Reliable Posterior Probability Estimation for Streaming Face Recognition

Abhijit Bendale
University of Colorado at Colorado Springs
abendale@vast.uccs.edu

Terrance Boult
University of Colorado at Colorado Springs
tboult@vast.uccs.edu

Abstract

Increasing access to large, non-stationary face datasets and corresponding demands to process , analyze and learn from this data. This has lead to a new class of online/incremental face recognition problems. While it is advantageous to build large scale learning systems when resources permit, a counter problem of learning with limited resources in presence of streaming data arises. We present a budgeted incremental support vector learning method suitable for online learning applications. Our system can process one sample at a time and is suitable when dealing with large streams of data. We discuss multiple budget maintenance strategies and investigate the problem of incremental unlearning. We propose a novel posterior probability estimation model based on Extreme Value Theory (EVT) and show its suitability for budgeted online learning applications (calibration with limited data). We perform thorough analysis of various probability calibration techniques with the help of methods inspired from meteorology. We test our methods on Labeled Faces in the Wild dataset and show suitability of the proposed approach for face verification/recognition

1 Introduction

The performance of a face recognition system relies heavily on its ability to handle spatial, temporal and operational variances [38]. Large scale face recognition systems have to adapt with changing environments. Face recognition systems often achieve excellent performance on benchmark datasets, but performance degrades significantly in operational environments [24]. One could build an application specific dataset, but this process is extremely cumbersome. Recently [22], [36] have adopted the approach to incrementally adapt learned models for face recognition with new data. As hardware (e.g. surveillance cameras) becomes cheaper, it is easier to capture more data to address problems such as self-occlusion, motion blur and illumination [10]. A typical image captured from a surveillance dataset is about 70 KB [4]. Analyzing a short video stream of about 10 minutes at 30fps amounts to processing of 1.2 GB

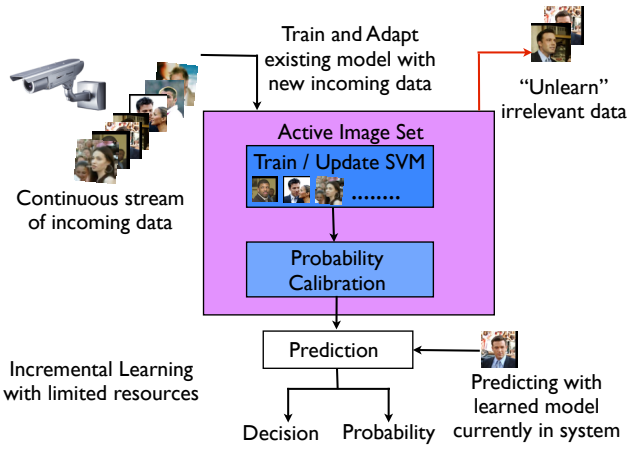

Figure 1. In streaming face recognition with limited resources, updating the existing system with incoming data, gradually adapting to variations and unlearning already learned samples, without compromising on accuracy can be extremely challenging. We present a system that can incrementally adapt to incoming samples and provide reliable posterior probability estimates.

of data. Updating existing learned models with such large and rapidly produced data poses significant challenge for a system with limited processing resources. While processing data on the cloud might be possible for some applications, many surveillance applications have constrained operational environments.

A common notion in many learning based vision systems is to learn with as much data as possible, to achieve best possible prediction performance during testing phase. This approach is useful for large scale face recognition when resources permit [35], [2], however poses multiple challenges when learning with limited resources [14]. The feature descriptors used for image representation are often high dimensional and data is generated faster than it can be processed/learned. Thus from practical application point of view it becomes important to handle not just every incoming sample, but all the "important" samples. This problem is further constrained in case of standalone or portable face

recognition systems [20]. Problem of online learning occurs in many consumer applications as well [15]. Updating the existing face recognition system with incoming data, gradually adapting to variations and possibly forgetting already learned samples can be extremely challenging. In this work, we study online learning on a fixed budget in the context of unconstrained face recognition. We consider a specific operational scenario where system is presented one sample at a time with a user-defined upper limit on maximum number of samples (budget size) it can retain in the memory.

We propose an incremental and online Support Vector Machine (SVM) learning system on a budget. Our system can process one example at a time and is based on work of Cauwenberghs et al [5]. This method maintains optimal SVM solution on all previously seen examples during training by incrementally updating Karush-Kuhn-Tucker (KKT) conditions. The system learns incoming data one sample at a time (Fig 1) until it reaches the maximum allowable budget size. In context of SVMs, the budget size implies maximum number of support vectors retained by the system. This set of support vectors is termed as active set. Once this size is reached, existing samples are incrementally removed from the active set. This process is called "unlearning".

In many face recognition applications, it is important to predict well calibrated probabilities [21], [28], [26] along with SVM decision scores (or class predictions). The problem of probability calibration is more pronounced for budgeted online learning, since calibration data is limited. Further, the calibration data changes regularly when model gets updated. While one can always re-calibrate with data present in the active set at a given time, a counter question about reliability of the calibration arises. Scheirer *et al*. [28] have demonstrated that the broader problem of recognition is consistent with the assumptions of statistical Extreme Value Theory (EVT). They show EVT provides a way to determine probabilities, regardless of the overall distribution of the data. They have shown the extremes or tail of a score distribution produced by a recognition/classification algorithm can always be modeled by an Extreme Value Theory (EVT) distribution. This distribution is shown to produce a reverse Weibull distribution when the data is bounded. This observation makes EVT a natural choice for probability calibration for budgeted online SVMs.

In the following sections, we discuss our modifications to the approach of [5] to make it suitable for budgeted online learning. We show via extensive experiments that our method works comparable (and often better) at really small budget sizes when compared with many off-the-shelf machine learning packages [9]. We develop EVT based calibration models for online budgeted learning for posterior probability estimation. We compare our method with a defacto standard in the community proposed by Platt [25]. We perform rigorous comparison of the proposed probability calibration techniques for extreme budget sizes to assess the reliability of estimated probabilities. Finally, we quantify our results by methods inspired from meteorology: reliability plots [34] and Brier Score analysis [19]. Our study suggests EVT calibration is well suited for online learning applications as it consistently yields more reliable probabilities. We test our methods on Labeled Faces in the Wild [12] dataset and show suitability of the proposed approach for large scale face verification/recognition. The contributions of this work can be summarized as follows:

1. A probability based budgeted incremental support vector learning and unlearning method.
2. A novel posterior probability estimation model based on EVT.
3. Analysis of posterior probability estimation models with limited calibration data.
4. Reliability analysis of various probability estimation techniques.

2 Related Work

Ozawa *et al*. [22] use incremental principal component analysis with resource allocating network with long term memory for constrained face recognition problem. Yan *et al*. [36] used incremental linear discriminant analysis with spectral regression. Their approach is suitable for incremental model adaptation but does not provide posterior probability estimates, as required in multiple applications.

Incremental Learning described in this work draws inspiration from work done by Cauwenberghs et al [5]. It maintains optimal SVM solution on all previously seen examples during training by incrementally updating Karush-Kuhn-Tucker (KKT) conditions. Wang et al [33] provide a thorough analysis of various budgeted online learning techniques in their work. In section 6, we compare our method with the ones mentioned in [33].

Posterior probability estimation from SVM decision scores is a well studied problem in computer vision and machine learning. Platt [25] proposed probability calibration for SVMs, which has also been applied to other learning algorithms. A comprehensive analysis of probability calibration techniques can be found in work of [21]. These methods were devised for batch supervised learning. They have been found to be effective when the entire training set is available for calibration. Calibration with limited data is a challenging problem as noted by Zadrozny et al [37] and [21]. Niculescu-Mizil et al [21] found that isotonic regression based calibration is prone to over-fitting. It performs worse than Platt scaling, when data is limited (less than 1000 samples). Majority of the online learning software packages [23], [29], [9] provide either a default Platt Scaling for posterior probability estimation or just decision score as output.

Recent work by Scheirer et al [28] has shown the extremes or *tail* of a score distribution produced by any recognition algorithm can always be modeled by an EVT distribution. These approaches were found to be useful for attribute fusion in image retrieval applications [26], scene classification in remote sensing applications [31] and biometric verification systems [27]. Probability calibration for budgeted online learning for SVMs presented in this work builds on top of the work of Scheirer *et al*. [28]. Reliability diagrams [34] are frequently used for assessing reliability of probability forecasts for binary events such as the probability of measurable precipitation in the area of weather forecasting. We introduce the tools such as reliability diagrams and Brier Scores [19] to assess the reliability of posterior probability estimation obtained by Platt calibration and EVT based calibration.

3 Incremental Support Vector Machines

Let us assume we have a set of training data $D = \{(x_i, y_i)\}_{i=1}^{k}$, where $x_i \in \mathcal{X} \subseteq \mathcal{R}^n$ is input and $y_i \in \{+1, -1\}$ is the output class label. Support Vector Machines learn the function $f(x) = w^T\phi(x) + b$, where $\phi(x)$ denotes a fixed feature space transformation. The dual formulation of this problem involves estimation of α_i, where α are the Lagrange multipliers associated with the constraints of the primal SVM problem. These coefficients are obtained by minimizing a convex quadratic objective function under the constraints

$$\min_{0\leq\alpha_i\leq C} : W = \frac{1}{2}\sum_{i,j}\alpha_i Q_{ij}\alpha_j - \sum_i \alpha_i + \sum_i y_i\alpha_i \quad (1)$$

where b is the bias (offset), Q_{ij} is the symmetric positive definite kernel matrix $Q_{ij} = y_i y_j K(x_i, x_j)$ and C is the nonnegative user-specified slack parameter that balances model complexity and loss of training data. The first order conditions on W reduce to the KKT conditions, from which following relationships are obtained:

$$\begin{aligned} y_i f(x_i) > 1 &\Rightarrow \alpha_i = 0 \\ y_i f(x_i) = 1 &\Rightarrow \alpha_i \in [0, C] \\ y_i f(x_i) < 1 &\Rightarrow \alpha_i = C \end{aligned} \quad (2)$$

and $\sum_{i=1}^{k} y_i\alpha_i = 0$. These conditions partition the training data into three discrete sets: margin support vectors ($\alpha_i \in [0, C]$), error support vectors ($\alpha_i = C$) and ignored vectors. Decision score for test sample x_t is obtained using $f(x) = w^T\phi(x) + b$ where

$$w = \sum_{i=1}^{l} y_i\alpha_i\phi(x_i) \quad (3)$$

and l is total number of support vectors (consisting of margin support vectors and error support vectors). This is the traditional batch learning problem for SVM [32].

The incremental extension for SVM was suggested by Cauwenberghs *et al*. [5]. In this method, the KKT conditions are preserved after each training iteration. For incremental training, when a new training sample (x_c, y_c) is presented to the system, the Lagrangian coefficients (α_c) corresponding to this sample and positive definite matrix Q_{ij} from the SVM currently in the memory undergo a small change Δ to ensure maintenance of optimal KKT condition (details of these increments can be found in [17], [5]).

4 SVM for Streaming Face Recognition

In this section, we describe how we extend the incremental SVM for streaming face recognition application. Throughout the course of this work, we consider the case of binary classification. As noted earlier, operational face recognition systems have to learn from a continuous stream of data. The incoming data is processed continuously till the user prescribed budget size B (i.e. Active Set) is reached. At every stage, the classifier is updated by methodology described in section 3. Once the prescribed budget size is reached (e.g. if the system runs out of memory), the process of "unlearning" starts. Two system design questions at this stage are: (i) How to select a sample from current active set to unlearn? (ii) How to update existing SVM solution?

The process of unlearning starts with determining a particular sample to unlearn. In the past, machine learning researchers have considered methods like randomly removing support vector [6], removing the oldest support vector [8] or removing support vector that yields smallest prediction error using leave one out error methodology [5]. Removing oldest sample from memory have found to be useful for applications that exploit temporal coherence [1] (eg. tracking). A related study in this area [33] analyzed multiple budget maintenance strategies. They concluded that removing random SV (support vector) or removing oldest SV from the current active set yields poor performance. They suggested removing support vector with smallest norm with respect to current SVM solution (a comparison with this method is shown in Fig 4). We consider an alternative budget maintenance strategy for our work.

When the prescribed limit is reached, the training samples currently in the active set are used to calibrate probabilities based on Platt scaling [25]. For two class classification, a probability of 0.5 determines random chance (before threshold estimation process). The most extreme probability is obtained by the equation $max(abs \mid 0.5 - p(f(x_i)) \mid_{i=1}^{l})$, where l is the total number of samples in active set, $p()$ is calibrated probability and $f()$ is the current SVM solution in memory. The corresponding sample is determined by the system as the sample to unlearn. Once the training sample to unlearn is determined, it is incrementally unlearned from the existing model. This process is combined as part of the entire update process, thus at each update stage probabilities are calibrated using the training

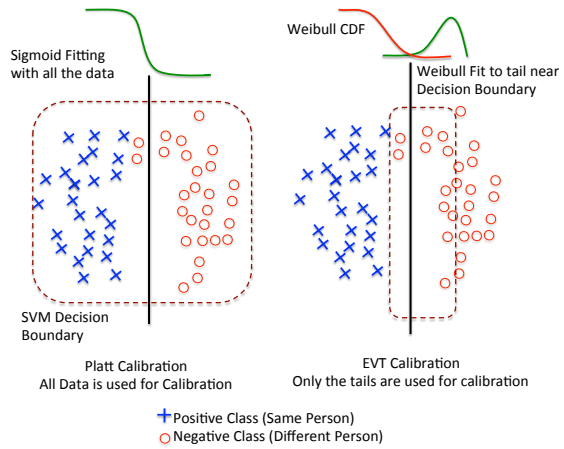

Figure 2. Platt posterior probability estimation [25] method uses all the available training data for calibrating the model. EVT based posterior probability estimation [28] uses the tail of the data near decision boundary for calibration. In streaming face recognition applications when limited data is available for calibration, EVT based methods yield robust posterior probability estimations

data currently in the active set. There are a number of reasons for using calibrated probabilities to determine sample to unlearn. If samples closest to decision boundary are removed, system becomes less robust at handling noisy data. Sigmoid nature of Platt scaling helps avoid these issues [3]. As we compute absolute distance from the mean, even in case of imbalanced data, the sample that is most probable to be included in a particular class is removed. This process helps the system to focus on samples around the decision boundary: an area considered to be most informative for discriminative learning methods [32]

To answer the second question (how to update existing SVM solution), we make a key observation in the method of [5]. Cauwenberghs *et al*. [5] note that update process is reversible: thus when a sample is to be removed from the system, the Lagrangian corresponding to the training sample is assigned to zero. The matrix Q_{ij} is decremented to maintain KKT condition. After every incremental/decremental stage, these optimality conditions (i.e. α_i, Q_{ij}) are saved as the part of the learned model. Although this process adds a small overhead on disk space it guarantees an optimal solution on previously seen data.

In sec 6 we compare the proposed SVM based unlearning method with budgeted stochastic gradient descent method proposed by Wang *et al*. [33]. The method maintains a fixed number of support vectors in the model and incrementally updates them during stochastic gradient descent (SGD) training. Budget maintenance is achieved by removal of support vector with smallest norm. The goal is to minimize degradation of weight vector Δ_i after removal of support vector (here Δ_i is the degradation obtained by removing ith support vector from the system).

5 Probability calibration for Streaming Face Recognition

5.1 Sigmoid Based Calibration

Many computer vision applications require prediction of posterior class probability $p(y = 1|x)$ [13], [26]. Probability calibration method suggested by Platt [25] is the most commonly adopted method for many machine learning tasks. The method proposes approximating the posterior by a sigmoid function. The parameters for calibration are estimated using the entire training set. Maximum likelihood estimation is used to solve for the parameters A and B. When a test sample x_i is to be tested with respect to a learned model $f(x)$, the posterior probability is given by

$$p(y = 1|x) = \frac{1}{1 + exp(Af(x) + B)} \tag{4}$$

In our experiments, we use a version of Platt's method modified by [18] to avoid issues related to numerical difficulties.

5.2 EVT based Calibration

For streaming face recognition the data available in the active set is constantly changing as the system learns from new incoming samples and unlearns existing samples (see Sec 4). This implies data available for calibration purpose for posterior probability estimation changes when the model undergoes an update process. Niculescu-Mizil *et al*. [21] carried out extensive experiments with various learning methods and calibration methods. They noted that posterior probability estimation was more reliable for methods like Platt [25] and isotonic regression when large number of training samples were available for calibration. *For streaming face recognition applications, when data is scarce for smaller budget sizes, how does one obtain reliable posterior probability estimation ?*

In this work, we build on top of a probability calibration model based on Extreme Value Theory for SVMs first proposed by [28], [26]. They noted that the general recognition problem itself is consistent with the assumptions of statistical extreme value theory (EVT), which provides a way to determine probabilities, regardless of the overall distribution of the data. The extremes or tail of a score distribution produced by any recognition algorithm can always be modeled by an EVT distribution, which is a reverse Weibull if the data are bounded. Fig 2 shows the difference between Platt calibration and EVT based calibration. The figure illustrates a toy scenario for a batch learning application. For batch learning, when all the data is available for calibration, Platt calibration methods uses entire data for building posterior probability estimation model. For EVT based calibration, only the tail of the data is used for building estimation model. We use this key observation to build a posterior

probability estimation model based on EVT for streaming face recognition.

Given a SVM decision function $f(x_i)$, and a test sample x_i, we have two independent estimates for posterior probability estimation $P(c \mid f(x_i))$, where c is the class under consideration. The first P_η is based on Weibull cumulative distribution function (CDF) derived from positive class (match $\mathcal{M}$) data. The second, P_ψ, is based on reverse Weibull CDF from negative (non-match $\check{\mathcal{N}}$) estimate, which is equivalent to rejecting the Weibull fitting on 1-vs-all negative (non-match) data. We consider the case of $P_\eta(c \mid f(x_i))$ for our experiments. The Weibull distribution has 3 parameters: scale λ_c, shape κ_c and location τ_c and is given by:

$$W_c(z) = \frac{\tau_c}{\lambda_c}\left(\frac{z-\tau_c}{\lambda_c}\right)^{\kappa_c - 1} e^{\left(\frac{z-\tau_c}{\lambda_c}\right)^{\kappa_c}} \tag{5}$$

with $z > \tau_c$, $\lambda_c > 0$ and $\kappa_c > 0$. For this work we use LibMR provided by [28], which uses maximum likelihood estimate to estimate the Weibull parameters. To estimate the probability at any point x_i belonging to a class c, we use CDF derived from $\mathcal{M}$, given below:

$$P_\eta(c \mid f(x_i)) = 1 - e^{-\left(\frac{-f(x_i)-\tau_\eta}{\lambda_\eta}\right)^{\kappa_\eta}} \tag{6}$$

It is interesting to note that Platt's [25] original observation *The class-conditional densities between the margins are exponential*, is roughly consistent with the Weibull-based CDFs in eqn 5 . The difference is that the Weibull CDF, while generally exponential, 1) has more parameters, 2) will have different parameter fits for the positive and negative classes using only positive or negative data respectively. In addition, EVT allows us to use a very different fitting process using only the extrema, which is more suitable when calibrating with limited data. Eqn 6 models the likelihood of the sample x_i being from match distribution ($\mathcal{M}$).

The overall approach can be summarized as follows:

1. Step 1: Incrementally learning all incoming samples till predefined budget size B is reached
2. Step 2: Once budget size B is reached
   (a) Calibrate data currently in buffer Platt probabilities. Get sample with probability farthest from random chance ($max||0.5 - p(x_i)||$).
   (b) Incrementally unlearn sample with max difference from 0.5 from system
3. Step 3: Incrementally learn incoming training sample. Go to step 2.
4. Step 4: Calibrate probabilities with samples currently in the system. Obtain calibration parameters (3 parameters $z > \tau_c$, $\lambda_c > 0$ and $\kappa_c > 0$ for EVT calibration and 2 parameters A, B for Platt calibration)
5. At any stage, perform prediction with model currently in the system. Obtain probabilities based on calibration parameters wrt model currently in the system.

6 Experiments

Figure 3. The figure shows examples of images from Labeled Faces in the Wild [12] dataset. The images considered in this dataset are taken in unconstrained settings with no control on pose, illumination, gender or context. The dataset contains images of about 5749 individuals with a total of 13233 images.

Dataset: In this section, we discuss our experiments on the problem of face verification in the wild. In face verification, given two face images, goal is to predict whether the images are from the same person. For evaluation, we used Labeled Faces in the Wild (LFW) [12] dataset which contains 13233 images of 5749 individuals, developed for face verification. View 1 of LFW is used for building models, feature selection and finding optimal operating parameters. View 2 consists of 6000 pairs of face images on which performance is to be reported. The 6000 image pairs are divided into 10 sets to allow 10-fold cross-validation. Overall classification performance is reported on "View 2" of the dataset. Fig 3 shows examples of images from LFW dataset.

Features: We use classifier scores obtained from attribute classification method described in [16]. Kumar *et al.* [16] compute visual describable visual attributes for face verification problem. Describable visual attributes are labels given assigned to an image to describe any aspect of its appearance (e.g. gender, hair color, ethnicity: Asian, ethnicity: European etc.). Kumar et al, compute attributes from each pair of images in LFW dataset. Image-pairs can be effectively verified based on presence/absence confidence on variety of these attributes. The total number of features per image used were 146.

Protocol: We follow the protocol as proposed for LFW dataset with minor modification. Training samples are incrementally added to the system, one sample at a time. A budget size is specified for each iteration. Once the prescribed budget size is reached, samples are incrementally unlearned from the system by the method described in Sec 4. For e.g. in a typical run, the system learns with 5400 training samples incrementally. At the end of the learning process, the data present in the active set is saved for probability calibration. Calibration for unlearning is done with

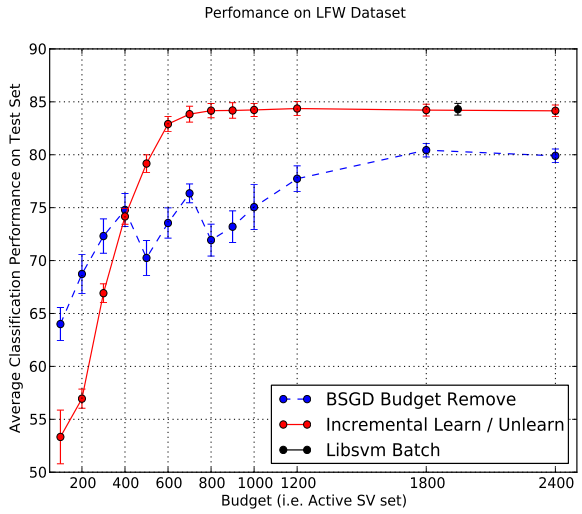

Figure 4. Performance of a leading budgeted stochastic gradient descent method with removal strategy for budget maintenance compared with incremental learning method proposed in this work on LFW face dataset. For each method, average and standard over 10 splits of LFW is shown. Performance obtained with LIBSVM [7] in batch mode is shown. See Sec 6

data in the system at any given point of time. Test performance is predicted on 600 pairs, as proposed in LFW protocol. At no time instant, the system is allowed to exceed the prescribed budget size B.

We conducted a number of experiments to assess the performance of the proposed system. Fig 4 shows comparison of performance of incremental learning/unlearning algorithm proposed and a leading off-the-shelf technique for budgeted online learning [9]. In their previous work, Wang *et al*. [33] found budgeted stochastic gradient descent (BSGD-remove) algorithm with removal strategy (sample to be removed based on smallest norm with respect to current SVM solution) to be a leading method. The performance of BSGD-remove and the proposed algorithm improves as the size of budget is increased. For very small budget sizes, proposed incremental learning/unlearning method performs worse than BSGD-remove. The performance saturates around budget size 1200 for incremental learning and 1800 for BSGD. For reference, the performance obtained with batch learning (1950 support vectors) with LIBSVM [23] is plotted. The performance of both the methods is significantly worse when budget size is less than the number of feature dimensions used for training. Thus, the proposed methods perform better at most budget sizes compared to a BSGD-remove. Henceforth, in our probability calibration experiments, we show the performance on data obtained from incremental learning/unlearning method presented in this work.

Probability Calibration: In earlier section 5, we discussed posterior probability estimation methods for streaming face recognition problem. We first discuss a methodology to evaluate calibration methods followed by our experiments. Reliability diagrams are frequently used in meteorology to assess the probability forecasts for binary events such as probability of measurable precipitation. On X-axis of reliability diagram, mean predicted value is plotted and on Y-axis fraction of positives is plotted. This chart effectively tells the user how often (as a percentage) of predicted probability actual event occurs. Thus, the diagonal line joining $[0, 0]$ and $[1, 1]$ represents ideally calibrated system [34] [11]. Discrete Brier skill score (BSS_D) measures the accuracy of probabilistic predictions. This measure is often used in weather forecasting to assess performance of a system plotted on a reliability diagram. The general formulation of the score is

$$BSS_D = \frac{1}{N}\sum_{t=1}^{N}(f_t - o_t)^2 \tag{7}$$

where f_t is the predicted probability and o_t is the actual outcome (binary) and N is total number of events. The score ranges from $[0, 1]$ with 0 representing ideally calibrated system and 1 representing worst possible calibration. Fig 5 shows BSS_D plotted for different budget sizes for different calibration methods. For a fixed budget size, model was learned incrementally with all the 5400 training samples for each LFW split. The data in the active set at the end of the complete run is used to calibrate model (i.e. in case of Platt, this data is used to estimate A and B params as described in sec 5.1 and estimating the Weibull CDF parameters scale λ_c, shape κ_c and location τ_c in Sec 5.2). Once the models for both methods are calibrated, posterior probability is estimated for each example in the test set. This process is repeated for each test set and each budget size. The figure 5 represents average and standard error over all the splits for these runs.

When data available is limited for calibration, EVT-based calibration gives more reliable posterior probability estimates compared to Platt's method. When more data is available, the performance of both the method converges. EVT calibration focusses only on tails of distribution. As budget size increases the amount of calibration data needed by Platt calibration reaches closer to entire training set. In case of EVT, only tails of distribution are required (irrespective of whether all or only part of the data is available for calibration). These trends are reflected more clearly in reliability diagram shown in Fig 6. When all training data is available for calibration, both Platt and EVT calibration oscillates around ideal calibration line, suggesting well calibrated models. When the models are calibrated using only support vectors (this was obtained when budget size was equal to total training samples and using only support vectors for calibration), the calibration process, appears to os-

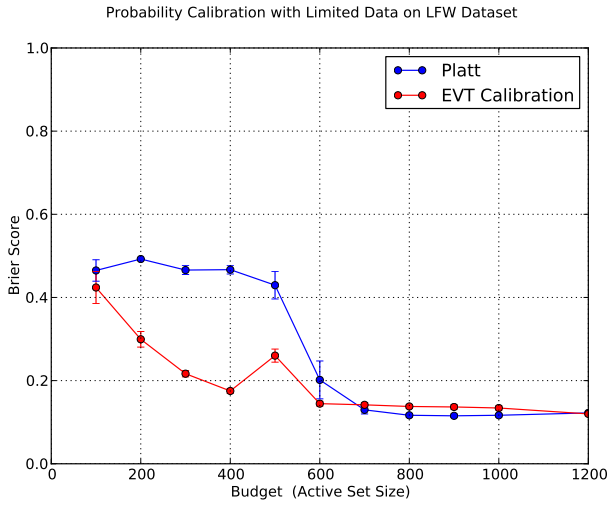

Figure 5. Brier Score Analysis for varying budget sizes. When limited data is available for calibration EVT based calibration gives more reliable posterior probability estimates compared to Platt's method. When more data is available, the performance of both the method converges. (The discrete Brier Skill Score ranges from $[0, 1]$ with 0 representing ideally calibrated system and 1 representing worst possible calibration.)

cillate further from the ideal calibration line. For smaller budget size, this phenomenon is amplified.

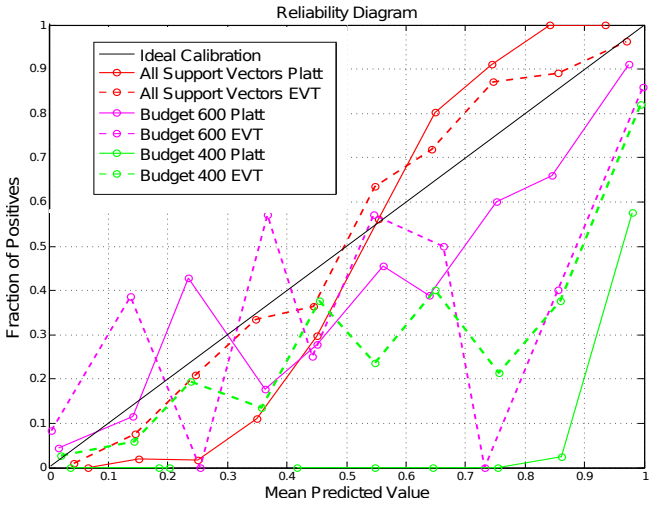

Figure 6. As the amount of calibration data reduces, the reliability of calibration for both Platt and EVT decrease but the EVT degrades most slowly, e.g. consider the green solid (platt) vs green dashedEVT). The EVT is much closer to the diagonal, which is ideal reliability. EVT calibration provides robust posterior estimations when limited data is available for calibration. See BSS_Ds for methods mentioned above in table 6

In the table 6, BSS_Ds for the examples plotted in 6 are given for reference.

Iteration Type	Calibration Method	BSS_D
All Training Data	Platt	**0.1024**
All Training Data	EVT	0.1385
Support Vectors	Platt	0.1345
Support Vectors	EVT	**0.1057**
Budget 600	Platt	0.1285
Budget 600	EVT	**0.1080**
Budget 400	Platt	0.4939
Budget 400	EVT	**0.1769**

7 Discussion

In this work , we presented a method suitable for streaming face recognition problem. We build on existing work on incremental learning and adapt it to incrementally unlearn training samples. Our system can operate on user-specified budget and still yield comparable (and often better) performance to existing off-the-shelf budgeted online learning methods. We proposed a novel posterior probability estimation method based on statistical extreme value theory for streaming face recognition problem. We carry out thorough analysis of the proposed method with respect to state-of-the-art estimation technique and show our method provides more reliable estimates for lower budget sizes. Finally we demonstrate our results on a unconstrained face verification problem and show the suitability of the proposed method to handle long streams of data.

Although there many advantages to the proposed incremental unlearning and EVT based calibration method, there are some limitations. The incremental learning algorithm of [5] scales with Budget size $O(B^2)$. Hence, for really large budget sizes, other alternative learning approaches might be practical. The space requirement of the algorithm is primarily due to saving the state of the optimality (KKT) conditions. Approaches such as Pegasos [30] save only SVM decision boundary, however compromise significantly on accuracy [33]. The proposed calibration techniques should be useful for online learning, irrespective of the choice underlying budget SVM learning algorithm.

We considered a particular approach for unlearning of training sample (described in 4). For streaming face recognition, when the system gains input from multiple cameras [10], a fast filtering mechanism could be devised based on correlation or Nearest Neighbour approach. This could ensure fast processing of incoming data, without explicitly adding all the training samples to the learned model to adapt. A combination of proposed calibration technique with methods in consumer face recognition offer an extremely interesting future direction [15]

References

[1] O. Arandjelovic and R. Cipolla. Incremental learning of temporally-coherent gaussian mixture models. *BMVC*, 2005. 3

[2] L. Best-Rowden, B. Klare, J. Klontz, and A. Jain. Video-to-video face matching: Establishing a baseline for unconstrained face recognition. *BTAS*, 2013. 1

[3] B. Biggio, B. Nelson, and P. Laskov. Support vector machines under adversarial label noise. *ACML*, 2011. 4

[4] S. Blunsden and R. Fisher. The behave video dataset: ground truthed video for multi-person behavior classification. *Annals of BMVA*, 2010. 1

[5] G. Cauwenberghs and T. Poggio. Incremental and decremental support vector machine learning. *NIPS*, 2001. 2, 3, 4, 7

[6] N. Cesa-Bianchi and C. Gentile. Tracking the best hyperplane with a simple budget perceptron. *Machine Learning*, 2007. 3

[7] C.-C. Chang and C.-J. Lin. LIBSVM: A library for support vector machines. *ACM Transactions on Intelligent Systems and Technology*, 2:27:1–27:27, 2011. 6

[8] O. Dekel, S. S. Shwartz, and Y. Singer. The forgetron: A kernel based perceptron on a budget. *SIAM Journal of Computation*, 2008. 3

[9] N. Djuric, L. Lan, S. Vucetic, and Z. Wang. Budgetedsvm: A toolbox for scalable svm approximations. *JMLR*, 2013. 2, 6

[10] J. Harguess, C. Hu, and J. Aggarwal. Fusing face recognition from multiple cameras. *WACV*, 2009. 1, 7

[11] H. Hartmann, T. Pagano, S. Sorooshiam, and R. Bales. Confidence builder: evaluating seasonal climate forecasts from user perspectives. *Bull Amer. Met. Soc.*, 2002. 6

[12] G. Huang, M. ramesh, T. Berg, and E. Learned-Miller. Labeled faces in the wild: A database for studying face recognition in unconstrained environments. *University of Massachusetts, Amherst Tech Report*, 2007. 2, 5

[13] A. J. Joshi, F. Porikli, and N. Papanikolopoulos. Scalable active learning for multi-class image classification. *IEEE TPAMI*, 2012. 4

[14] E. Kakula and F. Shaw. Towards a mobile biometric test framework - presentation. *International Biometric Performance Conference*, 2012. 1

[15] A. Kapoor, S. Baker, S. Basu, and E. Horvitz. Memory constrained face recognition. *CVPR*, 2012. 2, 7

[16] N. Kumar, A. Berg, P. Belhumeur, and S. Nayar. Attribute and simile classifiers for face verification. *ICCV*, 2009. 5

[17] P. Laskov, C. Gehl, S. Kruger, and K. Muller. Incremental support vector learning: Analysis, implementation and applications. *JMLR*, 2006. 3

[18] H. Lin, C. Lin, and R. Weng. A note on plattâĂŹs probabilistic outputs for support vector machines. *Machine Learning*, 2007. 4

[19] A. W. M. Liniger and C. Appenzeller. The discrete brier and ranked probability skill scores. *Monthly Weather Review, American Meteorological Society*, 2007. 2, 3

[20] J. Matai, A. Irturk, and R. Kastner. Design and implementation of an fpga-based real-time face recognition system. *Intern Symp on Field-Programmable Custom Computing Machines*, 2011. 2

[21] A. Niculescu-Mizil and R. Caruana. Predicting good probabilities with supervised learning. *ICML*, 2006. 2, 4

[22] S. Ozawa, S. Lee-Toh, S. A. S. Pang, and N. Kasabov. Incremental learning for online face recognition. *IJCNN*, 2005. 1, 2

[23] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830, 2011. 2, 6

[24] N. Pinto, J. DiCarlo, and D. Cox. How far can you get with a modern face recognition test set using only simple features? *CVPR*, 2009. 1

[25] J. Platt. Probabilities for sv machines. *Advances in Large Margin Classifiers*, 1999. 2, 3, 4, 5

[26] W. Scheirer, N. Kumar, P. Belhumeur, and T. Boult. Multi-attribute spaces: Calibration for attribute fusion and similarity search. *CVPR*, 2012. 2, 3, 4

[27] W. Scheirer, A. Rocha, R. Micheals, and T. Boult. Robust fusion: Extreme value theory for recognition score normalization. *ECCV*, 2010. 3

[28] W. Scheirer, A. Rocha, R. Micheals, and T. Boult. Meta-recognition: The theory and practice of recognition score analysis. *IEEE TPAMI*, 2011. 2, 3, 4, 5

[29] D. Sculley. Combined regression and ranking. *ACM SIGKDD*, 2010. 2

[30] S. Shalev-Shwartz, Y. Singer, N. Srebro, and A. Cotter. Pegasos: Primal estimated sub-gradient solver for svm. *ICML*, 2007. 7

[31] W. Shaoa, W. Yangab, and G.-S. Xiab. Extreme value theory-based calibration for the fusion of multiple features in high-resolution satellite scene classification. *International Journal of Remote Sensing*, 2013. 3

[32] V. Vapnik. The nature of statistical learning theory. *Springer Verlag*, 1995. 3, 4

[33] Z. Wang, K. Crammer, and S. Vucetic. Breaking the curse of kernelization: Budgeted stochastic gradient descent for large-scale svm training. *JMLR*, 2012. 2, 3, 4, 6, 7

[34] D. S. Wilks. On the combination of forecast probabilities for consecutive precipitation periods. *Academic Press*, 1995. 2, 3, 6

[35] L. Wolf, T. Hassner, and I. Maoz. Face recognition in unconstrained videos with matched background similarity. *CVPR*, 2011. 1

[36] J. Yan, Z. Lei, D. Yi, and S. Li. Towards incremental and large scale face recognition. *IEEE IJCB*, 2011. 1, 2

[37] B. Zadrozny and C. Elkan. Obtaining calibrated probability estimates from decision trees and naive bayesian classifiers. *ICML*, 2001. 2

[38] W. Zhao, R. Chellappa, P. J. Phillips, and A. Rosenfeld. Face recognition: A literature survey. *ACM Comput. Survey*, 2003. 1