Perceptual Annotation: Measuring Human Vision to Improve Computer Vision

Walter J. Scheirer, *Member, IEEE*, Samuel E. Anthony, *Student Member, IEEE*, Ken Nakayama, and David D. Cox, *Member, IEEE*

Abstract—For many problems in computer vision, human learners are considerably better than machines. Humans possess highly accurate internal recognition and learning mechanisms that are not yet understood, and they frequently have access to more extensive training data through a lifetime of unbiased experience with the visual world. We propose to use visual psychophysics to directly leverage the abilities of human subjects to build better machine learning systems. First, we use an advanced online psychometric testing platform to make new kinds of annotation data available for learning. Second, we develop a technique for harnessing these new kinds of information—"perceptual annotations"—for support vector machines. A key intuition for this approach is that while it may remain infeasible to dramatically increase the amount of data and high-quality labels available for the training of a given system, measuring the exemplar-by-exemplar difficulty and pattern of errors of human annotators can provide important information for regularizing the solution of the system at hand. A case study for the problem face detection demonstrates that this approach yields state-of-the-art results on the challenging FDDB data set.

Index Terms—Machine learning, psychology, visual recognition, face detection, support vector machines, regularization, citizen science, psychophysics, psychometrics

1 Introduction

For many classes of problems, the goal of computer vision is to solve visual challenges for which human observers have effortless expertise—face and object recognition, image segmentation, and medical image analysis, to name just a few. However, there exists a large class of problems where human performance dramatically outshines current efforts. This occurs even in areas where computer vision has been considered to be highly successful, such as the case of face detection. For example, digital cameras identify faces quickly and accurately, yet when compared to human ability to detect faces in challenging views and environments, no extant algorithm comes close to matching human performance.

There is an obvious gap between current state-of-the-art computer vision applications and human performance. While current methods are improving year by year, there is the concern that such methods will asymptote well below the level of human performance. In this article, we provide a new approach that relies on a heretofore untapped source of information, one that significantly improves performance at a rate beyond current methods. In addition, we argue that this method can be of considerable assistance even for emerging solutions that are not well-studied, as it supplies fundamental information likely to be useful for all algorithms.

Before describing the details of this untapped information, we step back and outline what we believe to be a primary concept of importance, that of the general notion of "learnability," as it applies to people. Consider various stages of expertise in the domain of recognizing a person's origin from their speech, taking the United States as an example. For a newly arrived foreigner from China, recognizing that someone is from the Deep South is perhaps the only such competence. In other words for the novice, distinguishing the Northern and Southern accent is "learnable." Other distinctions, say between a typical Midwestern accent and an East Coast accent are not "learnable." However, the information is there since most Americans can easily make this finer distinction. Further, there are distinctions that are extremely subtle, ones that for most people are not "learnable," say the distinction between people who originated from different parts of Brooklyn. However, some, say a latter day Prof. Henry Higgins (of My Fair Lady) whose speciality is spoken English, would have no difficulties.

How would we teach a new arrival to identify accents? We could start with the easiest distinctions, and when those were acquired, proceed with finer ones. We would never suggest that the novice learn all distinctions at the same time. We would use a graduated approach to learning. However, despite the work of Valiant [2] in formally defining a closely related concept of "learnability" for algorithmic purposes, in the field of machine learning, something akin to a "sink or swim" procedure has been traditionally adopted. For example, in the case of face detection, learning algorithms are presented with images that are labeled "face" and "no face," with little or no effort to tailor the learning to the human ability to learn from particular images. Even worse, there could be images in the training set that even humans cannot discriminate. Little effort is made to take into consideration the rich details of human competence. What we are suggesting here is something more intuitive. Since the point of these machine learning algorithms is to achieve performance levels comparable to that of humans, the human is the obvious standard of reference.

Nonetheless, the reference to human performance is often nonexistent or impoverished. If there is any reference, it is simply to compare overall performance, say measuring human accuracy and comparing it with that of the machine for an extended task with many items. There is much more information about human capacities that is of direct value. For example, some images are learnable and some are not. This learnability also varies with experience. Something that is initially not learnable can be learnable at a later training session. And learnability itself can be further fractionated. Some things are easily and quickly learned; some take more time. Such detailed information reflecting human capacity, which we call a *perceptual annotation*, is something that can be effectively used in conjunction with current algorithms. The key approach to accomplish this is to use the results obtained from the discipline of human psychophysics.

Visual psychophysics was one of the earliest techniques developed for the empirical investigation of internal mental capacities. From the time of its development in the mid-19th century researchers were able to accurately characterize the bounds of human visual capacity. In broadest outline, psychophysics allows the probing of psychological and perceptual thresholds through the manipulation of the characteristics of visual stimuli presented to a subject. The careful management of stimulus construction, ordering and presentation allow perceptual thresholds to be determined precisely—the canonical early example involved the determination of the minimum threshold for stimulation of an individual retinal photoreceptor.

The efficacy of psychophysics as a tool for understanding difficult problems in vision has not gone unnoticed by the computer vision community. Sinha et al. [3] have studied human recognition performance under challenging circumstances (low resolution, changing pose, occlusion, and various artificial distortions), emphasizing that humans perform remarkably well where

- *W.J. Scheirer and D.D. Cox are with the School of Engineering and Applied Sciences, Department of Molecular and Cellular Biology, and the Center for Brain Science, Harvard University, Cambridge, MA 02138. E-mail: {wscheirer, davidcox}@fas.harvard.edu.*
- *S.E. Anthony and K. Nakayama are with the Department of Psychology and the Center for Brain Science, Harvard University, Cambridge, MA 02138. E-mail: {santhony, ken}@wjh.harvard.edu.*

Manuscript received 19 Aug. 2013; revised 13 Dec. 2013; accepted 27 Dec. 2013. Date of publication 1 Jan. 2014; date of current version 10 July 2014.
Recommended for acceptance by F. Fleuret.
For information on obtaining reprints of this article, please send e-mail to: reprints@ieee.org, and reference the Digital Object Identifier below.
Digital Object Identifier no. 10.1109/TPAMI.2013.2297711

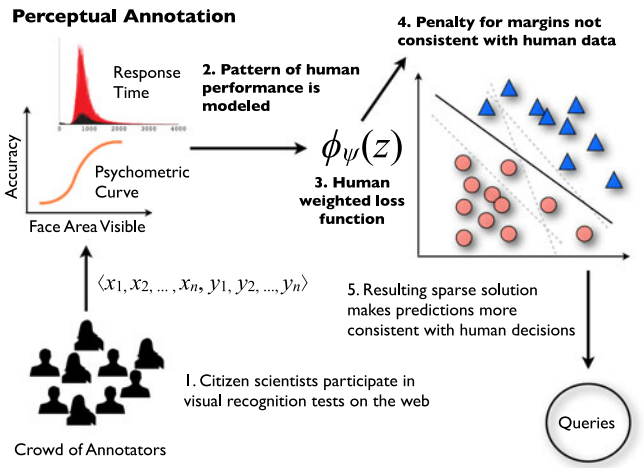

Fig. 1. Standard approaches for incorporating humans into the machine learning process [1] have focused on individual human annotators for labeling difficult or ambiguous training data for continual improvement of a class model. Here we propose a new approach, wherein sets of queries are posed to crowds of citizen scientists on the web. In the framework of psychophysical experiments we can model patterns of error, which can be translated to human weighted loss functions that apply penalties for margins that are not consistent with human data during training. Steps that are different from traditional supervised learning or active learning are highlighted in bold.

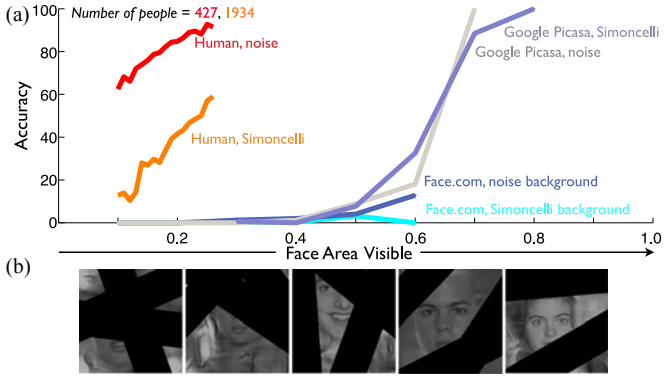

Fig. 2. (a) For many problems in computer vision, humans are still considerably better than machines. These psychometric curves for face detection performance on our "Face in the Branches" test show that as the visible percentage of the face increases, both human and computer performance improves, but in all conditions human performance reaches a high level of accuracy at a level of face visibility where the comparison algorithms (Google's Picasa algorithm, and the face.com algorithm, recently acquired by Facebook) were unable to successfully detect. Note that the curves for humans have been normalized so that performance ranges from zero to one hundred percent accuracy; non-normalized chance accuracy (e.g., random guessing) on a three alternative forced choice task is 33 percent. This normalization allows for a direct comparison of performance with the algorithms, which were given discrete stimuli and asked to make a binary decision. (b) Example occluded stimuli with Portilla-Simoncelli backgrounds arranged from left to right in order of decreasing difficulty and increasing face area visible.

machines currently fail. O'Toole et al. have examined human recognition performance for realistic biometric evaluations, leading to hypothesized upper bounds on challenge problems [4] and goals for scenarios with varying illumination [5]. Further, O'Toole et al. have formulated strategies for fusing human estimates of facial similarity with machine estimates in the context of difficult pair matching problems [6]. These studies have established good baselines for recognition and have made some inroads at augmenting automated approaches, but they have not been used to directly inform learning algorithms to any great extent.

A variety of existing methods have explored a more direct incorporation of humans into the machine learning process, albeit outside of the framework of conventional psychophysics. Active learning [1] is a prominent example of such an approach, wherein training set quality is enhanced by placing human annotators "in-the-loop" with a machine learning system [7], [8], [9], [10], [11], [12], [13]. However, while these algorithms yield significant improvements over traditional supervised learning, they are still largely restricted to improving the quality of training data based on simple class labels. Other studies have used other kinds of human-derived data, e.g., eye movements used to define discriminative image regions for feature extraction [14], [15], structured domain knowledge from human experts [16], [17], [18], and models of the typical human annotation process itself [19]. Similarly, Chen et al. [20] used human performance to constrain decoding of fMRI brain data in those same subjects. Such methods are consonant with the spirit of our approach, however, they are largely tied to specific niches and specific problem formulations.

In order to more completely capture information from human expertise, our approach relies on the collection of a psychophysical "item response" curve from a group of human subjects. This curve, which is described at length in Section 2, captures an exemplar-by-exemplar synopsis of the broad patterns of errors displayed by a population of human subjects performing a difficult task. We describe methods for incorporating this item response data into the objective function of support vector machines, effectively using human performance to guide and regularize a problem's solution. An overview of our approach is shown in Fig. 1.

The contributions of this work are threefold:

1. The use of advanced online psychophysical testing technologies to change the nature and depth of annotation data available for learning by using principled methods of psychometric measurement.
2. A novel model of "human weighted loss" for SVM that incorporates patterns of human performance over the training data, and produces sparse solutions that are more consistent with human performance.
3. A case study in face detection that highlights the effectiveness of perceptually annotated classifiers as filters for "off-the-shelf" detectors. Our results exceed those of the best published algorithms on the FDDB data set [21].

2 Visual Psychophysics Using TestMyBrain

The problem of face detection makes an excellent first case study for a number of reasons. First, it has not received the same level of attention as other components of face processing in the psychophysical literature. Second, computer algorithms for face detection, while mature, have not been informed by human behavior in any significant measure. And fundamental to this work, face detection is a specific case where a large gap between human and machine performance persists (Fig. 2).

There is reason to believe that humans have a specialized ability to detect faces in the environment; people with impaired face recognition skills often have unimpaired face detection ability [22], and a preference for face-like stimuli is present in newborns, well before face recognition abilities have emerged [23]. However, face detection performance is not well explained by models that use low-level salience clues to predict attentional focus [24]. This failure seems to indicate that processes based on higher-order image features (e.g., the eye region) are involved, raising the question of how detection relates to the face-specific recognition modules well-studied in the behavioral [25] and neuroimaging [26] literatures. It is possible that face detection is a largely independent early function with its own specialized brain region; the occipital face area, which has been shown to represent spatial configural information about faces [27] is one possible candidate.

Given such evidence that humans have a specialized pattern recognition mechanism for the detection of faces, operationalization of that latent ability (via the creation of psychophysical measures) allows us to compare human and computer ability on a one-to-one basis. The collection of crowdsourced web data is a well-understood technique in computer vision. However, when working with a latent trait like face detection, it is necessary to deploy more sophisticated measures of human ability to successfully model the observable behavior. Online visual psychophysics is essentially a species of crowdsourcing, but with differences in implementation, motivation, and the nature of collected data that are together vital to the success of our approach. Psychophysics refers to the specific use of a varying physical parameter (e.g., the relative occlusion level of the images) in order to measure a psychological parameter (e.g., the ability to detect the face in the images), or, more generally, to the use of manipulated stimulus presentation to investigate the limits of a cognitive or perceptual ability.

For our psychophysical measures we used the popular TestMyBrain website.[1] TestMyBrain has been used to gather data from more than 600,000 subjects in over 150 countries. The website is specifically designed to capture all of the psychometric measures that would be available to a lab-based experimenter; the data we capture closely matches what is recorded in a traditional, well-controlled lab setting [28]. Importantly, the TestMyBrain subject pool is vast and heterogeneous, which protects against the risk of subject saturation. This differentiates it from other popular crowdsourcing platforms such as Mechanical Turk. The relatively small and computer savvy Mechanical Turk subject pool has led to difficulty mounting experiments where prior ignorance of the experimental conditions is necessary [29], and may skew the results of Turk experiments compared to either lab studies or population studies with a demographically broader pool. Additionally, subjects who participate in experiments on TestMyBrain are motivated by an interest in learning about their own cognitive ability or a desire to participate in academic research as citizen scientists. They are supplied with a detailed explanation of their results and the goal of the experiment in which they have participated; there is much motivation to be as accurate as possible so to maximize a personal score relative to the population. These factors led us to conclude that the perceptual annotations gathered on this platform would provide a maximally general characterization of human ability.

The first test that we developed, "Face in the Branches," is a three alternative forced choice task. In each of the 102 trials presented to a subject, three side-by-side 300×300 pixel images (subtending about 13 degree of visual angle at a 30 inch viewing distance) are shown, and subjects must select the image that contains the face by pressing the 1, 2 or 3 key on their keyboard. One of the three images contained a face selected from a set of fifty male and female frontal face images that were tested for detectability by importing them into Google's Picasa software and confirming a successful detection. In some tests, the images were presented for 450 ms, and in others for 900 ms. There were five visual conditions in all. In four of the five conditions (the "noise" conditions), the images were presented on top of a background of noise matched to the amplitude statistics of the spatial frequency-domain face images. In the fifth condition, the images were presented on top of a background of Portilla-Simoncelli textures [30] that matched the second-order statistics of the face images while scrambling the spatial relations among local features. Each condition included either 1,000 or 2,448 target occluded face images. The fifth Portilla-Simoncelli condition provided the images used in Sections 3 and 4; it had the higher (2,448) number of face images.

In all conditions, the faces contained within the target images varied in size from 50 to 250 pixels in height, and were randomly positioned so that they were fully within the bounds of the larger image. Each of the faces within the target images was occluded so that between 10 and 30 percent of the image remained visible. This range was chosen based on an a priori judgment that this level of occlusion would provide the maximum discrimination of human performance.

Because this test was based on the manipulation of a physical parameter (the area of the face that is visible), it was possible to generate an item response curve characterizing human accuracy as the visible area increased. This curve could then be compared to a curve generated from the performance of state-of-the-art black box face detection algorithms (Google's Picasa algorithm, and the face.com algorithm, recently acquired by Facebook). In this comparison, an item response curve that approaches the upper left of the plot represents better overall performance, and the distance along the y-axis between two curves is a relative measure of the difference in performance. In Fig. 2a we show the results of this comparison. For the stimuli with noise-matched backgrounds, human performance was nearly perfect with only 40 percent of the face visible. By changing the background to the more closely matched Portilla-Simoncelli noise textures (see Fig. 2b) we were able to reduce human performance significantly. However, the curves for both of the algorithms are much farther to the right on the x-axis than the human curves; in all conditions, algorithms yield essentially no successful detections at levels of face visibility where human performance is essentially perfect.

Having established the superiority of human performance on our generated occluded stimuli, our next step was to use the human data we had collected to generate perceptually annotated training samples. However, the occluded stimuli, while a useful measure of human face detection ability, subtend a very small portion of the space of potential face images. In order to create an annotated data set that captured human ability across a wider range of challenging face detection situations, we created an additional psychophysical test.

The second test, "Fast Face Finder," took the form of a present-absent task. Stimuli were face images from the AFLW [31] data set that had been cropped to the dimensions of the outermost facial landmarks and converted to grayscale. Each face was resized to be 250 pixels in width, maintaining the original aspect ratio. Of the 25,993 landmarked faces in AFLW, 4,461 target face images (randomly sampled from the 10,496 images in the set that were not detected by Google's Picasa software) and corresponding foils (generated by sampling equally-sized images from non-face regions of the original Flickr images and then converting them to grayscale) were presented to subjects. Each subject performed two blocks of 102 trials each consisting of 34 face trials and 68 non-face trials. The images were presented for 50 ms; when the time expired, the subject had to press 1 for face or 0 for non-face.

For both tests, accuracy and reaction time were recorded on all trials. This data, accuracy and response time over the population per image, was the raw material used to generate the per-image perceptual annotations required for our machine learning approach. The learning details are described in the next section.

3 Perceptual Annotation for SVM

In any solution to a classification problem, there is some notion of risk involved that indicates the penalties incurred if a prediction is incorrect. The fundamental problem in statistical learning [32] seeks to find a classification function f that minimizes the ideal risk $R_{\mathcal{I}}$:

$$\underset{f}{\operatorname{argmin}} \left\{ R_{\mathcal{I}}(f) := \int_{\mathbb{R}^d \times \mathbb{N}} \phi(x, y, f(x)) P(x, y) \right\}. \quad (1)$$

1. http://www.testmybrain.org.

$R_{\mathcal{I}}$ is composed of two terms, the joint distribution $P(x, y)$ of data x and labels y, and the loss function $\phi(x, y, f(x))$, which assigns the cost of misclassification. Our first step towards a human-regularized support vector machine has been to address the issue of the loss function. A prediction during training can be calculated as the output of the classifier for a particular training sample multiplied by its label: $z = yf(x)$. Typically, a loss function that applies a linearly increasing cost for misclassifications (one-sided error) has been desirable because the minimum of its expected risk coincides with the optimal Bayes error [33]. This is embodied by the hinge loss function, which is defined as:

$$\phi_h(z) = \max(0, 1 - z). \tag{2}$$

However, the non-linear nature of psychometric curves for visual recognition tasks suggests a model that is much different than linear loss growth when $z < 1$.

All training samples are not created equal. We consider per-sample weights on two subsets of the training data instead of a global model over all of the training data (i.e., all of our data doesn't have to be conditioned over the psychometric curve). This is important because we want some number of samples to represent typical images that are easy for both humans and machines to classify to form the basis of our training data. It is the more challenging examples that require special treatment through perceptual annotation. Thus, assume a set of perceptually annotated training examples $\mathcal{P} = (x_i, y_i, c_i)_{i=1\ldots m}$ with $(x_i, y_i) \in \mathbb{R}^d \times \{-1, 1\}$ and $c_i \in \mathbb{R}$. Similarly, assume a set of typical training examples $\mathcal{T} = (x_j, y_j, c_j)_{j=1\ldots n}$. Combined, these two sets form our training data $X = \mathcal{P} \cup \mathcal{T}, m + n = L$.

Human weighted loss can be defined by making use of a mapping function M that associates each data point x with a cost c:

$$\phi_\psi(x, z) = \max(0, (1 - z) + \mathrm{M}(x, z)), \tag{3}$$

where

$$\mathrm{M}(x, z) = \begin{cases} c_x, & \text{if } z < 1, \\ 0, & \text{otherwise.} \end{cases}$$

The cost value c can take on one of two types of values: a static penalty, or a statistic from a point on the psychometric curve corresponding to the measurements for x (e.g., accuracy or reaction time). All perceptually annotated training samples are weighted according to their difficulty, reflected in the chosen statistic, while the typical training examples are weighted by a static cost that is smaller than the smallest perceptual annotation in the training set. For the experiments presented in Section 4.2, we fix c for each non-perceptually annotated training sample to 0 (they strictly follow the hinge loss function). This forces solutions that more aggressively follow human margins, since a much higher cost is associated with the perceptually annotated samples.

For SVM, the standard linear formulation of the classification function is defined as $f(x) = w^{\mathrm{T}} \cdot x + b$, where w and b are parameters of the model (the weight vector and bias term, respectively). To separate the training data in the linear binary case, we solve the following optimization problem:

$$min \frac{1}{2}\|w\|^2 + C \sum_{l=1}^{L} \phi_\psi(x_l, y_l f(x_l)), \tag{4}$$

where the parameter C controls the tradeoff between errors on the training data and margin maximization. The solution f represents a collection of support vectors that form a decision boundary that is strongly influenced by the perceptually annotated training examples via ϕ_ψ.

An interesting aspect of the formulation in Eq. (4) is that it is not convex, which is a controversial issue within the machine learning community. We emphasize, however, that this is both biologically consistent and not a practical computational limitation. Convexity is desirable because it guarantees a globally minimum solution, but it can also restrict us to "shallow" solutions for what are inevitably complex and hierarchical problems in computer vision. Bengio and LeCun have investigated the potential of non-convex loss formulations in-depth [34], specifically in the context of deep learning architectures. Biological visual systems themselves are composed of many layers of adaptive non-linear components [35], which are likely not amenable to a convex formulation [34]. Since our loss function models human behavior, which is the measurable output of such neural machinery, we have no expectation that the formulation should be convex.

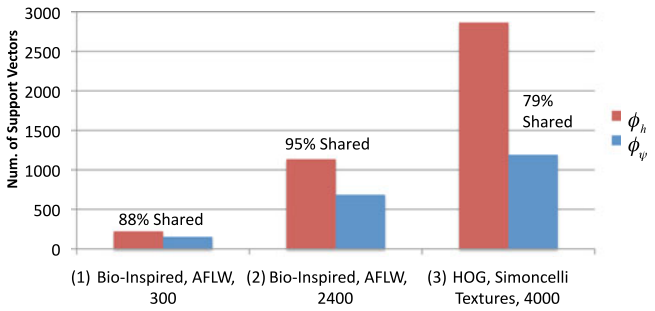

Fig. 3. Number of support vectors selected by SVMs with hinge loss (ϕ_h) and human weighted loss (ϕ_ψ) during training for the same SVM parameter C. In all experiments described in this paper, we observed a solution for human weighted loss that was sparse compared to the corresponding classifier trained with hinge loss. All bars are the average for the 10 classifiers for each experiment (error bars reflecting standard error were too small to be visible).

Specifically relevant to our development of human weighted loss, prior work by Collobert et al. [33] has shown that using a non-convex loss function with SVM reduces space constraints and training time. Similarly, we found that all of our solutions for the experiments in Section 4.2 took no longer to compute than the corresponding solution produced using hinge loss, were more accurate, and sparser (often by an order of magnitude number of support vectors; see Fig. 3). A property of the hinge loss function is that all misclassified training examples become support vectors. If we assume a smooth approximation of hinge loss, the function differentiates to 0 in the flat region ($z > 1$), thus correct classifications do not become support vectors. Several strategies exist for enforcing some measure of sparsity during training. One can make the loss function flat before a predefined threshold in the region where $z < 1$, as was done by Collobert et al. [33]. Alternatively, one can reduce the number of training errors by learning better margins.

Since our objective is to minimize training error through human-influenced regularization, we achieve sparsity by a solution that is a better fit to the training data, rather than through any explicit sparsity-inducing mechanism. This is in contrast to [33], where higher accuracy is not expected. Such implicit sparsity is another biologically-consistent aspect found in brain inspired modeling [36]. An examination of the support vectors learned by both hinge loss and human weighted loss for all experiments revealed that most of the support vectors selected by human weighted loss are shared with those selected by hinge loss (percentages in Fig. 3).

4 Perceptual Annotation for Face Detection

Face detection is interesting from a psychology perspective (Section 2), but it is also a highly relevant and current problem in real-world "in the wild" computer vision, where occlusion, pose variation and noise present in unconstrained imagery confound even the best algorithms.

4.1 Augmenting a Face Detector with Perceptual Annotation

For our experiments, a complete face detection software pipeline incorporating the perceptual annotation learning element described in Section 3 was implemented.[2] For training and testing, we compute features over image patches at a fixed resolution. Since an exhaustive scan of an image using a sliding window and SVM at multiple scales is prohibitively expensive computationally, we have designed a detection algorithm that leverages a standard cascade of Haar features (the ubiquitous Viola-Jones detector [37]) as a first stage. By relaxing the neighborhood scoring constraints of the face detector found in the OpenCV Library [38] (setting this parameter to 0) and increasing the number of scales searched by the algorithm (setting this parameter to <1.1), we collect a larger number of candidate face patches. A perceptually annotated linear SVM, which is more accurate than a Haar cascade, is used as a second stage filter. Patches that are positively identified by the SVM are grouped into neighborhoods, filtered for redundancy, and scored to produce a set of final detection predictions. This second stage filter approach is generic enough to be applied to any detector, not just the Viola-Jones approach we consider here for simplicity and reproducibility.

We examined two different feature types for this work. The first is the well-known dense grid of SIFT features (HOG [39]), which we generated using the VLFeat library [40]. This results in 10,369-dimension histogram bins that are used as feature vectors for learning. We selected this approach because it is the most common and best performing off-the-shelf feature for detection tasks. The second is the multi-layer biologically inspired features of Cox and Pinto [41] meant to mimic the early stages of visual processing, which we generated using the software developed by the authors for that work. Briefly described, the approach consists of multiple stacked layers of linear-nonlinear processing stages, with each stage applying a series of thresholding, saturation, pooling and normalization operations. This process results in 4,097-dimension feature vectors. We selected this approach because of its strong recognition performance [41], and because it lets us build a model that is overall more biologically consistent.

For the perceptually annotated classifiers, we required a set of annotations collected by the TestMyBrain website. Over the course of seven and a half weeks, we collected 337,932 annotations from 3,250 different online research subjects for 4,255 unique images from AFLW by conducting the "Fast Face Finder" test. In a separate collection over the course of two weeks, we gathered 41,650 annotations from 410 different online research subjects for 2,448 unique images from the Portilla-Simoncelli textures set by conducting the "Face in the Branches" test. In both tests, we recorded subject reaction time and accuracy, which after aggregation at the population level, serve as weights c_x in Eq. (3). We sample randomly for perceptually annotated training data from the images seen by between 50 and 77 annotators for the AFLW set, and between 6 and 21 annotators for the Portilla-Simoncelli set.

4.2 Experimental Results

In the following experiments, we make use of data from the FDDB set [21], the most current benchmark for unconstrained face detection. FDDB consists of 2,845 images that contain 5,171 annotated faces, split across 10 different folds for cross-validation style testing. The set includes a wide range of challenges including occlusions, large pose variation, and low resolution and out-of-focus faces (see examples in Fig. 7), making it quite suitable for investigating the potential of new detection models. Our first goal was to determine if there was an observable effect when replacing the hinge loss function of Eq. (2) with the human weighted loss function of Eq. (3) in the linear SVM formulation. We also wanted to assess the impact (if any) of a chosen data set, feature, or measure on accuracy.

2. Code and data are available at http://www.perceptualannotation.org.

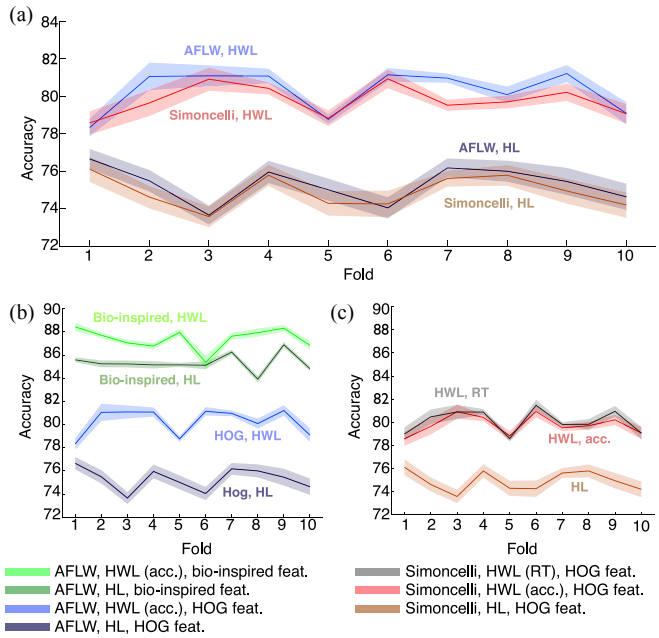

Fig. 4. An increase in accuracy is achieved when the hinge loss (HL) function of a linear SVM is replaced with a human weighted loss (HWL) function. Each curve represents an exhaustive tenfold cross validation experiment where a classifier for each fold of FDDB was trained on 200 images (100 +/−) from that fold and 100 images (50 +/−) from an outside data set. Classifiers were tested on 1,000 images (500 +/−) from each fold not used for training, for a total of 90 classification tests. All classifiers making use of the same data sets saw the exact same training data, and all classifiers were trained with the same SVM C parameter. Shaded regions represent standard error. Other possible configurations not shown did not differ significantly in pattern of results. (a) Accuracy increases when HL is replaced with HWL loss, using either the AFLW or Portilla-Simoncelli perceptually annotated data. Performance did not significantly differ between these data sets. (b) Biologically-inspired features outperform HOG features. Baseline performance increases in both cases when HL is replaced with HWL. (c) HWL using either accuracy or reaction time from the psychometric measure. Both improved baseline performance and did not significantly differ from each other.

To do this, we defined a large-scale classification task using partitions from all folds of FDDB. For each fold, we randomly sampled 500 positive face patches (this represents nearly all of the positive detections for a particular fold—we sample to keep the data uniform across folds), as defined by the ground-truth provided with the data set, and also randomly sampled 500 negative patches that did not overlap with the ground-truth face regions. Each sampled patch from the images was then scaled to 30×30 pixels and processed for features. For training, we assessed different combinations of features (HOG and biologically-inspired), outside data sets of perceptually annotated data (faces obscured by Portilla-Simoncelli textures and AFLW), and measures of human performance (accuracy and reaction time). In each of these cases, we trained a classifier for each fold using 200 images (100 $+/-$) from that fold and 100 perceptually annotated images (50 $+/-$) from one of the outside data sets. To ensure a fair comparison, all classifiers making use of the same data sets saw the exact same training data, and all classifiers were trained with the same SVM C parameter, optimized during training via cross-validation. These classifiers were then tested on all of the data not from the fold used for training (nine tests per fold), for a total of 90 classification tests.

The results of this experiment are shown in Fig. 4. In all cases, we see definitive improvement when hinge loss is replaced with human weighted loss. With respect to the impact of the perceptual annotations (Fig. 4a), the tests with HOG features show similar performance, even though the AFLW and Portilla-Simoncelli sets are

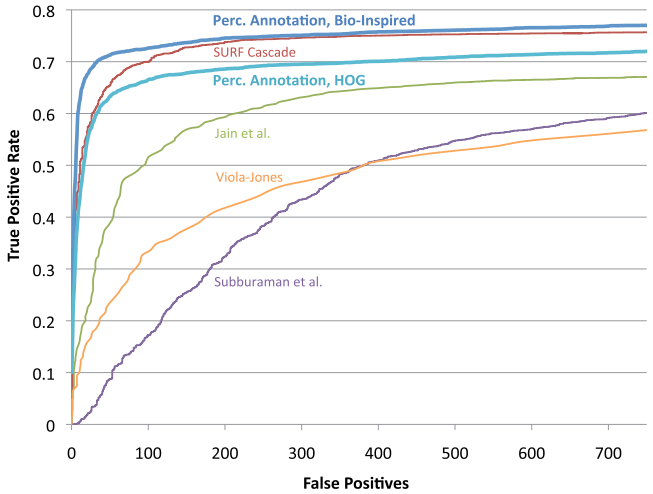

Fig. 5. FDDB results for the discrete score metric (Eq. (6)). The perceptually annotated classifiers trained using the biologically-inspired features of Cox and Pinto [41] produce the best results compared to all prior published approaches reporting on this data set [37], [42], [43], [44]. The biologically-inspired features are especially effective at reducing false positives at higher true positive rates. Note the large measure of improvement between the baseline Viola-Jones algorithm, which is used as a first stage by our detection approach, and perceptual annotation with both feature types.

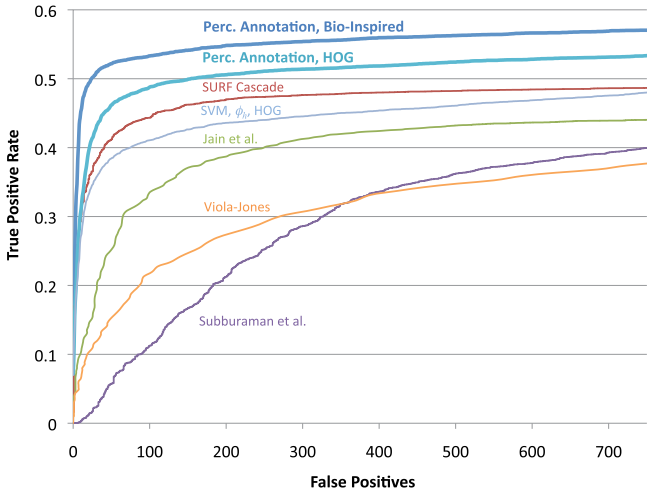

Fig. 6. FDDB results for the continuous score metric (Eq. (7)). For these curves, each individual score contributes to the final result. The perceptually annotated classifiers trained with both feature types yield the highest accuracy, producing much higher quality detections based on the criterion of Eq. (5), compared to prior published approaches reporting on this data set [37], [42], [43], [44]. The curve labeled "SVM, ϕ_h, HOG" highlights the difference in performance when the hinge loss function is replaced with human weighted loss for a well-known feature approach.

very different, and the Portilla-Simoncelli set is not obviously related to the test data. We submit that the two image sets capture different but useful aspects of human performance. The Portilla-Simoncelli images help the classifier identify features visible in frontal but occluded faces, while the AFLW images capture pose variation. Further, in the case of the Portilla-Simoncelli set, it is also conceivable that the learning is able to distinguish between face and closely resembling non-face texture in the training images, giving us some additional resistance against false positives. The choice of feature (Fig. 4b) impacts the resulting model performance to a much larger degree. Interestingly, we see very good interaction between the biologically-inspired features and the perceptually annotated data, with a large improvement over HOG features in this case. We also examined the effect of the chosen psychometric measure (Fig. 4c) on human weighted loss. Accuracy and reaction did not differ significantly.

With an established effect, we then moved on to assess the viability of perceptually annotated classifiers for an unconstrained face detection task. For these experiments, we used the standard FDDB protocols [21]. To calculate the degree of match between a detected region d_i and a ground-truth region l_j, the ratio of intersected areas to joined areas is used:

$$S(d_i, l_j) = \frac{area(d_i) \cap area(l_j)}{area(d_i) \cup area(l_j)}. \quad (5)$$

From the ratio score S, a discrete decision score y_i can be calculated by using a function δ that assigns a score of 1 to the detected region if $S > 0.5$ and 0 otherwise:

$$y_i = \delta_{S(d_i,l_j)>0.5}. \quad (6)$$

An alternative strategy is to treat the ratio score as the decision score itself. This is useful for determining the quality of detections, where the ratio matters:

$$y_i = S(d_i, l_j). \quad (7)$$

Using the algorithm described in Section 4.1, we collected detections for all folds. The classifiers trained with the biologically-inspired features made use of 1,800 images (900 $+/-$) sampled from all folds not used for testing, and 600 images (300 $+/-$) from the perceptually annotated AFLW set, while the HOG classifiers made use of 3,600 images (1800 $+/-$) from FDDB and 400 images (200 $+/-$) from the perceptually annotated Portilla-Simoncelli textures set. We chose to highlight the utility of both perceptually annotated data sets, with the expectation that performance would increase in both cases, based on our results in Fig. 4. Through cross-validation on the training sets, we determined that a patch size of 30×30 was suitable for the biologically inspired features, and 40×40 was suitable for the HOG features. The perceptual annotations in both cases incorporated accuracy as a measure of human performance. All scores S and y_i were calculated using the software provided by the maintainers of FDDB [21].

The results for the discrete test are shown in Fig. 5. The perceptually annotated classifiers trained using the biologically-inspired features of Cox and Pinto produce the best results compared to all prior published approaches reporting on this data set [37], [42], [43], [44]. Perhaps more meaningful are the results for the continuous test, shown in Fig. 6, where the quality of score matters. Both sets of perceptually annotated classifiers produce results that exceed the state-of-the-art here, indicating a strong preference for patches that minimize the surrounding background—something that is important for a subsequent task such as face verification or identification. Moreover, we note that our best result for the biologically-inspired features exceeds that of the "black box" commercial systems reporting on this same test.[3] Finally, to demonstrate a continued positive effect for human weighted loss on the detection task, we include an additional comparison curve in Fig. 6 for a set of SVM classifiers with the original hinge loss function preserved (grey curve).

5 Discussion

This article represents a first implementation of a class of learning algorithms that incorporates measured manifestations of perceptual human knowledge at training time. By seeking out new perspectives from psychology, we have shown that large-scale visual psychophysics allows us to take advantage of

3. Found on: http://vis-www.cs.umass.edu/fddb/results.html.

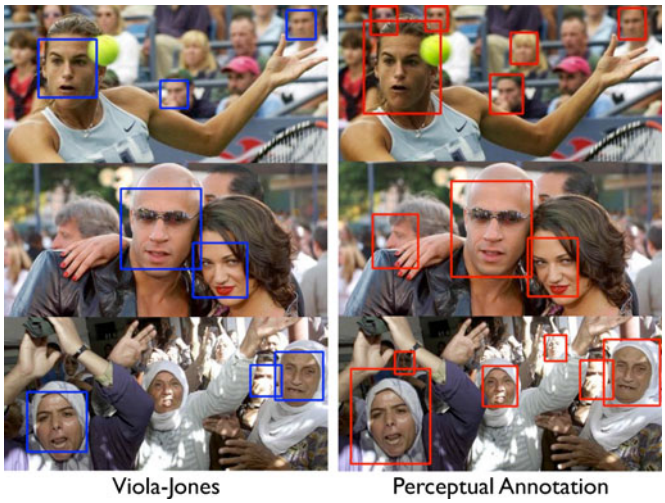

Fig. 7. Visual examples of detected FDDB faces from the full perceptual annotation approach and the baseline Viola-Jones algorithm. Perceptually annotated classifiers are better at detecting low-resolution and occluded faces, as well as those that are highly impacted by artifacts such as non-uniform illumination. Note that recall is not perfect for the perceptually annotated classifiers: at least one face is missed in each image shown above.

annotations that are far more descriptive than typical class labels in a supervised context. Our initial formulation places its emphasis on a risk calculation that considers misclassification penalties on a distance plus per example basis, which yields sparse solutions with margins more consistent with human behavior. The notion of a non-convex loss function like the one in Eq. (3) is indeed controversial, but as Bengio and LeCun [34] state, it " ... may be an unavoidable property of learning complex functions from weak prior knowledge."

With just a boosted cascade of Haar features as a basis, we have shown that perceptually annotated classifiers are able to filter candidate face windows to an extent of accuracy that exceeds all prior published approaches on the challenging, unconstrained FDDB data set. Beyond this base formulation, there is much potential for the general principle of perceptual annotation with respect to data collection, algorithms and applications. Measurements can be made via fMRI and EEG in humans, and electrophysiology in other animals that can recognize objects. Learning is also not constrained to SVM: alternative formulations for boosting, random forests, and neural networks (among others) are possible. Various combinations of annotation and learning strategies can be applied to applications as diverse as general object recognition, visual attribute assignment, face recognition, and segmentation. Considering all of these elements, we have merely scratched the surface of what these vastly richer forms of annotation can accomplish.

Acknowledgments

This work was supported by NIH Grant R01 EY01363, US National Science Foundation (NSF) IIS Award #0963668 and a gift from the Intel Corporation. Walter J. Scheirer and Samuel E. Anthony contributed equally to this work.

References

[1] B. Settles, *Active Learning*, Morgan & Claypool, 2012.
[2] L.G. Valiant, "A Theory of the Learnable," *Comm. ACM*, vol. 27, no. 11, pp. 1134-1142, 1984.
[3] P. Sinha, B. Balas, Y. Ostrovsky, and R. Russell, "Face Recognition by Humans: Nineteen Results All Computer Vision Researchers Should Know About," *Proc. IEEE*, vol. 94, no. 11, pp. 1948-1962, 2006.
[4] A. O'Toole, P. Phillips, and A. Narvekar, "Humans versus Algorithms: Comparisons from the Face Recognition Vendor Test 2006," *Proc. Eighth IEEE Int'l Conf. Automatic Face and Gesture Recognition (AFGR)*, Sept. 2008.
[5] A. O'Toole, P. Phillips, F. Jiang, J. Ayyad, N. Penard, and H. Abdi, "Face Recognition Algorithms Surpass Humans Matching Faces across Changes in Illumination," *IEEE Trans. Pattern Analysis and Machine Intelligence*, vol. 29, no. 9, pp. 1642-1646, Sept. 2007.
[6] A. O'Toole, X. An, J. Dunlop, V. Natu, and P. Phillips, "Comparing Face Recognition Algorithms to Humans on Challenging Tasks," *ACM Trans. Applied Perception*, vol. 9, no. 4, pp. 16:1-16:13, Oct. 2012.
[7] S. Vijayanarasimhan and K. Grauman, "What's It Going to Cost You? Predicting Effort vs. Informativeness for Multi-Label Image Annotations," *Proc. IEEE Conf. Computer Vision and Pattern Recognition (CVPR)*, June 2009.
[8] G. Kunapuli, R. Maclin, and J. Shavlik, "Multi-Level Active Prediction of Useful Image Annotations for Recognition," *Proc. Advances in Neural Information Processing Systems (NIPS)*, Dec. 2009.
[9] B. Settles, M. Craven, and S. Ray, "Multiple-Instance Active Learning," *Proc. Advances in Neural Information Processing Systems (NIPS)*, Dec. 2008.
[10] G. Qi, X. Hua, Y. Rui, J. Tang, and H. Zhang, "Two-Dimensional Active Learning for Image Classification," *Proc. IEEE Conf. Computer Vision and Pattern Recognition (CVPR)*, June 2008.
[11] S. Vijayanarasimhan, P. Jain, and K. Grauman, "Far-Sighted Active Learning on a Budget for Image and Video Recognition," *Proc. IEEE Conf. Computer Vision and Pattern Recognition (CVPR)*, June 2010.
[12] S. Vijayanarasimhan and K. Grauman, "Large-Scale Live Active Learning: Training Object Detectors with Crawled Data and Crowds," *Proc. IEEE Conf. Computer Vision and Pattern Recognition (CVPR)*, June 2011.
[13] A. Biswas and D. Jacobs, "Active Image Clustering: Seeking Constraints from Humans to Complement Algorithms," *Proc. IEEE Conf. Computer Vision and Pattern Recognition (CVPR)*, June 2012.
[14] J. Deng, J. Krause, and L. Fei-Fei, "Fine-Grained Crowdsourcing for Fine-Grained Recognition," *Proc. IEEE Conf. Computer Vision and Pattern Recognition (CVPR)*, 2013.
[15] E. Vig, M. Dorr, and D. Cox, "Space-Variant Descriptor Sampling for Action Recognition Based on Saliency and Eye Movements," *Proc. 12th European Conf. Computer Vision: Part VII (ECCV)*, 2012.
[16] K. Ganchev, J. Graa, J. Gillenwater, and B. Taskar, "Posterior Regularization for Structured Latent Variable Models," *J. Machine Learning Research*, vol. 11, pp. 2001-2049, 2010.
[17] G. Kunapuli, R. Maclin, and J. Shavlik, "Advice Refinement for Knowledge-Based Support Vector Machines," *Proc. 25th Conf. Neural Information Processing Systems (NIPS)*, Dec. 2011.
[18] K. Small, B. Wallace, C. Brodley, and T. Trikalinos, "The Constrained Weight Space SVM: Learning with Ranked Features," *Proc. 28th Int'l Conf. Machine Learning (ICML)*, 2011.
[19] P. Welinder, S. Branson, S. Belongie, and P. Perona, "The Multidimensional Wisdom of Crowds," *Proc. 24th Ann. Conf. Neural Information Processing Systems (NIPS)*, Dec. 2010.
[20] D. Chen, S. Li, Z. Kourtzi, and S. Wu, "Behavior-Constrained Support Vector Machines for fMRI Data Analysis," *IEEE Trans. Neural Networks*, vol. 21, no. 10, pp. 1680-1685, Oct. 2010.
[21] V. Jain and E. Learned Miller, "FDDB: A Benchmark for Face Detection in Unconstrained Settings," Technical Report UM-CS-2010-009, Univ. of Massachusetts, Amherst, 2010.
[22] L. Garrido, B. Duchaine, and K. Nakayama, "Face Detection in Normal and Prosopagnosic Individuals," *J. Neuropsychology*, vol. 2, no. 1, pp. 119-140, 2008.
[23] E. Valenza, F. Simion, V.M. Cassia, and C. Umiltà, "Face Preference at Birth," *J. Experimental Psychology: Human Perception and Performance*, vol. 22, no. 4, pp. 892-903, 1996.
[24] L. Itti and C. Koch, "Computational Modelling of Visual Attention," *Nature Rev. Neuroscience*, vol. 2, no. 3, pp. 194-203, 2001.
[25] B. Duchaine and K. Nakayama, "The Cambridge Face Memory Test: Results for Neurologically Intact Individuals and an Investigation of Its Validity Using Inverted Face Stimuli and Prosopagnosic Participants," *Neuropsychologia*, vol. 44, no. 4, pp. 576-585, 2006.
[26] N. Kanwisher, J. McDermott, and M.M. Chun, "The Fusiform Face Area: A Module in Human Extrastriate Cortex Specialized for Face Perception," *J. Neuroscience*, vol. 17, no. 11, pp. 4302-4311, 1997.
[27] D. Pitcher, V. Walsh, and B. Duchaine, "The Role of the Occipital Face Area in the Cortical Face Perception Network," *Experimental Brain Research*, vol. 209, no. 4, pp. 481-493, 2011.
[28] L. Germine, K. Nakayama, B. Duchaine, C. Chabris, G. Chatterjee, and J. Wilmer, "Is the Web As Good As the Lab? Comparable Performance from Web and Lab in Cognitive/Perceptual Experiments," *Psychonomic Bull. and Rev.*, vol. 19, pp. 847-857, 2012.
[29] J. Chandler, P. Mueller, and G. Paolacci, "Nonnaïveté among Amazon Mechanical Turk Workers: Consequences and Solutions for Behavioral Researchers," *Behavior Research Methods*, vol. 7, pp. 1-19, http://link.springer.com/article/10.3758%2Fs13428-013-0365-7, 2013.
[30] J. Portilla and E. Simoncelli, "A Parametric Texture Model Based on Joint Statistics of Complex Wavelet Coefficients," *Int'l J. Computer Vision*, vol. 40, no. 1, pp. 49-70, 2000.
[31] M. Kostinger, P. Wohlhart, P.M. Roth, and H. Bischof, "Annotated Facial Landmarks in the Wild: A Large-Scale, Real-World Database for Facial Landmark Localization," *Proc. IEEE Int'l Conf. Computer Vision Workshops*, pp. 2144-2151, 2011.
[32] A. Smola, "Learning with Kernels," PhD Dissertation, Technische Universität Berlin, Nov. 1998.

[33] R. Collobert, F. Sinz, J. Weston, and L. Bottou, "Trading Convexity for Scalability," *Proc. 23rd Int'l Conf. Machine Learning (ICML)*, 2006.
[34] Y. Bengio and Y. LeCun, "Scaling Learning Algorithms towards AI," *Large Scale Kernel Machines*, MIT Press, 2007.
[35] J.J. DiCarlo and D.D. Cox, "Untangling Invariant Object Recognition," *Trends in Cognitive Sciences*, vol. 11, pp. 333-341, 2007.
[36] B. Olshausen and D. Field, "Emergence of Simple-Cell Receptive Field Properties by Learning a Sparse Code for Natural Images," *Nature*, vol. 381, pp. 607-609, June 1996.
[37] P. Viola and M. Jones, "Robust Real-Time Face Detection," *Int'l J. Computer Vision*, vol. 57, pp. 137-154, 2004.
[38] "OpenCV Library," http://code.opencv.org, Dec. 2013..
[39] N. Dalal and B. Triggs, "Histograms of Oriented Gradients for Human Detection," *Proc. IEEE Conf. Computer Vision and Pattern Recognition (CVPR)*, 2005.
[40] A. Vedaldi and B. Fulkerson, "VLFeat: An Open and Portable Library of Computer Vision Algorithms," *Proc. Int'l Conf. Multimedia*, 2008.
[41] N. Pinto and D. Cox, "Beyond Simple Features: A Large-Scale Feature Search Approach to Unconstrained Face Recognition," *Proc. IEEE Int'l Conf. Automatic Face and Gesture Recognition (AFGR)*, 2011.
[42] J. Li, T. Wang, and Y. Zhang, "Face Detection Using SURF Cascade," *Proc. IEEE Int'l Conf. Computer Vision Workshop (ICCV)*, 2011.
[43] V. Jain and E. Learned-Miller, "Online Domain-Adaptation of a Pre-Trained Cascade of Classifiers," *Proc. IEEE Conf. Computer Vision and Pattern Recognition (CVPR)*, 2011.
[44] V.B. Subburaman and S. Marcel, "Fast Bounding Box Estimation Based Face Detection," *Proc. Workshop Face Detection of the European Conf. Computer Vision (ECCV)*, 2010.