Thus, the proposed dynamic batch selection frameworks can also be applied to problems with variable labeling costs.

A potential limitation of this framework is the selection of the cost parameter. A low value of this parameter results in a high permissible batch size and, consequently, a high accuracy and vice versa. It is thus a property of the system running the application and cannot be tuned to a particular data set. Converting the storage and labeling resources of a system to the same currency as the entropy and diversity terms (to derive the cost coefficient) may be a challenge.

As part of future work, we will explore other mechanisms of dynamic batch size computation (e.g., L_2 regularization as the penalty term). The problem of adaptive batch selection is closely related to finding the correct number of clusters in a clustering algorithm; recent work has addressed this problem using Dirichlet processes [37], [38], which we plan to investigate in our ongoing work. Our future work will also focus on deriving performance guarantees on the solution qualities for both the dynamic BMAL schemes. Furthermore, in the case of the SGD-based approach, the quadratic programming problem that needs to be solved as part of the optimization process can significantly increase the computation time (especially for large scale data). There have been recent efforts [39] to efficiently solve QP problems using a pivoting algorithm and the KKT conditions to significantly reduce computations. This can be judiciously used in our approach, making it meritorious even for large-scale data. We will explore this in our future work. Future work will also include designing a proper user interface for adaptive BMAL.

References

[1] A. Khosla, R. Hamid, C.-J. Lin, and N. Sundaresan, "Large-scale video summarization using web-image priors," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2013, pp. 2698–2705.

[2] S. C. H. Hoi, R. Jin, and M. R. Lyu, "Batch mode active learning with applications to text categorization and image retrieval," *IEEE Trans. Knowl. Data Eng.*, vol. 21, no. 9, pp. 1233–1248, Sep. 2009.

[3] S. C. H. Hoi, R. Jin, J. Zhu, and M. R. Lyu, "Batch mode active learning and its application to medical image classification," in *Proc. 23rd Int. Conf. Mach. Learn. (ICML)*, 2006, pp. 417–424.

[4] S. C. H. Hoi, R. Jin, and M. R. Lyu, "Large-scale text categorization by batch mode active learning," in *Proc. 15th Int. Conf. World Wide Web*, 2006, pp. 633–642.

[5] Y. Guo and D. Schuurmans, "Discriminative batch mode active learning," in *Advances of Neural Information Processing Systems (NIPS)*. Cambridge, MA, USA: MIT Press, 2007.

[6] Y. Guo, "Active instance sampling via matrix partition," in *Advances of Neural Information Processing Systems (NIPS)*. Cambridge, MA, USA: MIT Press, 2010.

[7] B. Settles, "Active learning literature survey," Dept. Comput. Sci., Univ. Wisconsin-Madison, Madison, WI, USA, Tech. Rep. 1648, 2010.

[8] Y. Baram, R. El Yaniv, and K. Luz, "Online choice of active learning algorithms," *J. Mach. Learn. Res.*, vol. 5, pp. 255–291, Mar. 2004.

[9] S. Tong and D. Koller, "Support vector machine active learning with applications to text classification," *J. Mach. Learn. Res.*, vol. 2, pp. 45–66, Mar. 2002.

[10] D. A. Cohn, Z. Ghahramani, and M. I. Jordan, "Active learning with statistical models," *J. Artif. Intell. Res.*, vol. 4, no. 1, pp. 129–145, 1996.

[11] A. Beygelzimer, S. Dasgupta, and J. Langford, "Importance weighted active learning," in *Proc. 26th Annu. Int. Conf. Mach. Learn. (ICML)*, 2009, pp. 49–56.

[12] Y. Freund, H. S. Seung, E. Shamir, and N. Tishby, "Selective sampling using the query by committee algorithm," *Mach. Learn.*, vol. 28, nos. 2–3, pp. 133–168, 1997.

[13] R. Liere and P. Tadepalli, "Active learning with committees for text categorization," in *Proc. Nat. Conf. Artif. Intell.*, 1997, pp. 591–596.

[14] Y. Guo and R. Greiner, "Optimistic active learning using mutual information," in *Proc. 20th Int. Joint Conf. Artif. Intell. (IJCAI)*, 2007, pp. 823–829.

[15] K. Brinker, "Incorporating diversity in active learning with support vector machines," in *Proc. 20th Int. Conf. Mach. Learn. (ICML)*, 2003, pp. 59–66.

[16] G. Schohn and D. Cohn, "Less is more: Active learning with support vector machines," in *Proc. 17th Int. Conf. Mach. Learn. (ICML)*, 2000, pp. 839–846.

[17] Z. Xu, K. Yu, V. Tresp, X. Xu, and J. Wang, "Representative sampling for text classification using support vector machines," in *Proc. Eur. Conf. Inf. Retrieval*, 2003, pp. 393–407.

[18] S. C. H. Hoi, R. Jin, J. Zhu, and M. R. Lyu, "Semi-supervised SVM batch mode active learning for image retrieval," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2008, pp. 1–7.

[19] S. C. H. Hoi, R. Jin, J. Zhu, and M. R. Lyu, "Semisupervised SVM batch mode active learning with applications to image retrieval," *ACM Trans. Inf. Syst.*, vol. 27, no. 3, 2009, Art. ID 16.

[20] Y. Grandvalet and Y. Bengio, "Semi-supervised learning by entropy minimization," in *Advances of Neural Information Processing Systems (NIPS)*. Cambridge, MA, USA: MIT Press, 2005.

[21] D. Shen, J. Zhang, J. Su, G. Zhou, and C.-L. Tan, "Multi-criteria-based active learning for named entity recognition," in *Proc. 42nd Annu. Meeting Assoc. Comput. Linguist. (ACL)*, 2004, Art. ID 589.

[22] A. Gretton, K. Borgwardt, M. Rasch, B. Schölkopf, and A. Smola, "A kernel method for the two-sample problem," in *Advances of Neural Information Processing Systems (NIPS)*. Cambridge, MA, USA: MIT Press, 2007.

[23] K. M. Borgwardt, A. Gretton, M. J. Rasch, H.-P. Kriegel, B. Schölkopf, and A. J. Smola, "Integrating structured biological data by kernel maximum mean discrepancy," *Bioinformatics*, vol. 22, no. 14, pp. e49–e57, 2006.

[24] J. Weston, A. Elisseeff, B. Schölkopf, and M. Tipping, "Use of the zero norm with linear models and kernel methods," *J. Mach. Learn. Res.*, vol. 3, pp. 1439–1461, Mar. 2003.

[25] J. Nocedal and S. J. Wright, *Numerical Optimization*. New York, NY, USA: Springer-Verlag, 1999.

[26] J. Wang, T. Jebara, and S. Chang, "Graph transduction via alternating minimization," in *Proc. 25th Annu. Int. Conf. Mach. Learn. (ICML)*, 2008, pp. 1144–1151.

[27] G. L. Nemhauser, L. A. Wolsey, and M. L. Fisher, "An analysis of approximations for maximizing submodular set functions—I," *Math. Programming*, vol. 14, no. 1, pp. 265–294, 1978.

[28] A. Krause and C. Guestrin, "Near-optimal nonmyopic value of information in graphical models," in *Proc. 21st Conf. Uncertainty Artif. Intell. (UAI)*, 2005, p. 5.

[29] J. Kim, D. Y. Ko, and S. Y. Na, "Implementation and enhancement of GMM face recognition systems using flatness measure," in *Proc. 13th IEEE Int. Workshop Robot Human Interact. Commun.*, Sep. 2004, pp. 247–251.

[30] C. M. Bishop, *Pattern Recognition and Machine Learning*, 1st ed. New York, NY, USA: Springer-Verlag, Oct. 2007.

[31] C. Sanderson, *Biometric Person Recognition: Face, Speech and Fusion*. Saarbrücken, Germany: VDM Verlag, Jun. 2008.

[32] S. Marcel, C. McCool, P. Matejka, T. Ahonen, and J. Cernocky, "Mobile biometry (MOBIO) face and speaker verification evaluation," Idiap Research Inst., Martigny, Switzerland, Tech. Rep. Idiap-RR-09-2010, 2010.

[33] P. Viola and M. Jones, "Rapid object detection using a boosted cascade of simple features," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Dec. 2001, pp. 511–518.

[34] H. K. Ekenel, M. Fischer, Q. Jin, and R. Stiefelhagen, "Multi-modal person identification in a smart environment," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2007, pp. 1–8.

[35] S. Chakraborty, V. Balasubramanian, and S. Panchanathan, "Dynamic batch size selection for batch mode active learning in biometrics," in *Proc. IEEE 9th Int. Conf. Mach. Learn. Appl. (ICMLA)*, Dec. 2010, pp. 15–22.

[36] K. Crammer, M. Kearns, and J. Wortman, "Learning from multiple sources," *J. Mach. Learn. Res.*, vol. 9, pp. 1757–1774, Jun. 2008.

[37] H. M. Wallach, S. T. Jensen, L. Dicker, and K. A. Heller, "An alternative prior process for nonparametric Bayesian clustering," in *Proc. Int. Conf. Artif. Intell. Statist. (AISTATS)*, 2010, pp. 1–8.

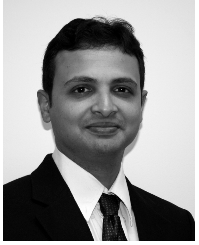
[38] G. Yu, R. Huang, and Z. Wang, "Document clustering via Dirichlet process mixture model with feature selection," in *Proc. ACM Conf. Knowl. Discovery Data Mining (KDD)*, 2010, pp. 763–772.

[39] Y. Liu and Z. Zhang, "A fast algorithm for linearly constrained quadratic programming problems with lower and upper bounds," in *Proc. Int. Conf. Multimedia Inf. Technol.*, Dec. 2008, pp. 58–61.

Shayok Chakraborty (M'08) received the Ph.D. degree in computer science from Arizona State University (ASU), Tempe, AZ, USA, in 2013.

He was with IBM India, Kolkata, India, as an Application Developer for one year until 2007. He was a Post-Doctoral Research Associate with Intel Labs, Hillsboro, OR, USA, from 2013 to 2014. He is currently a Post-Doctoral Researcher with the Department of Electrical and Computer Engineering, Carnegie Mellon University, Pittsburgh, PA, USA. His current research interests include machine learning, computer vision, data mining, and assistive technology. He has published his research in top-tier conferences and journals in computer vision, data mining, and machine learning.

Dr. Chakraborty has also delivered a tutorial on active learning at the 2013 International Conference on Multimedia and Expo. His Ph.D. dissertation on batch mode active learning was nominated for the Best Ph.D. Dissertation Award from the Department of Computer Science, ASU. He was a recipient of an Outstanding Graduate Student Award from the Graduate College at ASU, and was selected for a research internship with the Department of Machine Learning, Microsoft Research, Redmond, WA, USA.



Vineeth Balasubramanian (M'07) received the dual master's degree in mathematics and computer science from the Sri Sathya Sai Institute of Higher Learning, Anantapur, India, in 2001 and 2003, respectively.

He was with Oracle Corporation, for two years until 2005. Until 2013, he was an Assistant Research Professor with the Center for Cognitive Ubiquitous Computing, Arizona State University (ASU), Tempe, AZ, USA. He is currently an Assistant Professor with the Department of Computer Science and Engineering, IIT Hyderabad, Hyderabad, India. He has authored over 40 research publications in premier peer-reviewed venues, and holds three patents under review. His current research interests include pattern recognition, machine learning, computer vision, and multimedia computing with assistive and healthcare applications.

Dr. Balasubramanian is a member of the Association for Computing Machinery and the Association for the Advancement of Artificial Intelligence. His Ph.D. dissertation on the Conformal Predictions framework was nominated for the Outstanding Ph.D. Dissertation at the Department of Computer Science, ASU, in 2010. He was also awarded the Gold Medals for Academic Excellence in the bachelor's program in mathematics in 1999, and master's program in computer science in 2003. He also received Research Grants from the U.S. National Science Foundation.

Sethuraman Panchanathan (F'01) is currently the Senior Vice President of the Office of Knowledge Enterprise Development with Arizona State University (ASU), Tempe, AZ, USA, where he is also the Foundation Chair of Computing and Informatics and the Director of the Center for Cognitive Ubiquitous Computing. He was the Founding Director of the School of Computing and Informatics and was instrumental in founding the Biomedical Informatics department at ASU. He was the Chair of the Computer Science and Engineering department at ASU. He has authored over 430 research papers and articles. He has mentored over 125 graduate students, post-doctoral students, Research Engineers, and Research Scientists. His current research interests include ubiquitous computing environments for enhancing quality of life for individuals with disabilities, haptic user interfaces, face/gait analysis and recognition, medical image processing, media processor design, and human-centered multimedia computing.

Dr. Panchanathan is a member of the National Academy of Inventors and the Canadian National Academy of Engineering, and a fellow of the Society for Optics and Photonics. He was appointed to the National Science Board by President Obama in 2014.