Detection, Localization, and Sex Classification of Faces from Arbitrary Viewpoints and under Occlusion

Matthew Toews, *Member*, *IEEE*, and Tal Arbel, *Member*, *IEEE*

Abstract—This paper presents a novel framework for detecting, localizing, and classifying faces in terms of visual traits, e.g., sex or age, from arbitrary viewpoints and in the presence of occlusion. All three tasks are embedded in a general viewpoint-invariant model of object class appearance derived from local scale-invariant features, where features are probabilistically quantified in terms of their occurrence, appearance, geometry, and association with visual traits of interest. An appearance model is first learned for the object class, after which a Bayesian classifier is trained to identify the model features indicative of visual traits. The framework can be applied in realistic scenarios in the presence of viewpoint changes and partial occlusion, unlike other techniques assuming data that are single viewpoint, upright, prealigned, and cropped from background distraction. Experimentation establishes the first result for sex classification from arbitrary viewpoints, an equal error rate of 16.3 percent, based on the color FERET database. The method is also shown to work robustly on faces in cluttered imagery from the CMU profile database. A comparison with the geometry-free bag-of-words model shows that geometrical information provided by our framework improves classification. A comparison with support vector machines demonstrates that Bayesian classification results in superior performance.

Index Terms—Scale-invariant feature, viewpoint invariance, probabilistic modeling, visual trait, sex classification, faces, occlusion.

1 Introduction

Practical image processing applications must be able to robustly detect faces in arbitrary, cluttered images, and make inferences regarding their visual traits. For example, consider an intelligent vision system that must identify males in a crowded scene, as illustrated in Fig. 1. Image features associated with human face instances must first be detected and localized in the midst of unrelated clutter, occlusion, and viewpoint change, after which they can be used to classify traits such as sex for each person detected. Although the tasks of feature detection, localization, and classification are all inextricably linked in such a realistic image processing scenario, they are often treated in isolation in the current vision literature. For example, approaches to classifying facial traits such as sex typically assume frontal face data which have been prealigned and/or cropped from distracting clutter prior to classification [1], [2], [3], [4], [5], [6]. As a result, it is unlikely that they can be applied in arbitrary scenes, where the image features required for classification may be difficult to localize or may not even exist due to viewpoint change or occlusion (e.g., scarves, hairstyles). Likewise, recent work has shown that general 3D object classes can be detected and localized from arbitrary viewpoints and clutter using probabilistic models of local scale-invariant features [7], [8], [9], [10], [11]. However, it is unclear whether such models can be used to classify facial traits, or how effective such classification would be.

In this paper, we present an integrated framework for detecting, localizing, and classifying faces in terms of traits such as sex, from arbitrary viewpoints and under occlusion. Our approach, which in its preliminary form was presented in [14], is the first to propose learning facial traits from arbitrary viewpoints, and the first to embed all three computer vision tasks in a common framework. The framework is based on a general viewpoint-invariant appearance model derived from local scale-invariant features (e.g., SIFT), where features are probabilistically quantified in terms of their occurrence, appearance, and geometry within a common reference frame. Our approach involves first learning a viewpoint-invariant model of face appearance, after which learned facial features are used to train a Bayesian classifier of facial traits. Classifier training involves estimating the likelihood ratio of feature occurrence given trait presence versus absence, the underlying premise being that informative features are more likely than not to co-occur with the trait of interest. As our framework is invariant to viewpoint changes, it can be trained and tested on sets of image data acquired from arbitrary viewpoints, and does not require explicit modeling of multiple views or 3D structure.

This paper provides an in-depth exploration of facial trait classification, largely due to the availability of labeled public data from which difficult traits such as sex can be studied. The framework we present is potentially applicable to object classes other than faces, however, and we attempt

- *M. Toews is with the Department of Radiology, Surgical Planning Laboratory, Harvard Medical School, ASBI, L1-050, Brigham & Women's Hospital, 75 Francis St., Boston, MA 02115. E-mail: mt@bwh.harvard.edu.*
- *T. Arbel is with the Centre for Intelligent Machines, McConnell Engineering Bldg., Room 425, McGill University, 3480 University Street, Montreal, Quebec H3A 2A7, Canada. E-mail: arbel@cim.mcgill.ca.*

Manuscript received 4 Oct. 2007; revised 14 Apr. 2008; accepted 8 Sept. 2008; published online 16 Sept. 2008.
Recommended for acceptance by T. Darrell.
For information on obtaining reprints of this article, please send e-mail to: tpami@computer.org, and reference IEEECS Log Number TPAMI-2007-10-0671.
Digital Object Identifier no. 10.1109/TPAMI.2008.233.

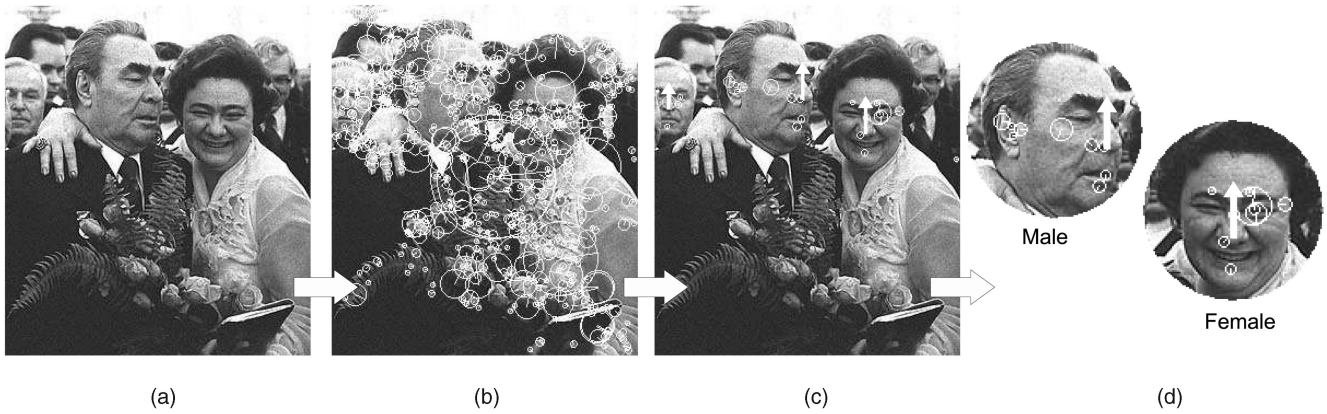

Fig. 1. Illustrating our framework for viewpoint-invariant detection, localization, and trait classification from arbitrary viewpoints. All three tasks are embedded in a viewpoint-invariant model derived from scale-invariant image features. In (b), scale-invariant features (white circles) are extracted from an image of a cluttered scene (a). Next, in (c), the viewpoint-invariant model is used to detect and localize face instances (small white arrows) and associated features. Finally, in (d), a Bayesian classifier is used to determine the sex of face instances from associated features. The image shown is from the CMU face database [12], and the probabilistic framework used is learned from 500 color FERET [13] face images acquired from arbitrary viewpoints.

to maintain a general discussion throughout. The remainder of this paper is organized as follows: In Section 2, we review related work in general object class detection and visual trait classification, in particular, the trait of sex in face images. In Section 3, we describe our framework for trait learning and classification based on probabilistic modeling of scale-invariant image features. In Section 4, we describe experimentation relating to face detection, localization, and sex classification. Using the standard color FERET database [13], we provide a quantitative performance evaluation of combined detection, localization, and sex classification of face images acquired from arbitrary viewpoints and in the presence of simulated occlusion, the first results of their kind in the literature. We compare viewpoint-invariant modeling and Bayesian classification with geometry-free bag-of-words modeling and support vector machine (SVM) classification, alternative approaches which could also be used. We show how our approach can be used to identify visual cues of sex in face images over a range of viewpoints, and demonstrate our system on a subset of difficult, cluttered face images from the CMU profile database [12]. Finally, in Section 5, we conclude with a discussion and pointers to future work.

2 Related Work

Visual traits are qualities of an object class identifiable from images, such as the make or model of cars, or the age or sex of faces. They represent a mechanism by which members of the same object class can be described or subdivided into meaningful categories. In order to classify visual traits from arbitrary viewpoints and in the presence of occlusion, there must be a means of reliably identifying and localizing the image features on which classification is based. To date, the majority of classification approaches utilize feature representations that cannot be easily localized from arbitrary viewpoints or in the presence of occlusion. Here, we review classification of the specific trait of sex from face images, and present recent work in feature detection that makes it possible to identify local image features from arbitrary viewpoints and in the presence of occlusion, thus providing a basis for our classification system.

2.1 Image Features

In order to classify visual traits of object classes such as faces, the image features used in classification and their associated object class instances must first be detected and localized in images. Although much work in visual trait classification assumes prelocalized object instances and image features, detection and localization are generally nontrivial in arbitrary scenes. This is because object class detection requires effective dealing with a wide range of appearance variation due to viewpoint changes, geometrical deformations such as translations, rotations, and scale changes, illumination changes, partial pattern occlusion, and multimodal intraclass variation (e.g., faces with/without sunglasses). In this paper, we seek a practical trait classification system based on image features that can be detected and localized in arbitrary imagery, particularly images acquired from arbitrary viewpoints and in the presence of occlusion.

Determining which features can be detected and used to classify visual traits in arbitrary imagery can be achieved by learning an appearance model from a set of training images. Early approaches advocated learning models based on global features, e.g., eigenfaces [15], but global features are poorly suited for coping with local appearance variation and occlusion and have been shown to be suboptimal for detection [16]. Researchers have increasingly turned to local image feature representations [17], which can be identified in the presence of partial pattern occlusion. Local Haar wavelets, for instance, have proven useful for frontal face detection using boosted classifiers [18], as they can be computed very efficiently using integral images, particularly for frontal, upright images. The integral image approach is less effective at coping with arbitrary viewpoints, as a battery of detectors must be used to model in-plane deformation parameters such as image scale and orientation and out-of-plane viewpoint changes [19].

Local scale-invariant features [20], [21], [22], [23], [24] offer an attractive alternative to Haar wavelets due to their high degree of invariance to in-plane transforms. Scale-invariant features are oriented image regions characterized geometrically within an image in terms of their location x, orientation θ, and scale σ. They can be efficiently extracted from images using scale-space pyramids in the presence of geometrical deformations such as translations, rotations, and scale changes and linear changes in illumination. Feature geometrical information obtained during the extraction process can be used to generate independent hypotheses as to the geometrical transform relating different images, without requiring an expensive explicit search over transform parameters. Scale-invariant features can be extracted from a variety of different underlying image characteristics, including derivatives in image scale [20] and space [22], phase [21], entropy [23], color moments [24], and others.

While invariant features can be used to reliably identify instances of the same scene or object in new images, i.e., the task of object detection, they cannot be used directly to obtain correspondence between different instances of the same class, for example, faces of different people. This is primarily due to intraclass appearance variability, e.g., changes in facial expression, facial hair, makeup, etc. Probabilistic modeling can be used to deal with these difficulties, the subject of the next section.

2.2 Modeling Appearance from Invariant Features

Probabilistic modeling and machine learning can be used in order to reliably identify invariant features and their associated object class instances in arbitrary, cluttered images [25], [26], [27], [28], [29], [30], [31], [32], [7], [33]. Probabilistic models describe the appearance of an object class in terms of a set of local features, including their appearances, occurrences, and geometries (e.g., image location, orientation, and scale). Models generally vary in terms of the assumptions made regarding interfeature geometrical dependencies, e.g., geometry independent models [27], [28], naive Bayes dependencies (i.e., star models) [25], [34], [7], [26], Markov (neighborhood) dependencies [9], [32], fully dependent models (i.e., constellation models) [35], and intermediate approaches [29]. Although geometrical dependence assumptions vary, most models make the assumption of conditional independence of individual feature appearances/occurrences given feature geometries and the object class. In this way, features can be efficiently identified independently in terms of their appearances and then used to construct geometrical hypotheses as to how they relate to form object classes.

Most approaches to invariant feature modeling are based on stable 2D feature configurations in the image plane, and are thus single viewpoint in nature. Recent approaches have extended modeling to general 3D object classes from arbitrary viewpoints, for a diverse range of object classes such as faces, motorbikes, shoes, etc. [25], [7], [8], [9], [10], [11]. To do this, models must have a means of accounting for the variable of viewpoint, or pose of the camera relative to the object class. This can be done by either explicitly modeling the variable of viewpoint or formulating the model in manner which is *invariant* to viewpoint change [36]. Modeling viewpoint explicitly involves describing appearance as a function of viewpoint, either by maintaining a set of views or aspects around the object class [37], [25], [8], [9] or by using 3D modeling to generate appearance from novel views [11], [10]. Detection is then performed by fitting data to the nearest view, requiring a search over viewpoint. The viewpoint-invariant approach relates image features to a geometrical structural description in a manner independent of viewpoint, e.g., using perspective invariants [7], [38], or parameterized volumetric primitives such as geons [39] or generalized cylinders [40]. As the variable of viewpoint is effectively marginalized from the formulation, detection can be achieved independently of viewpoint.

Model learning generally involves clustering techniques in order to identify scale-invariant features in different images, which are similar in terms of their appearances and their geometries relative to the object class. Although unsupervised learning techniques have been used for single-viewpoint models [35], [41], learning models over viewpoint variation typically requires a degree of additional supervision or data preparation in order to establish feature correspondences across neighboring viewpoints in addition to across object class instances. The multiview learning techniques of Thomas et al. [8], Savarse and Fei-Fei [10], and Yan et al. [11] require images captured from multiple viewpoints around each individual object class instance. As such, they are not directly applicable to arbitrary sets of images where a single object instance may not be seen more than once. The multiview model of Schneiderman and Kanade [25] and the flexible model of Kushal et al. [9] require object class instances to be localized within the image and sorted according to viewpoint. The object class invariant (OCI) model of Toews and Arbel [7] requires only localization of object class instances for learning, no explicit viewpoint information. For this reason, we adopt the OCI model in this paper, which we describe later in more detail.

2.3 Visual Trait Classification: Sex from Faces

Visual traits are abstract qualities of an object class identifiable from images, by which members of an object class can be described or categorized. While subcategories can be defined in a taxonomical fashion according to specific image features [42] or segmentations [43], visual traits are not generally defined by observable image features per se but rather by factors external to the image. The trait of sex, for example, is clearly defined by factors external to the image; however, the sex of a face can be inferred from an ensemble of image features. Thus, unlike subcategories defined solely by image features [42], [43], classification based on visual traits generally implies a learning process based on external training information.

A wide range of visual traits are used by humans in order to describe objects in images [44]. Nevertheless, learning and classifying visual traits of general object classes is not yet a current focus in computer vision, as evidenced in public image databases used for object analysis. Databases used for learning object class appearance with minimal supervision typically contain many images but small numbers of unique object class instances, e.g., 450 images of 26 different people [26] or

700 images of 10 different cell phones [10]. As a result, they are of limited use for learning and classifying traits over many different object class instances, a large population of people, for instance. Recent projects to create labeled databases such as PASCAL [45] or LabelMe [46] focus on labeling object identity and location, but provide little information regarding further traits. Even with accurate trait labeling, the number of images may be insufficient for learning traits in many cases. Consider the trait of motorcycle design, for instance. The PASCAL 2006 database [45] contains 275 motorcycle training instances while motorcycles can be classified into at least seven basic designs [47]. This leaves relatively few samples from which to learn motorcycle design traits from arbitrary viewpoints.

Due to the ubiquitous nature of face image analysis, one of the most common visual trait classification tasks is that of determining sex from face images. The wide range of published approaches to sex classification thus highlights the state-of-the-art in general trait classification. Trait learning has been tackled using spatially global feature representations such as templates [48], [6], principal components [5], independent components [4], or image intensities directly [2]. While most approaches utilize intensity data, 3D information may improve sex classification [49]. Much work has investigated the use of different machine learning techniques such as neural networks [6], SVMs [5], and boosted classifiers [2]. More recently, trait classification based on local features has emerged, using local regions [50] or Haar wavelets [3], [51]. In the interest of comparison, most approaches train and test on the standard FERET face database [13], which contains accurate labels for visual traits such as sex, age, and ethnicity.

To date, all published approaches to sex classification are based exclusively on single viewpoints, i.e., frontal faces [1], [2], [3], [4], [5], [6]. With the exception of [51], most approaches assume that, prior to classification, faces and facial features are precisely localized and background distraction such as hair and clothing is cropped away. For example, localization is performed by manually specifying eye locations [2] or using special-purpose frontal face alignment software [3], [5], and predefined facial masks are subsequently applied to remove background clutter. As a result, the reported classification error rates of 4-10 percent represent artificially low, ideal-case results. They offer little insight regarding classification performance in a general vision system where localization of faces and facial features required for classification is nontrivial. Indeed, a recent work evaluating the effect of artificial localization perturbations on classification accuracy showed that accuracy drops off rapidly with even small independent perturbations in scale and orientation (e.g., 5 degrees) [2]. An additional fact worth noting is that several published works reporting low error rates use different images of the same person in both classifier training and testing [3], [5]. As facial features arising from different frontal images of the same person are highly correlated, one cannot know whether the low classification error reported reflects the ability of the classifier to generalize to new, unseen faces or simply classification by recognition.

In the current literature, no work has yet addressed sex classification of faces from arbitrary viewpoints or in the presence of occlusion. Only a single approach has proposed a framework for general visual trait classification based on image features which can be detected and localized in the presence of partial occlusion [51], using boosted classifiers of Haar wavelet features [18]. The approach is single viewpoint (frontal faces), is not invariant to rotation, and the reported error rate of 0.21 reflects the increased difficulty of the combined task. Results obtained are based on proprietary training and testing databases, in which faces with ambiguous sex or in-plane orientations greater than 30 degrees are manually removed, and as such a direct comparison cannot be made.

3 Classifying Visual Traits of Faces

In realistic scenarios, visual trait classification is inseparable from detection and localization. Features must first be detected and localized before they can be used for classification. We propose embedding these three tasks within a general appearance model derived from local scale-invariant features, which can be used to detect, localize, and classify traits of faces in natural imagery captured from arbitrary viewpoints. By making use of recent research extending local invariant feature-based techniques to modeling 3D object classes [7], [8], [9], our approach is able to explicitly address visual traits from arbitrary viewpoints and in the presence of occlusion.

3.1 Viewpoint-Invariant Detection and Localization

Before visual traits can be learned or classified, the image features reflective of traits must first be detected and localized (i.e., associated with specific face instances) within the image, from arbitrary viewpoints and in the presence of partial occlusion. Recent literature contains several local feature-based models that offer a means by which this can be accomplished [7], [8], [9], the requirement being the ability to identify scale-invariant features in different images arising from the same underlying structure of the face. In this paper, we adopt the object class invariant (OCI) model, which was first presented as a means of viewpoint-invariant face detection in [7].

In the general case, the OCI model relates scale-invariant features to an OCI, an abstract 3D geometrical structure defined with respect to an underlying 3D object class, whose projection in the image plane maintains a consistent geometrical interpretation across different viewpoints and object class instances. The notion of an OCI is related to 3D primitives whose edges exhibit stable "nonaccidental" properties when projected onto the image plane from arbitrary views [39]. The projection of a 3D line segment, for instance, maintains a location, length, and orientation consistent with the 3D line segment from any viewpoint within a plane about the line. A 3D sphere projects to a 2D circle whose center and radius remain consistent with the sphere from arbitrary viewpoints. Early approaches to viewpoint-invariant detection focused on extracting these nonaccidental properties, or viewpoint-invariants, directly from the image. While this is feasible for classes of 3D shapes such as generalized cylinders [40] and planes [52],

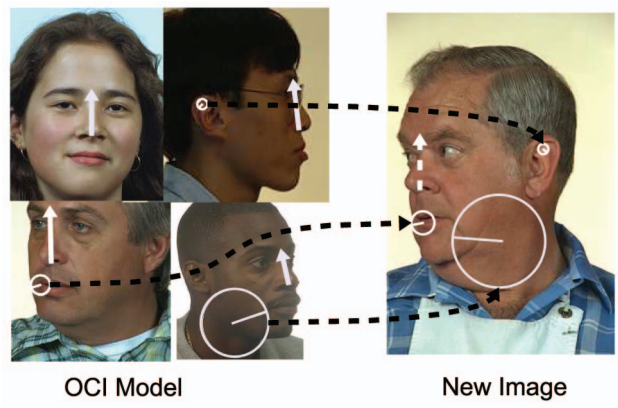

Fig. 2. The viewpoint-invariant OCI model relating scale-invariant features (white circles) to an OCI (solid white arrows). The OCI, defined here as a line segment from the base of the nose to the forehead, represents a viewpoint-invariant mechanism for grouping scale-invariant image features in images acquired from arbitrary viewpoints. A probabilistic model is learned from manually labeled OCIs in training images acquired from arbitrary viewpoints (four images to the left). Model instances can then be robustly detected and localized in a new image (right) acquired at an arbitrary viewpoint based on detected model features (dashed black lines) that agree on an OCI (dashed white arrow). Note that OCI shown here 1) exploits the symmetry of faces allowing mirror feature correspondence and 2) is not designed for overhead/underhead views.

viewpoint-invariants are difficult to extract directly from images and occur rarely in natural 3D objects [38].

Rather than extracting viewpoint-invariants directly, the OCI modeling methodology is to infer a viewpoint-invariant reference frame probabilistically from scale-invariant image features. A variety of viewpoint-invariant OCI parameterizations can be used, according to the degree of desired invariance. A 3D line segment can be used for object classes such as faces which are typically viewed from a coronal plane, as illustrated in Fig. 2. The projection of such an OCI maintains a location, orientation, and magnitude which are geometrically consistent with the head over a 360 degree range of viewpoint around the head (the OCI can be inferred even from rear views). The magnitude of the line segment vanishes in underhead or overhead views, however. Invariance from arbitrary viewpoints can be obtained via a 3D sphere centered about object class instances or a collection of orthogonal 3D line segments.

The appearance of a face can be described in terms of an OCI o and a set of scale-invariant image features $\{m_i\}$. A feature is denoted as $m_i : \{m_i^g, m_i^a, m_i^b\}$ and consists of variables of geometry m_i^g, appearance m_i^a, and occurrence m_i^b. Feature geometry $m_i^g : \{\sigma_i, \theta_i, x_i\}$ is a scale-invariant geometrical description of the feature in an image, including its scale σ_i, orientation θ_i, and absolute image (row, col) location x_i. Feature appearance m_i^a represents the image content within the region specified by the feature geometry and can be represented in a number of ways, e.g., principal components [15] or histograms of gradient orientations [20]. Feature occurrence m_i^b is a binary variable representing the presence or absence of a feature. The OCI is denoted as $o : \{o^g, o^b\}$ consisting of variables of geometry o^g and occurrence o^b. Geometry o^g is a viewpoint-invariant reference frame, which in the case of a 3D line segment OCI is equivalent to the geometry of a scale-invariant feature m_i^g, and o^b represents OCI presence or absence. Note that this OCI definition is similar to that of a model feature but lacks an appearance component, as an OCI is not directly observable and must be inferred from image data.

The relationship between OCI o and model features $\{m_i\}$ can be described probabilistically as

$$p(o|\{m_i\}) = \frac{p(o)p(\{m_i\}|o)}{p(\{m_i\})} = p(o)\frac{\prod_i p(m_i|o)}{p(\{m_i\})}, \qquad (1)$$

where the first equality results from Bayes rule and the second from the assumption of conditional feature independence given the OCI. With the conditional independence assumption, modeling focuses on $p(m_i|o)$ defining the relationship between an individual feature and the OCI, which can be expressed as

$$p(m_i|o) = p\left(m_i^a|m_i^b\right)p\left(m_i^b|o^b\right)p\left(m_i^g|o^b, o^g\right), \qquad (2)$$

under the assumptions of 1) conditional independence of feature appearance/occurrence $\{m_i^a, m_i^b\}$ and feature geometry $\{m_i^g\}$ given the OCI o, 2) conditional independence of feature appearance m_i^a and the OCI o given feature occurrence m_i^b, and 3) conditional independence of feature occurrence m_i^b and OCI geometry o^g given OCI occurrence o^b. In (2), $p(m_i^a|m_i^b)$ represents feature appearance given presence, and can be modeled as a Gaussian assuming additive noise. $p(m_i^b|o^b)$ represents the binomial probability of feature occurrence given reference frame occurrence. $p(m_i^g|o^b, o^g)$ represents the residual error in predicting the reference frame geometry from the feature geometry, and can be modeled as Gaussian assuming additive noise. Note that the scale parameters are treated in the log domain, and location parameters are normalized by the reference frame scale.

Learning the OCI model requires estimating the parameters of the distributions in (2) from data. This can be done applying a supervised learning technique to natural imagery acquired from arbitrary viewpoints as follows. First, OCIs are manually labeled in a set of training images, and scale-invariant features are automatically extracted in all training images. In the case of a linear OCI, labeling can be accomplished by drawing a line segment on training images which maintains a geometrically consistent interpretation across different viewpoints and face instances, e.g., the line from base of the nose to the forehead in face images shown in Fig. 2. With labeled OCIs and extracted features, learning proceeds by identifying clusters of features that agree in terms of their appearances and their geometries with respect to the OCI, where each such cluster represents a single underlying model feature m_i. As the number of clusters is initially unknown, iterative clustering techniques requiring initialization such as K-means [53] are infeasible. Instead, a robust clustering technique similar to the mean-shift algorithm [54] is used to identify dense clusters of features that agree in terms of appearance and geometry.

Clustering proceeds by treating each extracted feature as a potential model feature m_i. A feature m_j is said to agree geometrically with m_i if, when normalized according to their geometries m_i^g and m_j^g, their respective OCIs o_i^g and o_j^g

differ by less than a set of scale-invariant thresholds T^g: $\{T^\sigma, T^\theta, T^x\}$ in scale T^σ, orientation T^θ, and location T^x. These thresholds are applied independently; note that location difference is normalized according to the image scale of o_i^g and scale difference is calculated in the log domain. Features m_j that agree geometrically are considered as events $o^{b=1}$ and those that do not agree are considered as events $o^{b=0}$. Note that T^g represents the maximum acceptable error in predicting the OCI geometry, and thus, a single empirically determined threshold is applicable for all features.

Two features are said to agree in terms of appearance if the difference between their appearances m_i^a and m_j^a is less than an appearance threshold T_i^a. Features m_j that agree in appearance are considered as events $m_i^{b=1}$ and those that do not agree are considered as events $m_i^{b=0}$. Unlike the global geometrical threshold T^g, the appearance threshold T_i^a is feature-specific and determined by the image content of individual features. A low threshold will not capture the full range of appearance variability of geometrically similar features, whereas a high threshold will include the appearance ranges of geometrically unrelated features and lead to false correspondences. Here, T_i^a is automatically determined to maximize the likelihood ratio $\frac{p(m_i^{b=1}|o^{b=1})}{p(m_i^{b=1}|o^{b=0})}$, i.e., the ratio of geometrically agreeing versus disagreeing features. Note that this ratio can be considered a measure of feature distinctiveness [27]. After learning, poorly distinctive or redundant features can be pruned in order to improve performance.

Once learned, the OCI model can be used to automatically detect and localize faces in a new image as follows. Features are first extracted in the new image and matched to model features. An image feature m is said to match a model feature m_i if the difference in their appearance representations is less than the learned appearance threshold T_i^a. Each model-to-image match implies the geometry of an OCI o^g in the new image, and clusters of similar geometries o^g suggest the presence of a valid OCI. Different hypotheses o_i^g and o_j^g are considered as belonging to the same cluster if their difference is less than the global geometrical threshold T^g used in model learning. The hypotheses that a particular OCI cluster results from a true OCI instance $o^{b=1}$ or noise $o^{b=0}$ can be tested using a Bayes decision ratio

$$\begin{aligned}\gamma(o^g) &= \frac{p(o^g, o^{b=1}|\{m_i\})}{p(o^g, o^{b=0}|\{m_i\})}, \\ &= \frac{p(o^g, o^{b=1})}{p(o^g, o^{b=0})} \prod_i \frac{p(m_i|o^g, o^{b=1})}{p(m_i|o^g, o^{b=0})}.\end{aligned} \tag{3}$$

In (3), factor $\frac{p(o^g, o^{b=1})}{p(o^g, o^{b=0})}$ is a constant representing the prior ratio of valid versus invalid OCI o^g occurrences, and $\frac{p(m_i|o^g, o^{b=1})}{p(m_i|o^g, o^{b=0})}$ represents the likelihood ratio of a true versus false feature match.

An important issue is that of defining a suitable OCI reference frame when modeling faces or general object classes. While the OCI must maintain a consistent geometrical interpretation across viewpoints and object class instances, the particular OCI definition is arbitrary and can be specified in a number of ways. For the purpose of supervised learning, an OCI can be manually specified according to features of interest common to instances of a class. The line along the nose in Fig. 2 is easy to label when modeling faces in images acquired from arbitrary viewpoints around the head, for example. In terms of optimality, an OCI located centrally with respect to features in the image plane minimizes model localization error, as OCI localization error increases with the distance between the feature and the OCI origin [55]. An optimal OCI minimizing localization error can be derived in a data-driven manner by iteratively learning the model feature densities in (2), then reestimating OCI labels in training images by maximizing $\gamma(o^g)$ in (3). This iterative learning process is demonstrated later in experimentation in the context of face images.

3.2 Learning and Classifying Facial Traits

The OCI model described in the previous section can be learned from a set of natural, cluttered training images acquired from arbitrary viewpoints around the face, and used to identify occurrences of the same model features m_i in images of new faces. Our hypothesis is that some of these features bear information regarding visual traits, and that once detected and localized, can be used to classify faces. We propose training a classifier via a supervised learning procedure to classify faces in terms of visual traits, using the co-occurrence statistics of individual features with the trait of interest. To do this, we define $f_i = m_i^{b=1}$ to be the random event of positive occurrence of model feature i, and we expand the random event $o^{b=1}$ of positive OCI occurrence into a discrete random variable $c : \{c_1, \ldots, c_K\}$ over K trait values of interest, e.g., $sex : \{female, male\}$. A Bayesian classifier $\psi(c)$ can then be used to express the most probable trait classification given a set of M model feature occurrences $\{f_i\}$. Under the assumption that model features f_i are conditionally independent given trait c, this can be expressed as

$$\psi(c) = \frac{p(c|\{f_i\})}{p(\bar{c}|\{f_i\})} = \frac{p(c)}{p(\bar{c})} \prod_i^M \frac{p(f_i|c)}{p(f_i|\bar{c})},$$

or, equivalently,

$$\log \psi(c) = \log \frac{p(c)}{p(\bar{c})} + \sum_i^M \log \frac{p(f_i|c)}{p(f_i|\bar{c})}. \tag{4}$$

In (4), $\frac{p(c)}{p(\bar{c})}$ represents a prior ratio of trait value presence c versus absence $\bar{c}$ (e.g., male versus not male), controlling classifier bias toward different trait values, and $\frac{p(f_i|c)}{p(f_i|\bar{c})}$ expresses the likelihood ratio of trait value presence c versus absence $\bar{c}$ coinciding with observed feature f_i. The optimal Bayesian classification is to choose trait value c^* maximizing $\log \psi(c)$

$$c^* = \underset{c}{\operatorname{argmax}} \{\log \psi(c)\}. \tag{5}$$

Classifier training requires estimating $\frac{p(c)}{p(\bar{c})}$ and $\frac{p(f_i|c)}{p(f_i|\bar{c})}$ in (4). Features that are important to classification or highly

Fig. 3. Illustrating different instances of the same local feature bearing gender information (white circles), from data set with a male:female ratio of approximately 3:2. The forehead feature shown in (a) occurs in 30 males and three females and is indicative of male faces. The cheek feature shown in (b) occurs in zero males and eight females and is indicative of female faces. The nasal feature shown in (c) occurs in 16 males and 10 females and bears no information regarding sex.

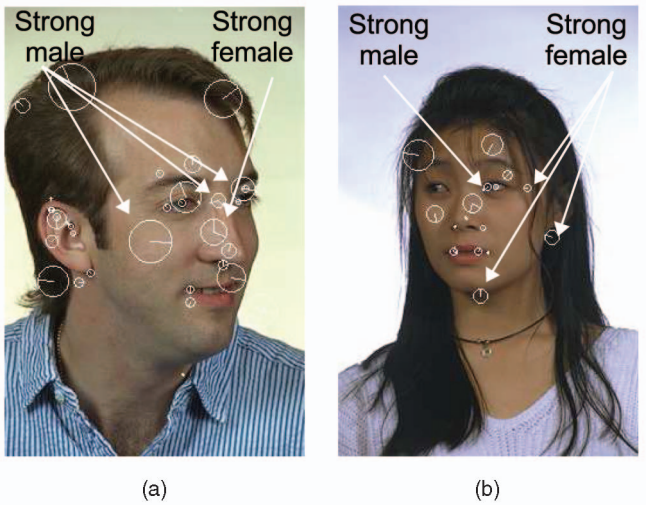

Fig. 4. Illustrating classification of the visual trait of sex from local features (white circles). A given face instance consists of a set of local features, a subset of which are reflective of either sex, and it is their ensemble which determines the final decision. To illustrate, we describe a feature as strongly male or female if its likelihood ratio of co-occurring with the indicated sex in training images is greater than 2:1. Of the 63 model features detected in (a), 15 are strongly male and one is strongly female, suggesting a male face. Of the 31 features detected in (b), seven are strongly female and one is strongly male, suggesting a female face. Many features, although very common in the class of face images, are uninformative regarding sex.

informative with regard to a particular trait value c_j will have high likelihood ratios, as illustrated in Fig. 3. The focus of our approach is to use these likelihood ratios to quantify the association of model features with visual traits, as illustrated in Fig. 4.

Estimating $\log\frac{p(f_i|c)}{p(f_i|\bar{c})}$**.** In order to estimate the likelihood ratios, we use a supervised learning process, based on observed model feature occurrences f_i and trait labels c_j for each training image. Discrete class-conditional likelihoods $p(f_i|c_j)$ can be represented as binomial distributions, parameterized by event counts [56]. During training, $p(f_i|c_j)$ is estimated from $p(c_j)$ and $p(f_i, c_j)$, the probability of observed joint events (f_i, c_j), using the definition of conditional probability

$$p(f_i|c_j) = \frac{p(f_i, c_j)}{p(c_j)}. \tag{6}$$

Term $p(c_j)$ is important in correcting bias in the training set. The most straightforward manner of estimating $p(f_i, c_j)$ is via maximum likelihood (ML) estimation, by counting the joint events (f_i, c_j) and normalizing with respect to their sum. ML estimation is known to be unstable in the presence of sparse data, leading to noisy or undefined parameter estimates. This is particularly true in models consisting of many local features, where feature occurrences are typically rare events. Bayesian maximum a posteriori (MAP) estimation can be used to cope with data sparsity, and involves regularizing estimates using a Dirichlet hyperparameter distribution [56]. In practice, Dirichlet regularization involves prepopulating event count parameters with samples following a prior distribution embodying assumptions regarding the expected sample distribution. Where no relevant prior knowledge exists, a uniform or maximum entropy prior can be used [57]. Although both ML and MAP estimates converge as the number of data samples increases, MAP estimation using a uniform prior will tend toward conservative parameter estimates while the number of data samples is low. The final estimator we use is

$$p(f_i|c_j) \propto \frac{k_{i,j}}{p(c_j)} + d_{i,j}, \tag{7}$$

where $k_{i,j}$ is the frequency of the joint occurrence event (f_i, c_j), $p(c_j)$ is the frequency of trait value c_j in the training data, and $d_{i,j}$ is the Dirichlet regularization parameter used to populate event counts. In the case of a uniform prior, $d_{i,j}$ is constant for all i, j. The proportionality constant for the likelihood in (7) can be obtained by normalizing over values of f_i, but is not required for likelihood ratios.

Estimating $\log\frac{p(c)}{p(\bar{c})}$**.** Although individual likelihood ratios have been corrected for training set bias by the estimator in (7), the Bayesian classifier in (4) will still exhibit bias due to the fact that the number of features f_i and their corresponding likelihood ratios associated with different traits are generally unequal. Given a set of M features to classify, $\log\psi(c)$ will be

higher a priori for trait values associated with a larger number of features or with features bearing higher likelihood ratios. This bias can be controlled by setting $\log\frac{p(c_j)}{p(\bar{c}_j)}$ for each trait value c_j such that the expected value of $\log\psi(c_j)$ based on a set of M features is zero

$$E[\log\psi(c_j)] = E\left[\log\frac{p(c_j)}{p(\bar{c}_j)} + \sum_i^M \log\frac{p(f_i|c_j)}{p(f_i|\bar{c}_j)}\right] = 0, \quad (8)$$

and thus,

$$\begin{aligned}\log\frac{p(c_j)}{p(\bar{c}_j)} &= -E\left[\sum_i^M \log\frac{p(f_i|c_j)}{p(f_i|\bar{c}_j)}\right] \\ &= -ME\left[\log\frac{p(f_i|c_j)}{p(f_i|\bar{c}_j)}\right],\end{aligned} \quad (9)$$

where the expectation in the right-hand side of (9) is taken with respect to the conditional probability of f_i given OCI occurrence $o^{b=1}$:

$$E\left[\log\frac{p(f_i|c_j)}{p(f_i|\bar{c}_j)}\right] = \sum_i p(f_i|o^{b=1})\log\frac{p(f_i|c_j)}{p(f_i|\bar{c}_j)}. \quad (10)$$

Thus, term $\log\frac{p(c_j)}{p(\bar{c}_j)}$ is the product of the expected likelihood ratio for trait c_j calculated during training from (10), and the number of features M associated with a detected face to be classified.

4 Experimentation

In this section, we present experimentation addressing the classification of face sex from arbitrary viewpoints and in the presence of occlusion. We begin by describing the experimental setup in Section 4.1. A qualitative view of model features as visual cues of sex is presented in Section 4.2. Quantitative experimentation consists of four main sections. Section 4.3 describes classification performance from arbitrary viewpoints while varying the amount of training data. Section 4.4 compares OCI modeling and Bayesian classification with the alternative approaches of bag-of-words modeling and SVM classification. Section 4.5 reports an analysis of classification performance in the presence of artificial occlusion. Finally, Section 4.6 details results for detection, localization, and sex classification on a subset of images from the cluttered CMU database.

4.1 Experimental Setup

Data. Our evaluation focuses on the performance of combined face detection, localization, and sex classification. All training is based on the standard, publicly available color FERET face image database [13]. Testing is performed on the FERET database and on a subset of the cluttered CMU profile database [12]. The FERET database consists of images of 994 unique subjects of various ethnicity, age, sex, acquired from various viewpoints, illumination conditions, with/without glasses, etc. The FERET database does not necessarily represent a challenging scenario for detection and localization; however, it is the standard for evaluating and comparing sex classification of faces. We build a database of 994 images, one for each FERET subject, where each subject image is chosen at random from a 180 degree viewpoint range (i.e., from left to right profile images). In this way, no subjects are duplicated in either testing or training data, in order to evaluate the generality of our approach. The male:female ratio in the database is approximately 3:2 (591:403). Fig. 5 shows the distributions of male and female data over viewpoint, which are consistent with the 3:2 ratio and do not exhibit significant sex-related bias. Images are converted to gray scale and processed at a resolution of 256×384 pixels.

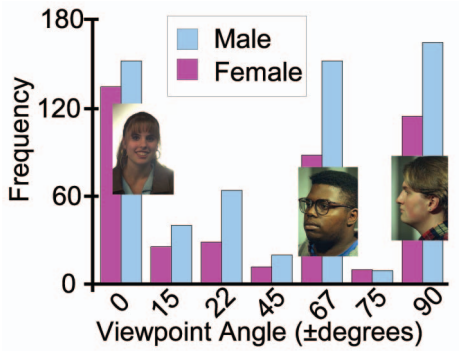

Fig. 5. The viewpoint distributions for the 403 female and 591 male unique FERET subject images used in experimentation. Note that, due to face symmetry, only the absolute value of the viewpoint is considered.

Scale-invariant feature extraction. Scale-invariant features are automatically extracted from all images used in training and testing. Although a variety of different features can be used, we use the scale-invariant feature transform (SIFT) technique [20] for feature detection and appearance description based on an implementation made public by the author. The SIFT feature detection method is based on identifying minima and maxima in a difference-of-Gaussian scale-space pyramid, and has been shown to outperform other techniques in terms of detection repeatability [58]. The SIFT appearance representation involves transforming the image content associated with features into a histogram of gradient orientations, and has been shown to be superior to other representations in terms of detection performance in a variety of natural image scenes [59].

Appearance model learning. As mentioned, recent literature contains several models that could be potentially used to learn the appearance of faces from local invariant features. We learn viewpoint-invariant OCI face models from training data using the procedure outlined in Section 3.1, with face OCIs manually labeled as line segments from the base of the nose to the forehead. To test the optimality of this OCI labeling, the iterative model learning and OCI reestimation procedure described in Section 3.1 is performed on a set of 500 FERET images. Fig. 6 illustrates how the OCI converges to an optimal value corresponding to a line located centrally within the head, thereby minimizing the distance between face features and the OCI. In [55], this optimal OCI definition results in face detection performance which is only marginally superior to that of the original OCI, indicating that the OCI labeled along the nose is already near-optimal for modeling the face.

Classifier training from modeled features. Once the OCI face model has been learned, model feature occurrences

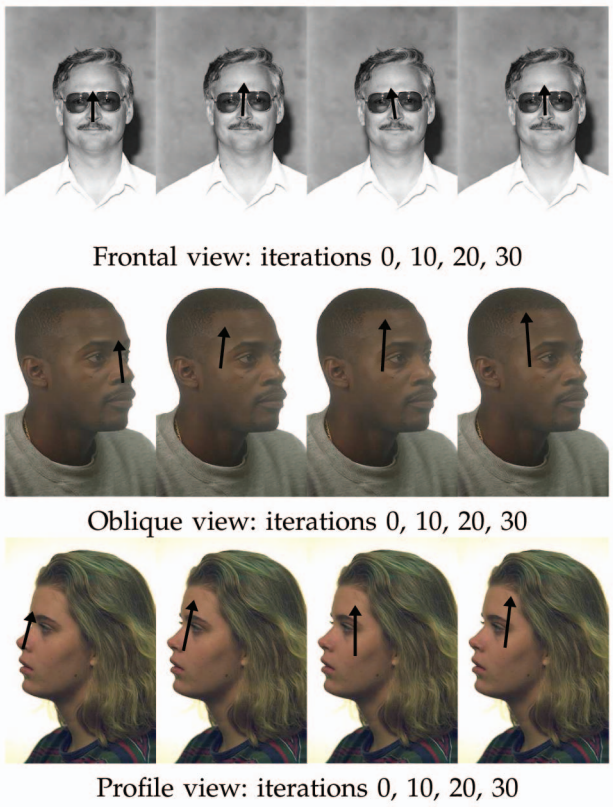

Fig. 6. The progression of iterative model learning and OCI reestimation in training images, for 0, 10, 20, and 30 iterations. In iteration 0, all OCIs are manually initialized as line segments from the base of the nose to the forehead. Little change occurs for OCIs after 30 iterations in frontal views, which are already approximately central to image features arising from the face. In oblique and profile views, OCI locations recede to the cheeks, minimizing the average distance to image features characteristic of these views (e.g., ears, cheeks, eyes). Note that the OCIs in all views remain consistent with 3D geometry of the face, corresponding to the 2D projections of the same 3D line segment located within the head.

identified in training images along with FERET sex labels are used to estimate likelihood ratios of the Bayesian trait classifier, as described in Section 3. In estimating likelihood ratios via (7), we used an empirically determined Dirichlet regularization parameter of $d_{i,j} = 2$, which maximizes training set classification performance.

Detecting and localizing faces in new images. Once the appearance model and classifier have been learned, fully automatic detection, localization, and classification proceed on testing images, as described in Section 3.1. Viewpoint-invariant face detection and localization are performed by determining the OCI instances in each of the testing images, maximizing the Bayes decision ratio in (3). The threshold values of T^g used to evaluate geometrical consistency for OCI localization are $T^\sigma = \log(1.5)$, $T^\theta = 20$ degree, and $T^x =$ OCI scale/2.

Classifying faces in new images. Once an OCI instance is detected in a new image, model features associated with the instance are then used to determine sex using the Bayesian classifier in (4). As faces are either male or female, determining face sex is a two-class problem and, thus, $\psi(male) = \psi(female)^{-1}$. A single-threshold ψ^* on $\psi(c)$ can be used such that faces are classified as either male if $\log \psi(male) > \psi^*$ or as female if $\log \psi(male) < \psi^*$. Classification results are reported in terms of the equal error rate (EER), i.e., the threshold at which the probabilities of misclassifying males and females are equal.

4.2 Identifying Visual Cues of Sex

As humans, we are generally capable of describing faces in terms of visual traits such as sex or age, however, it is often difficult to identify the visual cues that are operative in determining these traits. Most faces contain a variety of cues that could be construed as either male or female, and it is their ensemble which determines the final decision. The local feature-based approach provides insight in terms of what local image cues are most important in determining visual traits, insight which is not possible from other representations, e.g., global features or templates. By sorting features according to their likelihood ratios, the image regions most telling regarding the trait of sex can be visualized as in Fig. 7. Features on the ears and forehead are often indicative of males, as they are less visible in females due to generally longer female hair. Different features around the mouth, eyes, and cheeks can be strongly indicative of either males or females, possibly due to sex differences relating to facial stubble, makeup, and hair. In many cases, the relationship between facial features and sex may not be obvious or easily explained. Certain model features arising from the nose or cheeks, although very common in faces, are uninformative regarding sex. Note that although the male:female ratio in training data is 3:2, approximately twice as many sex-related features are identified for males as for females, suggesting a greater number of visual cues characteristic of the male sex.

4.3 Classifying Sex from Arbitrary Viewpoints

In order to evaluate sex classification from arbitrary viewpoints, 15 different trials of training, localization, and classification are performed. Five training set sizes of 100, 200, 300, 400, and 500 face images are used, and for each size, three training sets are randomly selected from 994 images. In this way, both cross validation and training efficacy can be investigated. Fig. 8 illustrates the classification error as a function of training set size. As expected, classification error decreases with an increase in the number of training data. This is primarily due to the emergence of more rare, yet sex informative, features. Examples of correctly classified faces are shown in Figs. 9a, 9b, and 9c, and examples of incorrectly classified faces are shown in Figs. 9d, 9e, and 9f. For trials based on 500 training images, approximately 3.6 percent of error cases are due to poor model localization, where the discrepancy between the localized and labeled OCIs is greater than the geometrical consistency threshold T^g.

Table 1 illustrates the distribution of classification EER over three ranges of viewpoint for trials involving 500 subjects, ranging from frontal (0 degree-22 degree) to profile (67 degree-90 degree) views. Note that the error rate for profile views is almost double that of frontal views, reflecting the difficulty of classifying nonfrontal faces. This appears to be due to the fact that nonfrontal views generally contain fewer model features. Note also that the EER for frontal views here (11.9 percent) is somewhat higher than error rates obtained by other frontal face classifiers (4-10 percent) [2], [3], [4], [5], [6].

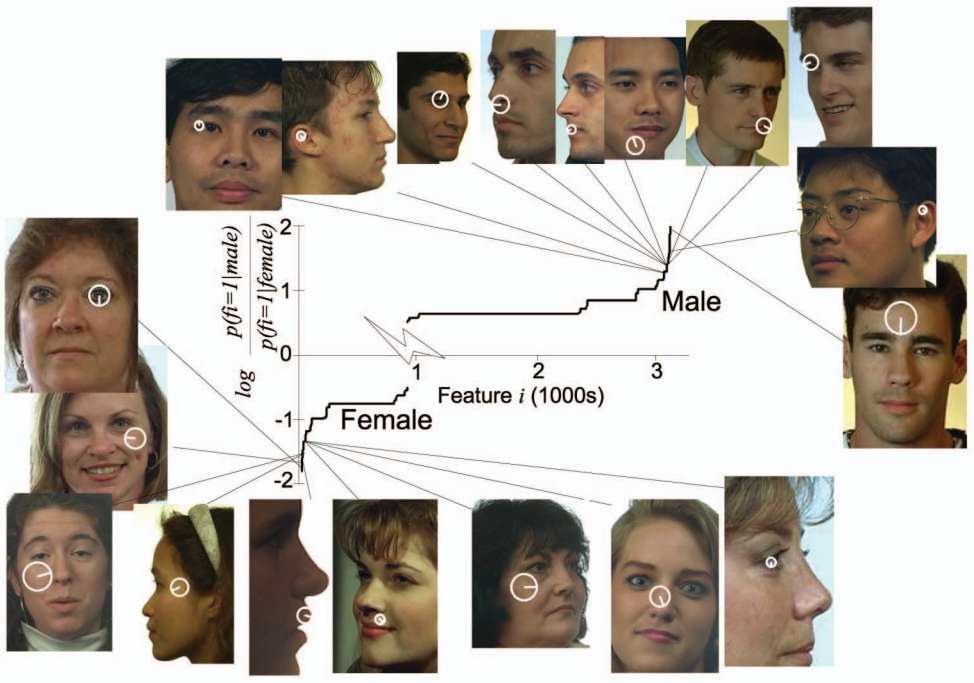

Fig. 7. Visual cues indicative of face sex, in the form of scale-invariant features. Features are sorted in increasing order of their log likelihood ratio $\log \frac{p(f_i=1|male)}{p(f_i=1|female)}$. Of approximately 15,000 features in a viewpoint-invariant face model learned from 500 randomly selected FERET images, approximately 3,000 features bear information regarding sex (i.e., $|\log \frac{p(f_i=1|male)}{p(f_i=1|female)}| > 0.5$). Features in the lower left occur more frequently in females, and features in the upper right occur more frequently in males. Face images shown illustrate instances of sex-informative features (white circles) with absolute log likelihood ratios ranging from 1.3 to 2.0. Although the male:female ratio in the training data is 3:2, approximately twice as many sex-reflective features are associated with males. Note that SIFT features shown generally arise from image content in a region slightly larger than the circles indicated in the images.

4.4 Comparison to Alternative Techniques

Given that FERET images contain faces which are reasonably centered within the image and relatively little background clutter, several questions arise regarding sex classification performance. First, how important is the localization of SIFT features via the OCI model to the classification result? Geometry-free bag-of-words (BOW) models [27], [60], [28] can be computed independently of viewpoint, and could potentially be used to classify FERET faces acquired from arbitrary viewpoints according to sex. Second, how effective is the Bayesian classification technique? Sex classification could be achieved by applying other black-box methods such as SVMs [61] to features identified by either the OCI or BOW models.

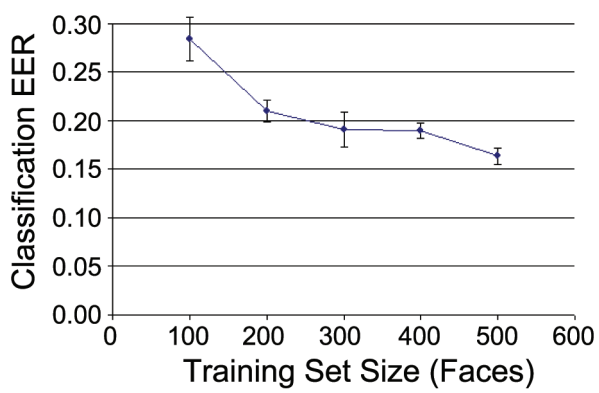

Fig. 8. The classification equal error rate (EER) as a function of the number of faces used in training. For each training set size, three different sets of training data are randomly selected and classification is performed on the remaining face images not used in training. The points and error bars indicate the EER mean and the standard deviation for the indicated training set size. Note that the mean EER generally diminishes with an increase in training set size. The minimum mean EER achieved here is 16.3 percent for a training set size of 500 images.

We investigate these questions by experimenting with both BOW modeling and SVM classification. BOW models are constructed from training images and used as a basis for sex classification. This is done by clustering SIFT features using the K-means algorithm, thereby defining a set of model features or "visual words." K-means requires defining the number of words, values of 1,000, 5,000, 10,000; and 15,000 words are used, in the same order of magnitude of values proposed in the literature [60], [28]. The BOW model is fit to new images by matching SIFT features extracted in the new image to their nearest neighbor model features, based on the Euclidean distance, and binary vectors of BOW feature occurrence are used in classification. Classification performance generally improves going from 1,000 to 10,000 words, after which 10,000 and 15,000 word models result in similar performance. Results for 15,000 word BOW models are reported here.

Sex classification is also performed via the SVM technique. Briefly, SVM classification is based on identifying hyperplanes which maximally separate feature data arising from different classes. The input to the SVM

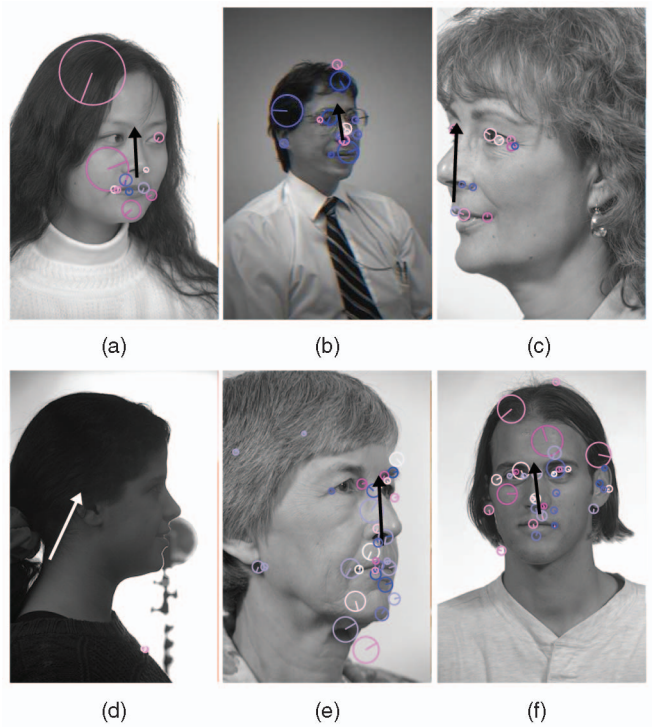

Fig. 9. Examples of correctly and incorrectly classified faces in trials involving 500 training images. Features shown are those detected and used in classification, and arrows indicate detected OCIs. Blue features indicate male characteristics and pink features indicate female characteristics, where the color saturation is proportional to the magnitude of the log likelihood ratio. Note that most faces contain features indicative of both male and female sexes. The top row illustrates correct classifications, where images (a) and (c) are female faces and image (b) is a male face. The bottom row illustrates incorrect classifications. Image (d) is misclassified as male due to model localization failure. Here, strong backlighting results in poor image contrast in the face, and thus few scale-invariant features. Image (e) is a female face misclassified as male, due to an excess of strong male characteristics. Image (f) is a male face misclassified as female, due to an excess of female features.

classifier is equivalent to that of the Bayesian classifier, i.e., a set of positive model feature occurrences for each face image to be classified along with sex labels for classifier training. The SVM is defined by a kernel function, two popular choices are the radial basis function (RBF) kernel [5] and the linear kernel [60]. Both kernels involve a free parameter C relating to the classification error margin and the RBF requires an parameter γ defining the width of the RBF kernel. A search over kernels and parameters is performed in order to determine the parameter combination which results in the best SVM sex classification performance. The optimal SVMs are RBF (parameters $\gamma = 1.0e-13$, $C = 1.0e12$) for BOW classification and linear (parameter $C = 20$) for OCI classification. The

TABLE 1
The Data Distribution and Mean EERs for Three Ranges of Face Viewpoint in All Trials Based on 500 Training Images

Viewpoint range	$(0° - 22°)$	$(22° - 67°)$	$(67° - 90°)$
Data %	33.3%	36.4%	30.2%
Mean EER	11.9%	15.6 %	19.9%

TABLE 2
Sex Classification EERs for Combinations of Models (BOW, OCI) and Classifiers (SVM, Bayesian)

Model	Classifier	EER
BOW	SVM	$24.7\% \pm 1.9\%$
BOW	Bayesian	$23.5\% \pm 1.3\%$
OCI	SVM	$18.8\% \pm 1.8\%$
OCI	Bayesian	$16.3\% \pm 0.9\%$

SVM implementation used is the open source LIBSVM package [62].

Experimentation tests all combinations of models (BOW, OCI) and classifiers (SVM, Bayesian). All trials are based on three random partitions of the data into 500 training and 494 testing images, and the results are shown in Table 2. In general, classification based on the OCI model is superior to that of the BOW model for either classifier. This suggests that the localizing features geometrically within a reference frame such as the OCI is important for classification. Furthermore, while Bayesian and SVM classification are similar for the BOW model, Bayesian classification is marginally superior for the OCI model.

Interestingly, while Bayesian classification outperforms SVMs for determining sex from face images, the opposite has been found for classifying images according to distinctly different scene categories such as faces, buildings, and cars [60]. One hypothesis for this difference is as follows: SVMs generally exploit dependencies between different image features when determining hyperplanes separating data. Such interfeature dependencies are likely more prominent in distinctly different scene categories than in face images of different sexes, where features tend to occur infrequently and independently in both sexes, and may be missing due to occlusion. A Bayesian classifier based on the assumption of conditionally independent features avoids relying on feature interdependencies, and therefore, results in improved performance.

4.5 Classifying Sex from Occluded Faces

Classifying faces according to visual traits in arbitrary scenes is complicated by occlusion, as features useful for classification may not be visible. In the case of face images, occlusion can arise from a number of factors, such as sunglasses, hats, hairstyles, scarves, crowds. The effect of occlusion has not been previously investigated in the context of face classification, as most previous work has assumed that facial features required for classification are visible and precisely localized. Classification based on local features is capable of coping with a significant degree of occlusion, as only a subset of features is required.

We test occlusion by artificially obscuring each FERET testing image with a black occluding circle, and then performing classification trials using the three classifiers trained on 500 images described in the previous section. The black circle is placed in the center of the images, thereby obscuring a variety of different facial regions in different images, as faces are approximately but not precisely centered in the FERET dataset. The degree of occlusion is varied by changing the radius of the occluding circle from 0 to 80 pixels. The occluding circle border is blended

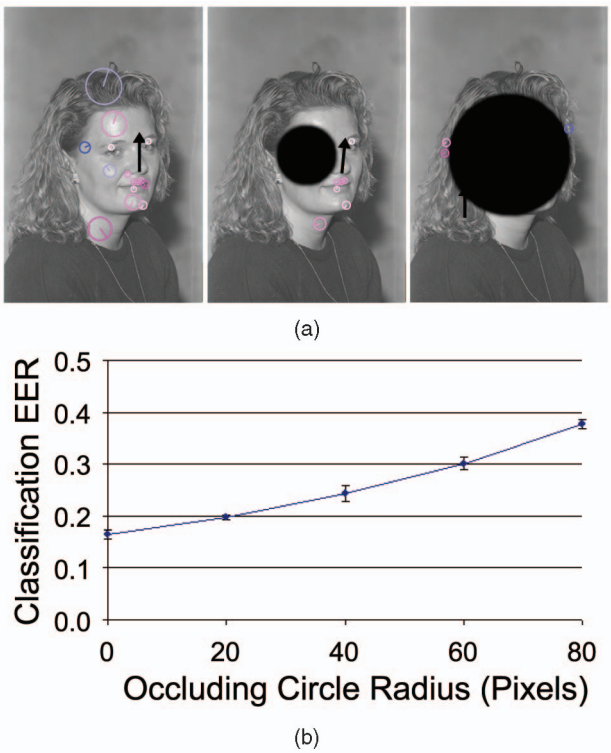

Fig. 10. (a) Examples of detection and localization for occluding circles of radii of 0, 40, and 80 pixels (left to right). (b) Classification error as a function of the degree of occlusion. For each occluding circle radius, three different classification trials are performed using the three classifiers in the previous section trained on 500 images, based on occluded images not used in training. The points and error bars indicate the mean and the standard deviations of the classification error for the indicated radius. Classification error starts at 16.3 percent with no occlusion and rises to approximately 37.7 percent for a radius of 80 pixels.

smoothly into the face images using a small Gaussian kernel with a standard deviation of 2 pixels, to simulate a more natural occluding contour.

Fig. 10 illustrates examples of occlusion and the classification error as a function of occluding circle radius. Note that classification performance degrades gracefully with an increase in the occluding circle radius, as sex-informative features can still be extracted and used for classification in nonoccluded regions of the face. Even at a reasonably large occluding circle of radius 40 pixels, classification error is approximately 25 percent. At an occluding radius of 80 pixels, classification error reaches approximately 0.4, the rate of female faces in the data set. Note that the standard deviation of classification error does not generally change significantly with the degree of occlusion, indicating that error variability is determined by the amount of training data and not by the number of features identified in the image.

4.6 Detecting, Localizing, and Classifying Faces in Clutter

In this section, we investigate localization, detection, and classification of faces in cluttered data. Despite the ubiquitous nature of face imagery, there is currently no standard database of cluttered faces captured from arbitrary viewpoints with sex labels. We thus perform experimentation using CMU profile database, a challenging database of cluttered face imagery used to benchmark face detection performance [12]. OCI model detection and localization performance is limited by the underlying feature detector, and modeling based on SIFT features is ineffective for low resolution faces which produce few detector responses. We thus select a subset of the largest CMU faces for testing as follows: We select all images from the CMU profile database containing a face with a distance of 19 pixels or greater from the eye to the nose, as determined by ground truth profile face labels. From this image set, we consider all faces (profile or otherwise) for which the distance from the point between the eyes to the nose is 19 pixels or greater. This results in a set of 132 faces, which is manually determined to contain 100 males and 32 females. Note the bias toward the male sex, from inspection the male:female ratio over the entire CMU profile data set appears to be approximately 6:1.

Detection and localization of CMU faces are performed using an OCI model learned from 500 FERET faces. Sex classification is then performed on faces correctly localized with respect to ground truth labels, using Bayesian and SVM classifiers trained on 500 FERET faces. We evaluate sex classification on the sets of 25, 50, and 100 faces detected with the highest precision, in order to illustrate the degradation of sex classification performance with detection precision. Fig. 11a illustrates the precision-recall characteristic of detection, along with sex classification error rates at the indicated values of precision. In order to prune multiple detection hypotheses arising from the same face, all hypotheses within the geometrical threshold T^g of hypotheses bearing locally maximal Bayes decision ratios are removed. Classification performance generally degrades with decreasing detection precision, as both are linked to the number of images features extracted in each localized face. Here, Bayesian classification generally outperforms SVM classification by larger margins than in trials on FERET imagery, suggesting that Bayesian classification generalizes more readily to the context of cluttered imagery. Fig. 11b illustrates an example of OCI detection, localization, and classification in a cluttered scene containing faces in arbitrary viewpoints and partial occlusion.

5 Discussion

In this paper, we present a general approach to learning and classifying visual traits of faces from arbitrary viewpoints and in the presence of occlusion. As a realistic classification scenario requires first detecting and localizing faces and associated features prior to trait classification, we base our classifier on a viewpoint-invariant appearance model of local scale-invariant features, which can be used to detect and localize faces in images acquired from arbitrary viewpoints. A Bayesian visual trait classifier is constructed from modeled features, where classifier training involves estimating the likelihood ratios of model feature occurrence given trait presence versus absence. Features associated with significantly nonzero likelihood ratios can be interpreted as visual cues reflective of the trait of interest.

We present the first experimental results for face sex classification from arbitrary viewpoints, based on the

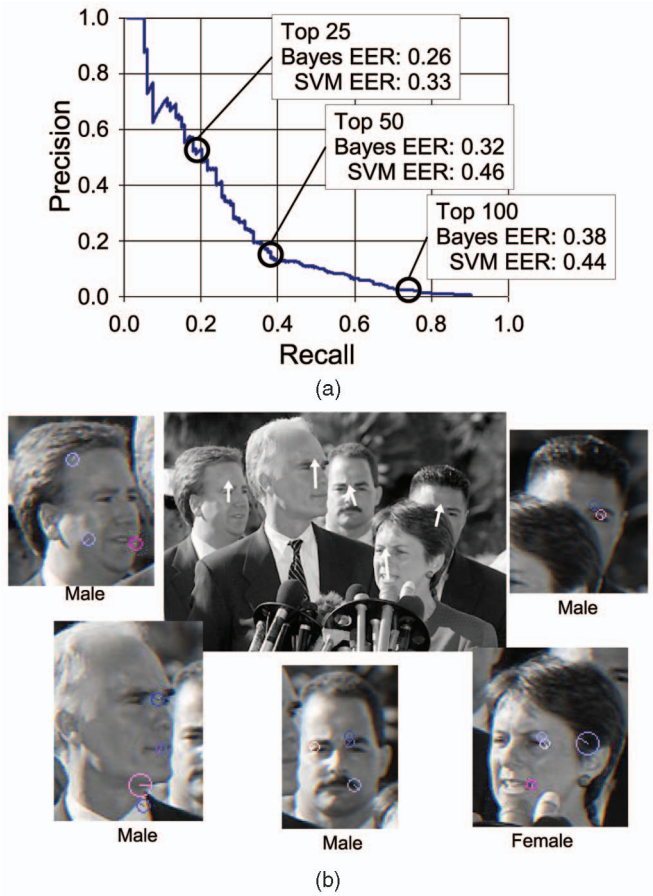

Fig. 11. (a) The precision-recall characteristic of detection and localization based on 132 faces of the CMU database. Classification EERs are displayed for the top 25, 50, and 100 detected faces. (b) An example of correctly localized and classified faces in a cluttered scene. White arrows overlaying the central image indicate correctly identified OCIs. Features overlaying thumbnails indicate instances of model features involved in localization and classification for each face. Blue features indicate male characteristics and pink features indicate female characteristics, where the color saturation is proportional to the magnitude of the log likelihood ratio. Note that successful classification is possible despite the high degree of occlusion in the rightmost face.

standard FERET database, obtaining an EER of 16.3 percent for over a 180 degree range of face viewpoint. Classification error is lowest in frontal views, and error for profile views is approximately twice that obtained for frontal views. The EER of 11.9 percent obtained by our classifier for frontal faces is higher than that of other approaches, suggesting that the resistance to occlusion offered by the local feature approach may come at the cost of slightly reduced classification performance under ideal circumstances. We present results of sex classification in the presence of simulated occlusion. Classification error degrades smoothly with an increase in the degree of occlusion, demonstrating the capacity of local feature-based classification to cope with missing features. Quantitative comparisons with geometry-free BOW model show that sex classification is significantly improved by the geometrical constraints afforded by the OCI model. Furthermore, Bayesian sex classification results in lower error than SVM classification, particularly in the cluttered, occluded scenes from the CMU profile database. This runs contrary to classification of general scenes, where SVM classification has been shown to be superior [60].

The framework we present is general and may prove useful in modeling and classifying a variety of different object classes and/or visual traits. We experimented with learning the trait of age, by dividing faces into less than/greater than 25 years of age, splitting the FERET data set approximately evenly. A somewhat high classification error rate of 23 percent was obtained from the framework trained on 200 frontal faces, indicating that age classification is a more difficult problem than sex. We also applied the general OCI framework to modeling brain anatomy in magnetic resonance imagery using an OCI defined according to a standard neuroanatomical reference frame [63], and subsequently achieved a brain sex classification accuracy of ≈ 80 percent.

The effectiveness of our framework in the general case depends on the appearance characteristics of the object class and the traits to be modeled. Viewpoint-invariant OCI modeling is most effective for object classes which produce similar distinctive image features across different instances, e.g., bicycles or cars. For these two classes, we have achieved similar detection performance with fewer training images than the method of [10] which models viewpoint information explicitly. As trait classification follows detection and localization, the cardinality of training sets required for classification is generally equal to or greater than for detection alone, and ultimately depends on the degree of information shared between image features and trait values. To illustrate, we investigated detection and design classification of motorcycles from the PASCAL 2006 database [45], using an OCI in the form of a sphere centered on the motorcycle. The average detection precision obtained was 0.159, which lies within the range of [0.153, 0.390] reported for other methods and could potentially be improved by further investigating feature selection techniques such as boosting. Classification was then performed based on motorcycle design labels of sport:offroad:moped:other, these were feasible to label manually from prototype examples provided in [47]. The numbers of instances were 56:32:21:166 in training and 81:37:21:135 in testing, and the classification EERs for the 50 testing motorcycles detected with the highest precision were sport $= 0.45$, offroad $= 0.4$, and moped $= 0.3$. These results were encouraging, particularly given the small numbers of training instances, the range of intradesign variability over viewpoint and interdesign ambiguity. Here, the moped was qualitatively most distinctive in appearance and also most effectively classified, while sport and offroad designs were qualitatively less distinct (e.g., due to design hybridization) and more difficult to classify.

Various future avenues exist for detection and visual trait classification from local scale-invariant features. The computational complexity of detection, localization, and classification is low, and the combined system should be implementable in real or near-real time. Continuous-valued traits such as age could potentially be modeled using continuous-valued likelihoods in a regression framework. Facial traits such as age or emotion could be modeled and used as a soft biometric in interactive image-based

applications, surveillance, or recognition. Classification performance could be potentially improved by incorporating feature types other than SIFT which offer complementary information [64] or by using alternative techniques to identify learned features more reliably, e.g., sliding-window-based scanning. In particular, reliable invariant feature extraction in low-image resolution images would improve modeling for small faces and objects. Whether multiple traits such as age and sex are best modeled independently or jointly is an open research question. The Bayesian classifier we present could be used for either, however, joint modeling may be computationally complex for large numbers of different traits.

REFERENCES

[1] E. Makinen and R. Raisamo, "Evaluation of Gender Classification Methods with Automatically Detected and Aligned Faces," *IEEE Trans. Pattern Analysis and Machine Intelligence,* vol. 30, no. 3, pp. 541-547, Mar. 2008.

[2] S. Baluja and H.A. Rowley, "Boosting Sex Identification Performance," *Int'l J. Computer Vision,* vol. 71, no. 1, pp. 111-119, 2007.

[3] Z. Yang, M. Li, and H. Ai, "An Experimental Study on Automatic Face Gender Classification," *Proc. Int'l Conf. Pattern Recognition,* pp. 1099-1102, 2006.

[4] A. Jain, J. Huang, and S. Fang, "Gender Identification Using Frontal Facial Images," *Proc. IEEE Int'l Conf. Multimedia and Expo,* pp. 1082-1085, 2005.

[5] B. Moghaddam and M. Yang, "Learning Gender with Support Faces," *IEEE Trans. Pattern Analysis and Machine Intelligence,* vol. 24, no. 5, pp. 707-711, May 2002.

[6] S. Gutta, H. Wechsler, and P. Phillips, "Gender and Ethnic Classification of Human Faces Using Hybrid Classifiers," *Proc. Int'l Conf. Automatic Face and Gesture Recognition,* pp. 194-199, 1998.

[7] M. Toews and T. Arbel, "Detection over Viewpoint via the Object Class Invariant," *Proc. Int'l Conf. Pattern Recognition,* vol. 1, pp. 765-768, 2006.

[8] A. Thomas, V. Ferrari, B. Leibe, T. Tuytelaars, B. Schiele, and L. Van Gool, "Towards Multi-View Object Class Detection," *Proc. IEEE Conf. Computer Vision and Pattern Recognition,* pp. 1589-1596, 2006.

[9] A. Kushal, C. Schmid, and J. Ponce, "Flexible Object Models for Category-Level 3D Object Recognition," *Proc. IEEE Conf. Computer Vision and Pattern Recognition,* 2007.

[10] S. Savarese and L. Fei-Fei, "3D Generic Object Categorization, Localization and Pose Estimation," *Proc. IEEE Int'l Conf. Computer Vision,* 2007.

[11] P. Yan, S.M. Khan, and M. Shah, "3D Model Based Object Class Detection in an Arbitrary View," *Proc. IEEE Int'l Conf. Computer Vision,* 2007.

[12] CMU Face Group "Frontal and Profile Face Databases," http://vasc.ri.cmu.edu/idb/html/face/, 2009.

[13] "Color FERET Face Database," www.itl.nist.gov/iad/humanid/colorferet, 2009.

[14] M. Toews and T. Arbel, "Detecting, Localizing and Classifying Visual Traits from Arbitrary Viewpoints Using Probabilistic Local Feature Modeling," *Proc. IEEE Workshop Analysis and Modeling of Faces and Gestures,* pp. 154-167, 2007.

[15] M. Turk and A.P. Pentland, "Eigenfaces for Recognition," *J. Cognitive Neuroscience,* vol. 3, no. 1, pp. 71-96, 1991.

[16] S. Ullman, M. Vidal-Naquet, and E. Sali, "Visual Features of Intermediate Complexity and Their Use in Classification," *Nature Neuroscience,* vol. 5, pp. 682-687, 2002.

[17] B. Heisele, T. Serre, and T. Poggio, "A Component-Based Framework for Face Detection and Identification," *Int'l J. Computer Vision,* vol. 74, no. 2, pp. 167-181, 2007.

[18] P. Viola and M. Jones, "Rapid Object Detection Using a Boosted Cascade of Simple Features," *Proc. IEEE Conf. Computer Vision and Pattern Recognition,* pp. 511-518, 2001.

[19] M. Jones and P. Viola, "Fast Multi-View Face Detection," Technical Report tr2003-96, pp. 1-10, 2003.

[20] D.G. Lowe, "Distinctive Image Features from Scale-Invariant Keypoints," *Int'l J. Computer Vision,* vol. 60, no. 2, pp. 91-110, 2004.

[21] G. Carneiro and A. Jepson, "Multi-Scale Phase-Based Local Features," *Proc. IEEE Conf. Computer Vision and Pattern Recognition,* vol. 1, pp. 736-743, 2003.

[22] K. Mikolajczyk and C. Schmid, "Indexing Based on Scale Invariant Interest Points," *Proc. IEEE Int'l Conf. Computer Vision,* vol. 1, pp. 525-531, 2001.

[23] T. Kadir and M. Brady, "Saliency, Scale and Image Description," *Int'l J. Computer Vision,* vol. 45, no. 2, pp. 83-105, 2001.

[24] F. Mindru, T. Tuytelaars, L. Van Gool, and T. Moons, "Moment Invariants for Recognition under Changing Viewpoint and Illumination," *Computer Vision and Image Understanding,* vol. 94, nos. 1-3, pp. 3-27, 2004.

[25] H. Schneiderman and T. Kanade, "Object Detection Using the Statistics of Parts," *Int'l J. Computer Vision,* vol. 56, no. 3, pp. 155-177, 2004.

[26] R. Fergus, P. Perona, and A. Zisserman, "Weakly Supervised Scale-Invariant Learning of Models for Visual Recognition," *Int'l J. Computer Vision,* vol. 71, no. 3, pp. 273-303, 2006.

[27] G. Dorko and C. Schmid, "Selection of Scale-Invariant Parts for Object Class Recognition," *Proc. IEEE Int'l Conf. Computer Vision,* pp. 634-640, 2003.

[28] J. Sivic, B.C. Russell, A.A. Efros, A. Zisserman, and W.T. Freeman, "Discovering Objects and Their Localization in Images," *Proc. IEEE Int'l Conf. Computer Vision,* pp. 370-377, 2005.

[29] D. Crandal, P. Felzenswalb, and D. Huttenlocher, "Spatial Priors for Part-Based Recognition Using Statistical Models," *Proc. IEEE Conf. Computer Vision and Pattern Recognition,* vol. 1, pp. 10-17, 2005.

[30] E. Bart, E. Byvatov, and S. Ullman, "View-Invariant Recognition Using Corresponding Object Fragments," *Proc. European Conf. Computer Vision,* pp. 152-165, 2004.

[31] P.F. Felzenszwalb and D. Huttenlocher, "Pictorial Structures for Object Recognition," *Int'l J. Computer Vision,* vol. 61, no. 1, pp. 55-79, 2005.

[32] G. Carneiro and D.G. Lowe, "Sparse Flexible Models of Local Features," *Proc. European Conf. Computer Vision,* vol. 3, pp. 29-43, 2006.

[33] E. Sudderth, A. Torralba, W. Freeman, and A. Willsky, "Describing Visual Scenes Using Transformed Objects and Parts," *Int'l J. Computer Vision,* vol. 77, nos. 1-3, pp. 291-330, 2008.

[34] B. Leibe, A. Leonardis, and B. Schiele, "Combined Object Categorization and Segmentation with an Implicit Shape Model," *Proc. ECCV Workshop Statistical Learning in Computer Vision,* 2004.

[35] M. Weber, M. Welling, and P. Perona, "Unsupervised Learning of Models for Recognition," *Proc. European Conf. Computer Vision,* vol. 1, pp. 18-32, 2000.

[36] I. Beiderman and P.C. Gerhardstein, "Recognizing Depth-Rotated Objects: Evidence and Conditions for 3D Viewpoint Invariance," *J. Experimental Psychology: Human Perception and Performance,* vol. 19, pp. 1162-1182, 1993.

[37] J. Koenderink, "The Internal Representation of Solid Shape with Respect to Vision," *Biological Cybernetics,* vol. 32, no. 4, pp. 211-216, 1979.

[38] J. Burns, R. Weiss, and E. Riseman, "The Non-Existence of General-Case View-Invariants," *Geometric Invariance in Computer Vision,* MIT Press, pp. 120-131, 1992.

[39] I. Beiderman, "Recognition-by-Components: A Theory of Human Image Understanding," *Psychological Rev.,* vol. 94, no. 2, pp. 115-147, 1987.

[40] J. Ponce, D. Chelberg, and W. Mann, "Invariant Properties of Straight Homogeneous Generalized Cylinders and Their Contours," *IEEE Trans. Pattern Analysis and Machine Intelligence,* vol. 11, no. 9, pp. 951-966, Sept. 1989.

[41] R. Fergus, P. Perona, and A. Zisserman, "Object Class Recognition by Unsupervised Scale-Invariant Learning," *Proc. IEEE Conf. Computer Vision and Pattern Recognition,* pp. 264-271, 2003.

[42] N. Ahuja and S. Todorovic, "Learning the Taxonomy and Models of Categories Present in Arbitrary Images," *Proc. IEEE Int'l Conf. Computer Vision,* pp. 1-8, 2007.

[43] B.C. Russell, A.A. Efros, J. Sivic, W.T. Freeman, and A. Zisserman, "Using Multiple Segmentations to Discover Objects and Their Extent in Image Collections," *Proc. IEEE Conf. Computer Vision and Pattern Recognition,* pp. 1605-1614, 2006.

[44] L. Fei-Fei, A. Iyer, C. Koch, and P. Perona, "What Do We Perceive in a Glance of a Real-World Scene?" *J. Vision,* vol. 7, no. 1, pp. 1-29, 2007.

[45] M. Everingham, A. Zisserman, C.K.I. Williams, and L. Van Gool, "The PASCAL Visual Object Classes Challenge 2006 (VOC2006) Results," www.pascal-network.org/challenges/VOC/voc2006/results.pdf, 2009.
[46] B.C. Russell, A. Torralba, K.P. Murphy, and W.T. Freeman, "LabelMe: A Database and Web-Based Tool for Image Annotation," *Int'l J. Computer Vision,* vol. 77, nos. 1-3, pp. 157-173, 2008.
[47] Motorcycle Safety Foundation Inc., *Basic Rider Course Handbook,* seventh ed., www.msf-usa.org/CurriculumMaterials/BRC_Handbook_Vs71_noprint.pdf, 2007.
[48] H.-C. Kim, D. Kim, Z. Ghahramani, and S.Y. Bang, "Appearance-Based Gender Classification with Gaussian Processes," *Pattern Recognition Letters,* vol. 27, pp. 618-626, 2006.
[49] A.J. O'Toole, T. Vetter, N.F. Troje, and H.H. Bulthoff, "Sex Classification Is Better with Three-Dimensional Structure than with Image Intensity Information," *Perception,* vol. 26, pp. 75-84, 1997.
[50] C. BenAbdelkader and P. Griffin, "A Local Region-Based Approach to Gender Classification from Face Images," *Proc. IEEE Conf. Computer Vision and Pattern Recognition,* 2005.
[51] G. Shakhnarovich, P.A. Viola, and B. Moghaddam, "A Unified Learning Framework for Real Time Face Detection and Classification," *Proc. Int'l Conf. Automatic Face and Gesture Recognition,* 2002.
[52] Y. Lamdan, J.T. Schwartz, and H.J. Wolfson, "On Recognition of 3D Objects from 2D Images," *Proc. IEEE Int'l Conf. Robotics and Automation,* pp. 1407-1413, 1988.
[53] R.O. Duda, P.E. Hart, and D.G. Stork, *Pattern Classification,* second ed. Wiley, 2001.
[54] D. Comaniciu and P. Meer, "Mean Shift: A Robust Approach toward Feature Space Analysis," *IEEE Trans. Pattern Analysis and Machine Intelligence,* vol. 24, no. 5, pp. 603-619, May 2002.
[55] M. Toews and T. Arbel, "Detecting and Localizing 3D Object Classes Using Viewpoint Invariant Reference Frames," *Proc. Int'l Conf. Computer Vision (ICCV Workshop) 3D Representation for Recognition,* 2007.
[56] M.I. Jordan, *An Introduction to Probabilistic Graphical Models,* in preparation.
[57] E. Jaynes, "Prior Probabilities," *IEEE Trans. Systems, Science, and Cybernetics,* vol. 4, no. 3, pp. 227-241, Sept. 1968.
[58] C. Schmid, R. Mohr, and C. Bauckhage, "Evaluation of Interest Point Detectors," *Int'l J. Computer Vision,* vol. 37, no. 2, pp. 151-172, June 2000.
[59] K. Mikolajczyk and C. Schmid, "A Performance Evaluation of Local Descriptors," *IEEE Trans. Pattern Analysis and Machine Intelligence,* vol. 27, no. 10, pp. 1615-1630, Oct. 2005.
[60] C. Dance, J. Willamowski, L. Fan, C. Bray, and G. Csurka, "Visual Categorization with Bags of Keypoints," *Proc. European Conf. Computer Vision (ECCV Workshop) Statistical Learning in Computer Vision,* 2004.
[61] C. Cortes and V. Vapnik, "Support-Vector Networks," *Machine Learning,* vol. 20, no. 3, pp. 273-297, 1995.
[62] C.-C. Chang and C.-J. Lin, "LIBSVM—A Library for Support Vector Machines," www.csie.ntu.edu.tw/cjlin/libsvm, 2007.
[63] M. Toews and T. Arbel, "A Statistical Parts-Based Appearance Model of Anatomical Variability," *IEEE Trans. Medical Imaging,* special issue on computational neuroanatomy, vol. 26, no. 4, pp. 497-508, Apr. 2007.
[64] A. Opelt, A. Pinz, M. Fussenegger, and P. Auer, "Generic Object Recognition with Boosting," *IEEE Trans. Pattern Analysis and Machine Intelligence,* vol. 28, no. 3, pp. 416-431, Mar. 2006.

Matthew Toews received the BEng degree in computer engineering from the University of British Columbia in 2000, and the MEng and PhD degrees in electrical engineering from McGill University in 2003 and 2008, respectively. He is currently a postdoctoral research fellow at the Brigham and Women's Hospital in the Department of Radiology of the Harvard Medical School, where he is supported by a National Science and Engineering Research Council (NSERC) postdoctoral fellowship. His research involves applying statistical modeling, machine learning, probability theory, and information theory in the contexts of computer vision and medical image analysis. His interests include computational anatomy, image registration, image similarity measurement, invariant feature methods, and object detection, classification, and recognition. He is a member of the IEEE.

Tal Arbel received the BEng, MEng, and PhD degrees in electrical engineering from McGill University, where she received the D.W. Ambridge Prize for the most outstanding PhD thesis in engineering and physical sciences, in 1992, 1995, and 2000, respectively. She was a postdoctoral fellow in the Brain Imaging Centre at the Montreal Neurological Institute (MNI) from 2000 to 2001, where her research in image-guided neurosurgery was supported by a National Science and Engineering Research Council (NSERC) postdoctoral fellowship. In 2001, she joined the Department of Electrical and Computer Engineering at McGill University as an assistant professor and became a member of the Centre for Intelligent Machines (CIM). She was subsequently awarded an NSERC University Faculty Award. Her interests lie in the development of probabilistic techniques in computer vision and in medical imaging, particularly in object recognition, active vision, image registration, and object detection, localization, and classification. In medical imaging, she has been particularly interested in the domains of neurology and image-guided neurosurgery, where she has developed techniques in image registration, medical image reconstruction (e.g., ultrasound), and tissue/pathology modeling and classification. In 2007, she served as the general chair of AI/GI/CRV/IS—the Joint Canadian Conference on Artificial Intelligence, the Graphics Interface Conference, the Canadian Conference on Computer and Robot Vision, and the Annual Canadian Conference on Intelligent Systems. She is a member of the IEEE and the IEEE Computer Society.