Detecting, Localizing and Classifying Visual Traits from Arbitrary Viewpoints Using Probabilistic Local Feature Modeling

Matthew Toews and Tal Arbel

Centre for Intelligent Machines
McGill University
Montreal, Canada
{mtoews,arbel}@cim.mcgill.ca

Abstract. We present the first framework for detecting, localizing and classifying visual traits of object classes, e.g.gender or age of human faces, from arbitrary viewpoints. We embed all three tasks in a viewpoint-invariant model derived from local scale-invariant features (e.g. SIFT), where features are probabilistically quantified in terms of their occurrence, appearance, geometry and relationship to 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 advantage of our framework is that it can be applied and evaluated in realistic scenarios, unlike other trait classification techniques that assume data that is single-viewpoint, pre-aligned and cropped from background distraction. Experimentation on the standard color FERET database shows our approach can automatically identify the visual cues in face images linked to the trait of gender. Combined detection, localization and gender classification error rates are a) 15% over a 180-degree range of face viewpoint and b) 13% in frontal faces, lower than other reported results.

1 Introduction

Practical visual processing applications must be able to robustly detect instances of object classes of interest in arbitrary, cluttered images, and make inferences regarding their visual traits. For example, consider an intelligent vision system that must identify all males in a crowded scene, as illustrated in Figure 1. Image features arising from human face instances must first be detected and localized in the midst of unrelated clutter and viewpoint change, after which they can be used to determine traits such as gender for each person detected. Although the tasks of detection, localization and classification are all inextricably linked in such realistic visual processing scenarios, they are typically treated in isolation in the current vision literature. For example, approaches to classifying facial traits such as gender typically assume frontal face data which has been precisely pre-aligned and cropped from distracting background clutter prior to classification [2,25,14,17,9]. As a result, it remains questionable whether such approaches can be applied in conjunction with automatic detection strategies in arbitrary, cluttered scenes where automatic face localization is non-trivial. Likewise, it is unclear whether

S.K. Zhou et al. (Eds.): AMFG 2007, LNCS 4778, pp. 154–167, 2007.

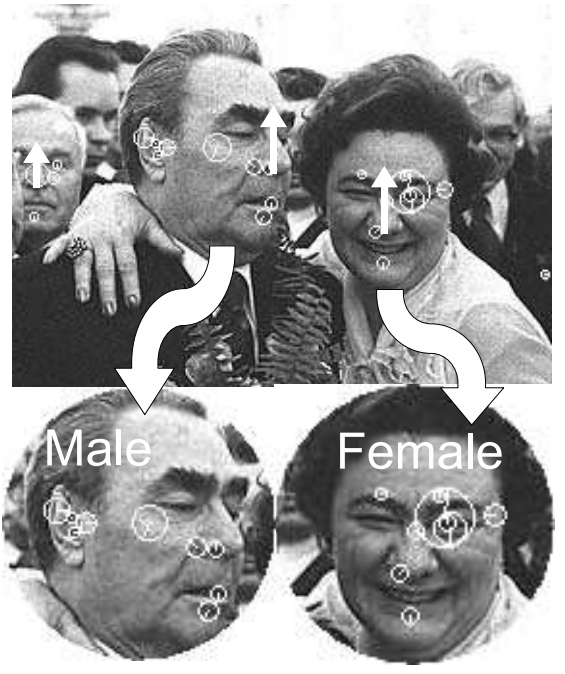


Fig. 1. Illustrating the output of our general framework for detection, localization and trait classification from arbitrary viewpoints. All three tasks are embedded in a viewpoint-invariant model derived from scale-invariant image features. Here, face instances (white arrows above) are first detected and localized from scale-invariant features (white circles) extracted in a cluttered scene. Features associated with each face instance are then used in a Bayesian classifier to determine face gender (lower insets). The image shown is from the CMU face database [5], and the probabilistic framework used is learned from 500 FERET [1] face images taken at arbitrary viewpoints.

recent general object class detection strategies [8,7,19,6,4,21] can be extended in order to learn and classify abstract visual traits such as gender from arbitrary viewpoints.

Our contribution in this paper is a general, integrated framework for detecting, localizing and classifying visual traits of object classes from arbitrary viewpoints. Our approach is the first to propose learning visual traits from arbitrary viewpoints, and the first to embed all three tasks in a general appearance model based on 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 set of model features related to the object class of interest, after which the same features are used to train a Bayesian classifier for visual traits. Classifier training involves estimating the likelihood ratio of feature occurrence given trait presence vs. absence, the underlying premise being that informative features are more likely than not to co-occur with the trait of interest. The resulting framework can be used to detect and localize and classify the traits of object class instances in the presence of viewpoint change, geometrical deformations such as translation, orientation

and scale changes, linear illumination changes, partial occlusion and multi-model intra-class variation (e.g. faces with/without sunglasses).

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 Section 3 we describe our approach to trait learning and classification based on probabilistic modeling of scale-invariant image features. Although our approach generalizes to a variety of object classes and visual traits, we experiment on the class of faces and the trait of gender in Section 4 using the standard color FERET database. We provide a quantitative performance evaluation for combined detection, localization and gender classification of faces images in both arbitrary and frontal viewpoint contexts, and show how our approach can identify visual cues of gender in face images over a range of viewpoint. A discussion follows in Section 5.

2 Related Work

2.1 Object Class Detection

The general detection task requires identifying and localizing instances of an object class, e.g. cars or faces, in images. General object class detection requires effectively dealing with a wide range of appearance variation due to viewpoint change, geometrical deformations such as translation, orientation and scale changes, illumination changes, partial pattern occlusion and multi-modal intra-class variation (i.e. faces with/without sunglasses). Such variation can only be realistically overcome by learning a model from a set of natural training images. Early approaches advocated learning models of global features, i.e. eigenfaces [22], but proved to be inefficient for detection over geometrical deformations and poorly suited for coping with local appearance variation and occlusion. To overcome these difficulties, researchers have increasingly turned to local image feature representations. Scale-invariant features [15,3,16,13] for instance can be robustly and efficiently extracted from scale-space pyramids in the presence of translation, orientation and scale geometrical deformations and illumination changes. As features are local, they can be used to determine correspondence between different images in the presence of partial occlusion. Geometrical information from the extraction process including feature location, orientation and scale can be used to generate independent hypotheses as to the geometrical transform relating different images, without requiring an expensive explicit search over transform parameters.

While scale-invariant feature correspondences cannot generally be established directly between different instances of the same object class due to intra-class variability, research has shown that learned probabilistic models of features can be used to reliably detect object class instances in arbitrary, cluttered images [8,7,19,6,4,21]. Such models describe the appearance of an object class in terms of a set of local features, including their appearances, occurrences and their geometries (i.e. image location, orientation and scale). Models generally vary in terms of the assumptions made regarding inter-feature geometrical dependencies, e.g. geometry independent models [7,19], naive Bayes dependencies [21,8], Markov dependencies [4], fully-dependent [24], and intermediate approaches [6]. Although geometrical dependence assumptions vary, most

models make the assumption of conditional independence of individual feature appearances/occurrences given feature geometry and object class.

Most approaches to invariant feature modeling are based on stable 2D feature configurations in the image plane, and are thus single-viewpoint in nature [8]. Multi-view [20] and viewpoint-invariant [21] representations have emerged to address object class detection and localization from arbitrary viewpoints. Modeling 3D object class appearance over viewpoint change is considerably more challenging than from single viewpoints, as correspondences must be established between different views in addition to different object class instances, and learning techniques typically employ a degree of external supervision. The multi-view modeling approach [20] requires a viewsphere sampled at regular angular intervals, for each of set of different object class instances. Such an approach is not well suited to learning from natural images taken from arbitrary viewpoints around arbitrary 3D object class instances, however. The viewpoint-invariant approach [21] relates features in different images via an object class invariant (OCI), a geometrical reference frame that is uniquely defined for each object class instance and invariant to projective image transform arising from viewpoint change. As the variable of viewpoint is effectively marginalized from the formulation, a viewpoint-invariant model can be learned from natural imagery taken from arbitrary viewpoints from labeled images.

2.2 Visual Trait Classification: Gender from Faces

Visual traits are abstract qualities of an object class identifiable from images, such as the make or model of cars, the age or gender of faces, etc. They represent a mechanism by which members of the same object class can be described or subdivided. Due to the ubiquitous nature of face image analysis, one of the most common visual trait classification tasks is that of determining gender from face images, and the wide range of published approaches highlights the state-of-the-art in general trait classification. Trait learning has been tackled from spatially global feature representations such as templates[14,9], principle components [17] or independent components [10]. Other more recent approaches used pixels as features [2] or Haar wavelets [25,18]. Machine learning techniques such as neural networks [9], support vector machines (SVMs) [17] and boosted classifiers [2] have been brought to bear. In the interest of comparison, most approaches train and test on the standard FERET face database [1] containing accurate labels for visual traits such as gender, age and ethnicity.

To date, all published approaches to trait classification are based exclusively on single viewpoints, i.e. frontal faces [2,25,10,17,9]. With the exception of [18], most approaches assume that, prior to classification, faces 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 [25,17], and pre-defined facial masks are subsequently applied to remove background clutter. As a result, classification error rates of 4% to 10% represent artificially low, ideal-case results, and offer little insight as to classification performance in a general vision system where object class localization is non-trivial. 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 (i.e. 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 [25,17]. 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.

2.3 Combined Detection, Localization and Trait Classification

To date only a single approach has proposed combined detection, localization and classification within an integrated framework suitable for general object classes [18], using boosted classifiers of Haar wavelet features [23] for all tasks. The approach is single-viewpoint (frontal faces) and not invariant to orientation, and the reported error rate of 21% reflects the increased difficulty of the combined task. This result is based a proprietary database, however, where faces with ambiguous gender are manually removed, as are faces whose in-plane orientation is greater than 30 degrees, and as such a direct comparison cannot be made. The general scale-invariant feature-based approach to modeling object classes offers an attractive alternative for combined detection, localization and classification, as it can provide invariance to viewpoint change, in addition to in addition to translation, orientation and scale changes. To date, the general scale-invariant feature-based modeling approach has not been investigated for visual trait classification, and classifying visual traits such as gender from faces from arbitrary viewpoints has not been addressed.

3 Classifying Visual Traits from Local Features

In realistic scenarios, visual trait classification is inseparable from detection and localization: features must first be detected and localized before they can be classified. We propose embedding all three tasks within a general appearance model derived from local scale-invariant features, which can be used to detect, localize and classify traits of object classes in natural imagery captured from arbitrary viewpoints. A model describing object class appearance is first learned, after which a Bayesian trait classifier is then trained from features in the model.

3.1 Viewpoint-Invariant Appearance Modeling

To effectively capture the subtleties of visual traits, we require a model that can be 1) effectively learned from arbitrary viewpoints 2) used to detect and localize individual object class instances in arbitrary viewpoints and 3) provide a rich, multi-modal description of object class appearance on which trait classification can be based. We thus avoid geometry-free models which are generally less suitable for localizing object class instances [7,19] and models consisting of relatively few features (e.g.10) [8] which may not provide a sufficiently rich image description for visual traits. While our trait classification approach is generally applicable to single-viewpoint models consisting of many features (e.g.100+) [4], effectively learning and classifying visual traits requires

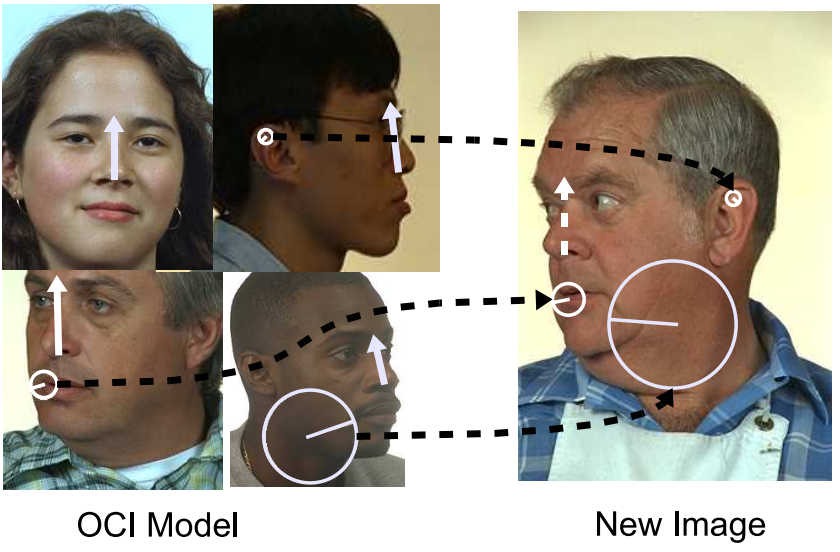


Fig. 2. Illustrating the viewpoint-invariant OCI model relating scale-invariant features (white circles) to an OCI (solid white arrows). The OCI, defined here a line segment from the base of the nose to the forehead, represents a viewpoint-invariant mechanism for grouping scale-invariant image features in images taken from arbitrary viewpoints. A probabilistic model (left) is learned from manually labeled OCIs in training images taken at arbitrary viewpoints. Model instances can then be robustly detected and localized in a new image (right) taken 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.

addressing the issue of viewpoint change. To do this, we adopt the viewpoint-invariant OCI model [21]. The model relates scale-invariant features to an OCI, an abstract 3D geometrical structure that is uniquely defined with respect to each 3D object class instance and invariant to projective transform arising from viewpoint change, as illustrated in Figure 2. The probabilistic model formulation adopts the assumption of conditional independence of feature geometries and appearances/occurrences, and the naive Bayes assumption of conditional independence of individual features i, given the OCI o. Under these assumptions, the posterior probability of o given feature geometry $G : \{g_i\}$, appearance $A : \{a_i\}$ and occurrence $F : \{f_i\}$ data can be expressed as:

$$\begin{aligned} p(o|G, A, F) &\propto p(o)p(G|o)p(A, F|o), \\ &\propto p(o)\prod_i p(g_i|o)p(a_i|f_i)p(f_i|o), \end{aligned} \tag{1}$$

where distributions $p(g_i|o)$, $p(a_i|f_i)$ and $p(f_i|o)$ over individual feature geometries, appearances and occurrences are learned from a set of training data containing features and labeled OCIs. Novel object class instances can be detected and localized in new

images by maximizing the posterior probability in equation (1) with respect to o based on detected model features.

3.2 Visual Trait Classification

Once a viewpoint invariant model has been learned for a given object class, we seek to identify model features indicative of visual traits using the co-occurrence statistics of individual features with the trait of interest. To do this, we consider the random event $f_i = 1$ signifying the occurrence of model feature i, and we expand the random event of object class occurrence $o = 1$ into a discrete random variable $c : \{c_1, \ldots, c_K\}$ over K trait values of interest, e.g.$gender : \{female, male\}$. A Bayesian classifier $\gamma(c)$ can then be used to express the most probable trait classification given a set of model feature occurrences $\{f_i\}$:

$$\gamma(c_j) = \frac{p(c_j|\{f_i\})}{p(\bar{c_j}|\{f_i\})} = \frac{p(c_j)}{p(\bar{c_j})} \prod_i \frac{p(f_i|c_j)}{p(f_i|\bar{c_j})}. \tag{2}$$

where $\frac{p(c_j)}{p(\bar{c_j})}$ is the prior ratio of trait value presence c_j vs. absence $\bar{c_j}$ (e.g. male vs. not male or female), and $\frac{p(f_i|c_j)}{p(f_i|\bar{c_j})}$ expresses the likelihood ratio of trait value presence c_j vs. absence $\bar{c_j}$ coinciding with feature observation f_i. Features that are important to classification or highly informative with regard to trait value c_j will have high likelihood ratios. The focus of our approach is to use these likelihood ratios to quantify the association of model features with visual traits, as illustrated in Figure 3.

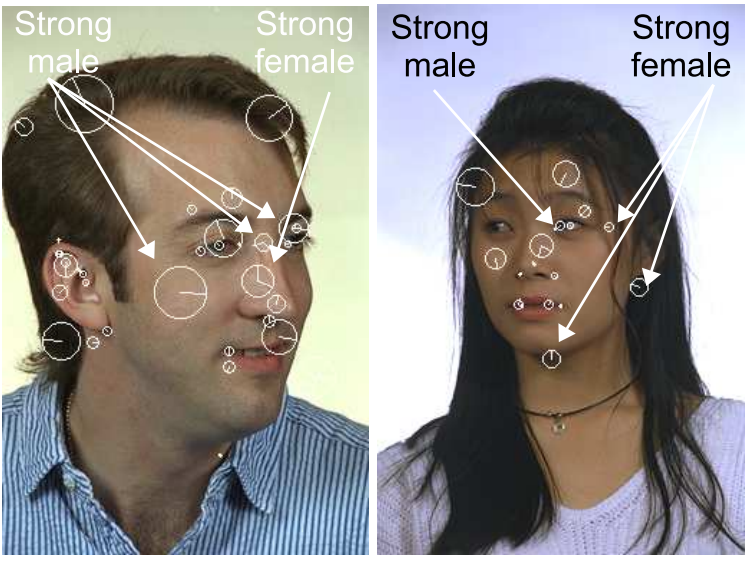
In order to estimate the likelihood parameters, we use a supervised learning process, based on observed model feature occurrences f_i and trait labels c_j for each training image. Discrete class-conditional likelihoods $p(f_i|c_j)$ can be represented as binomial distributions, parameterized by event counts [12]. 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:

$$p(f_i|c_j) = \frac{p(f_i, c_j)}{p(c_j)}. \tag{3}$$

The most straightforward manner of estimating $p(f_i, c_j)$ is via ML (maximum likelihood) 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 MAP (maximum a posteriori) estimation can be used to cope with data sparsity, and involves regularizing estimates using a Dirichlet hyperparameter distribution [12]. In practice, Dirichlet regularization involves pre-populating 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 [11]. Although both ML and MAP estimates converge as the number of data samples increases, MAP estimation using a uniform prior will tend towards conservative parameter estimates while the number of data samples is low. The final estimator we use becomes:

$$p(f_i|c_j) \propto \frac{k_{i,j}}{p(c_j)} + d_{i,j}, \tag{4}$$

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 equation (4) can be obtained by normalizing over values of f_i, but is not required for likelihood ratios.



(a) (b)

Fig. 3. Illustrating classification of the visual trait of gender from local features (white circles). A given face instance consists of a set of local features, a subset of which are reflective of either gender, 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 gender in training images is greater than 2:1. Of the 63 model features detected in image (a), 15 are strongly male and 1 is strongly female, suggesting a male face. Of the 31 features detected in image (b), 7 are strongly female and 1 is strongly male, suggesting a female face. Many features, although very common in the class of face images, are uninformative regarding gender.

4 Experimentation

For the purpose of experimentation, we consider the combined task of detection, localization and gender classification of faces from arbitrary viewpoints. To compare with the results in the literature, we also provide results for a model trained from frontal faces only. Experimentation is based on the standard, publicly available color FERET face image database [1] for both training and testing. The FERET database consists

of images of 994 unique subjects of various ethnicity, age, gender, taken from various viewpoints, illumination conditions, with/without glasses, etc. We process images at a resolution of 256x384 pixels and no subjects are duplicated in either testing or training data, in order to evaluate the generality of our approach.

Learning proceeds as follows: an initial local feature-based model is trained on randomly-selected subsets of face images using the supervised OCI technique, and the remaining images are used for testing. Model learning requires approximate labeling of an OCI in the form of a line segment from the base of the nose to the forehead as in Section 3.1, and automatically extracting scale-invariant features in each training images. Although a variety of different scale-invariant features can be used, we use the SIFT (scale-invariant feature transform) technique [15] for feature detection and appearance description based on robust implementation made public by the author. Once the model has been learned, model feature occurrences identified in the training set along with FERET gender labels are used to estimate likelihood ratios of the Bayesian trait classifier as described in Section 3. In estimating likelihood ratios via equation (4), we used a Dirichlet regularization parameter of $d_{i,j} = 2$ which maximizes training set classification performance.

Once the framework has been learned, combined detection, localization and classification proceed on the remainder of FERET faces not used in training. Scale-invariant features are first extracted in all testing images, after which detection and localization are performed by determining the most probable OCI instance in each of the testing images based on extracted image features. Model features contributing the OCI instance are then used to determine gender using the Bayesian classifier in equation (2), using a prior trait ratio of $\frac{p(c_j)}{p(\bar{c_j})} = 1$. Note that the FERET database does not necessarily represent the most challenging test for model detection, as most faces are clearly visible, but it does allow evaluating whether or not facial features can be automatically localized with sufficient accuracy for subsequent trait classification. Qualitative experimentation was performed on cluttered imagery from the CMU profile database [5], as illustrated in Figure 1), demonstrating the viability of the system in difficult detection/localization contexts. Performance is generally better for higher resolution faces, where the number of SIFT features extracted is sufficient for reliable detection and classification. Classification appears correct in most cases, although ground truth gender labels are unavailable and difficult to determine in many cases.

4.1 Locating, Detecting and Classifying from Arbitrary Viewpoints

In order to investigate the trait of gender from arbitrary viewpoints, a database of all 994 unique FERET subjects was selected, where each subject image is chosen at random from a 180 degree viewpoint range (i.e. from left to right profile images). Figure 4 illustrates the viewpoint distribution in the dataset. We trained on two randomly selected subsets of 331 and 497 images (1/3 and 1/2 of the data), and performed combined detection, localization and gender classification on the remaining images.

Table 1 summarizes the results obtained, where our approach achieves an error rate of 15% based on 497 training images. Misclassification rates were 12% and 17% for males and females, respectively, examples can be seen in Figure 5. The detection and localization error rate prior to classification was 3.6%, where the discrepancy between

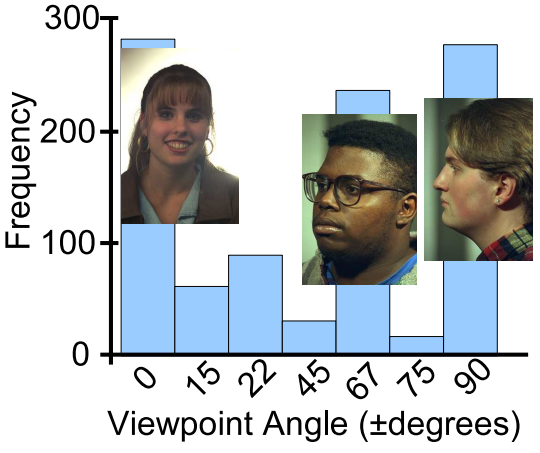


Fig. 4. A histogram illustrating the viewpoint distribution for the 994 unique FERET subject images used in experimentation

Table 1. Error rates for localization, and combined detection, localization and classification over viewpoint. Our Bayesian classifier obtains an error rate of 15% when trained on 497 of 994 images. Localization error rates are based on discrepancy threshold between localized and manual labeled OCIs in testing images.

Database	Loc. Error	Combined Error
FERET (497 training)	3.6%	15%
FERET (331 training)	4.5%	19%

the localized and labeled OCIs was greater than a threshold in scale, orientation and location of $log(1.5)$, 20 degrees and OCI scale/2 pixels, respectively. It is possible that increasing the training data size by several hundred more images would further reduce the error rate by several percentage points.

4.2 Identifying Visual Cues of Gender

As humans, we are all capable of determining visual traits such as face gender of a face image with reasonable certainty. What is more difficult is to identify the visual cues that are operative in making the determination - 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 gender can be visualized as in Figure 6. In a viewpoint invariant model, ear features are more indicative of males, as they are less visible due to generally longer female hair. Several features around the mouth are indicative of males, indicative of

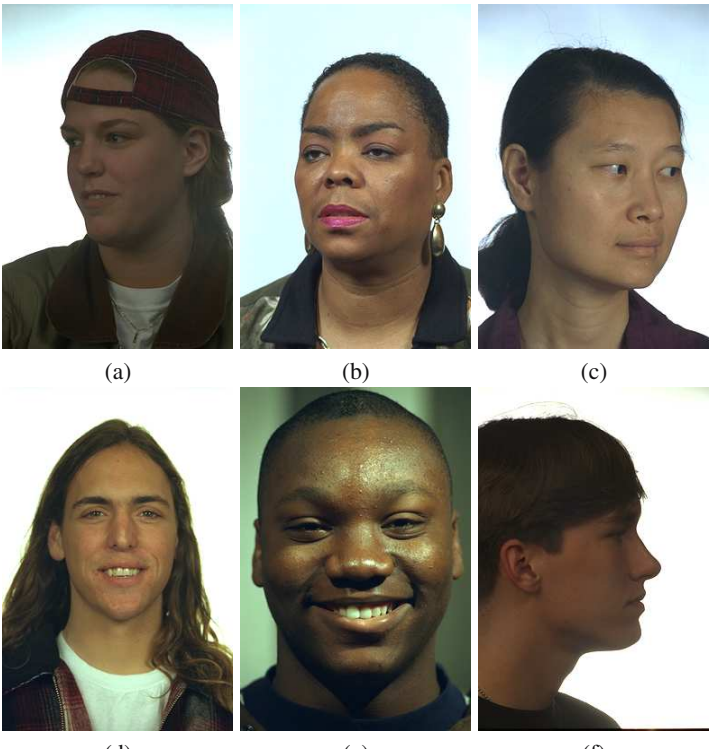


Fig. 5. Illustrating several misclassification examples. Images (a)-(c) are misclassified as male, while (d)-(f) are misclassified as female. Misclassification can occur due to faces containing a disproportionately high number of features indicative of the opposite sex, e.g. (a)-(e), or a lack of gender-informative features e.g. (f).

beards or facial stubble. Females are distinguished by features arising from hairlines, eyes (possible from makeup) and lips. In contrast, certain model features arising from nostrils or cheeks, although very common in the class of face images, were generally less informative regarding gender. Note that although the male:female ratio in training data was close to 1:1, approximately twice as many gender-related features were identified for males as for females, suggesting a greater number of visual cues characteristic of the male gender.

4.3 Locating, Detecting and Classifying from Frontal Views

In order to compare our general approach results in the literature, we also trained and tested on a restricted set of frontal faces. We used the 925 standard FERET frontal images labeled "*_fa.*", models were trained from randomly selected subsets of 100 and 200 images. Table 2 for the combined task of detection, localization and classification, error rates for classification only (i.e. faces pre-aligned and cropped prior to classifica-

Table 2. Published error rates for combined detection, localization and classification for frontal faces. Our Bayesian classifier trained on 200 FERET faces in (a) achieves the lowest error rate of 13%, in comparison with (b) the boosted Haar wavelet classifier [18] of 21%. Results for (b) are based on a proprietary database, however, so a precise comparison cannot be made. Results for classification only (c), i.e. with faces pre-aligned and cropped, represent an ideal-case baseline and are included for completeness.

Task	Method	Features	Database	Error Rate
Combined detection, localization and classification	(a) Bayesian classifier	Scale-invariants	FERET (200 training)	13%
			FERET (100 training)	16%
	(b) Adaboost [18]	Haar wavelets	Proprietary (≈3000 training)	21%
Classification only	(c) Various [2,25,10,17,9]	Various	FERET	4%-10%

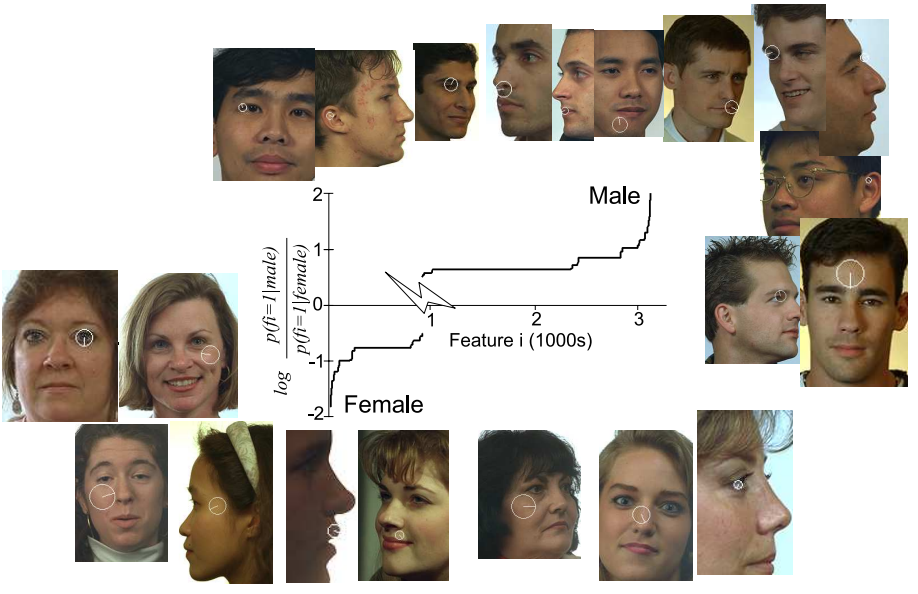


Fig. 6. Illustrating visual cues indicative of face gender, in the form of scale-invariant features. Features are sorted in increasing order of their log likelihood ratio $log(\frac{p(f_i=1|male)}{p(f_i=1|female)})$. Of approx. 15,000 features in a viewpoint invariant face model learned from 497 randomly selected FERET images, approx. 3000 features bear information regarding gender (i.e. $|log(\frac{p(f_i=1|male)}{p(f_i=1|female)})| > 0.5$). Features lying to the left of the graph occur more frequently in female subjects and those to the right more frequently in male subjects. Face images shown illustrate instances of gender-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 was 28:22, approximately twice as many gender-reflective features are associated with males.

tion) represent an ideal-case baseline and are included for completeness. For the more difficult combined task, our Bayesian gender classifier achieves an error rate of 13% for training based on 200 randomly selected FERET subjects. The misclassification rates are 9.7% and 15% for males and females, respectively. Note that training on only 100 subjects results a marginally higher error rate of 16%, suggesting that the majority of the information regarding face gender is captured from on the order of several hundred subjects. Note that significantly fewer training images are required to obtain similar error rates to modeling viewpoint.

5 Discussion

In this paper, we present the first approach addressing learning and classification of visual traits of object classes from arbitrary viewpoints. As a realistic scenario requires first detecting and localizing object class instances prior to trait classification, we embed all three tasks within a single viewpoint-invariant model of general object class appearance that can be used for combined detection, localization and classification from arbitrary viewpoints. Our approach involves first learning a model of object class appearance, then training a Bayesian classifier for visual traits from model features. Classifier training involves estimating the likelihood ratio of positive feature occurrence given trait presence vs. absence, where features associated with significantly non-zero likelihood ratios indicate visual cues reflective of the trait of interest. We provide the first experimental results on a standard, publicly available database (FERET) for the combined tasks of detection, localization and gender classification of faces. We obtain an error rate of 15% for the combined task over a 180 degree range of face viewpoint, and an error rate of 13% for frontal faces.

Various future avenues exist for learning visual traits from general appearance models based on local scale-invariant features. Computational complexity of detection, localization and classification is low and the combined system should be implementable in real or near-real time. Visual traits of faces can be used as a soft biometric in interactive image-based applications, surveillance or recognition, and the framework could possibly extend to classifying traits such as age or emotion. We experimented with learning the trait of age, by dividing faces into less than/greater than 25 years of age, splitting the data set approximately evenly. A somewhat high classification error rate of 23% was obtained framework trained from the framework trained on 200 frontal faces, indicating that age classification is a more difficult problem than gender, particularly when faces are split at the 25-year old mark. Useful traits can potentially be learned from of different object classes, for example the make or model of cars and motorbikes. Whether different traits such as age and gender are best modeled independently or jointly is an open research question, the Bayesian classifier we present could be used for either although joint modeling becomes computationally complex for large numbers of different traits. Continuous-valued traits such as age could potentially be modeled using continuous-valued likelihoods. A variety of different scale-invariant feature types other than SIFT could be incorporated to potentially improve classification performance by highlighting different image characteristics.

References

1. Feret face database. http://www.itl.nist.gov/iad/humanid/colorferet
2. Baluja, S., Rowley, H.A.: Boosting sex identification performance. IJCV 71(1), 111–119 (2007)
3. Carneiro, G., Jepson, A.: Multi-scale phase-based local features. In: CVPR, vol. 1, pp. 736–743 (2003)
4. Carneiro, G., Lowe, D.G.: Sparse flexible models of local features. In: Leonardis, A., Bischof, H., Pinz, A. (eds.) ECCV 2006. LNCS, vol. 3953, pp. 29–43. Springer, Heidelberg (2006)
5. CMU Face Group. Frontal and profile face databases, http://vasc.ri.cmu.edu/idb/html/face/
6. Crandal, D., Felzenswalb, P., Huttenlocher, D.: Spatial priors for part-based recognition using statistical models. In: CVPR, vol. 1, pp. 10–17 (2005)
7. Dorko, G., Schmid, C.: Selection of scale-invariant parts for object class recognition. In: ICCV, pp. 634–640 (2003)
8. Fergus, R., Perona, P., Zisserman, A.: Weakly supervised scale-invariant learning of models for visual recognition. IJCV 71(3), 273–303 (2006)
9. Gutta, S., Wechsler, H., Phillips, P.: Gender and ethnic classification of human faces using hybrid classifiers. In: FGR, pp. 194–199 (1998)
10. Jain, A., Huang, J., Fang, S.: Gender identification using frontal facial images. In: ICME (2005)
11. Jaynes, E.: Prior probabilities. IEEE Transactions of systems, science, and cybernetics 4(3), 227–241 (1968)
12. Jordan, M.I.: An Introduction to Probabilistic Graphical Models (2003) (in preparation)
13. Kadir, T., Brady, M.: Saliency, scale and image description. IJCV 45(2), 83–105 (2001)
14. Kim, H.-C., Kim, D., Ghahramani, Z., Bang, S.Y.: Appearance-based gender classification with gaussian processes. PRL 27, 618–626 (2006)
15. Lowe, D.G.: Distinctive image features from scale-invariant keypoints. IJCV 60(2), 91–110 (2004)
16. Mikolajczyk, K., Schmid, C.: Indexing based on scale invariant interest points. In: ICCV, pp. I: 525–531 (2001)
17. Moghaddam, B., Yang, M.: Learning gender with support faces. PAMI 24(5), 707–711 (2002)
18. Shakhnarovich, G., Viola, P.A., Moghaddam, B.: A unified learning framework for real time face detection and classification. In: FGR (2002)
19. Sivic, J., Russell, B.C., Efros, A.A., Zisserman, A., Freeman, W.T.: Discovering objects and their localization in images. In: ICCV, pp. 370–377 (2005)
20. Thomas, A., Ferrari, V., Leibe, B., Tuytelaars, T., Schiele, B., Van Gool, L.: Towards multi-view object class detection. In: CVPR (2006)
21. Toews, M., Arbel, T.: Detection over viewpoint via the object class invariant. In: ICPR, pp. 765–768 (2006)
22. Turk, M., Pentland, A.P.: Eigenfaces for recognition. CogNeuro 3(1), 71–96 (1991)
23. Viola, P., Jones, M.: Rapid object detection using a boosted cascade of simple features. In: CVPR, vol. 1, pp. 511–518 (2001)
24. Weber, M., Welling, M., Perona, P.: Unsupervised learning of models for recognition. In: Vernon, D. (ed.) ECCV 2000. LNCS, vol. 1842, pp. 18–32. Springer, Heidelberg (2000)
25. Yang, Z., Li, M., Ai, H.: An experimental study on automatic face gender classification. In: ICPR, pp. 1099–1102 (2006)