Jointly Estimating Demographics and Height with a Calibrated Camera

Andrew C. Gallagher
Eastman Kodak Company
andrew.gallagher@kodak.com

Andrew C. Blose
Eastman Kodak Company
andrew.blose@kodak.com

Tsuhan Chen
Cornell University
tsuhan@ece.cornell.edu

Abstract

One important problem in computer vision is to provide a demographic description a person from an image. In practice, many of the state-of-the-art methods use only an analysis of the face to estimate the age and gender of a person of interest. We present a model that combines two problems, height estimation and demographic classification, which allows each to serve as context for the other. Our idea is to use a calibrated camera for measuring the height of people in the scene. Height is measured by jointly inferring across anthropometric dimensions, age, and gender using publicly available statistics. The height estimate provides context for recognizing the age and gender of the subject, and likewise age and gender conditions the distribution of the anthropometric features for estimating height.

The performance of our method is explored on a new database of 127 people captured with a calibrated camera with recorded height, age, and gender. We show that estimating height leads to improvements in age and gender classification, and vice versa. To the best of our knowledge, our model produces the most accurate automatic height estimates reported, with the error having a standard deviation of 26.7 mm.

1. Introduction

The goal of this paper is to describe a person's height and demographics from an image. In computer vision research, algorithms exist to identify the age and the gender of people. Broadly speaking, these algorithms build statistical models for the image appearance of a person for different demographic categories, and these models are employed to categorize the image of a previously unseen face. With few exceptions, demographic recognition is performed solely based on facial appearance. In practice, however, facial appearance does not provide enough information to solve this problem with the desired level of accuracy.

Similarly, several researchers have investigated the problem of estimating the height of a standing or walking human. In some cases, the problem has been addressed solely as a metrology problem, using similar techniques than can be applied for measuring any other vertical object.

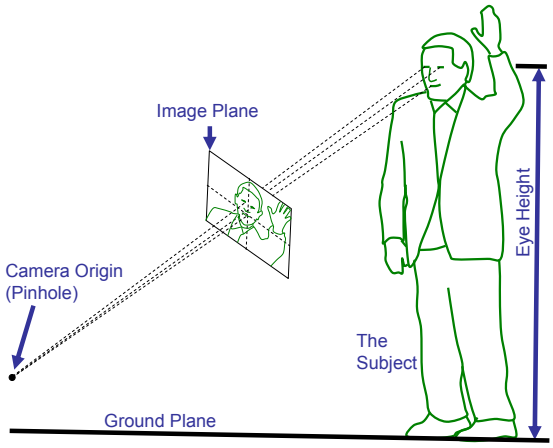


Figure 1. Measuring human height with a calibrated camera. Calibration provides the relationship between the image and world coordinate systems. Facial key points fall on rays from the camera center and corresponding images of the points. Anthropometric data, conditioned by the estimated age and gender, provides distributions on distances between key points to infer distance from subject to camera, height, age, and gender.

The goal of our paper is to unite these two sub-problems (height measurement and demographic estimation) into a common framework employing a probabilistic model to allow evidence gathered for each sub-problem to reduce the uncertainty about the other. Our approach is to combine facial appearance with height estimation to improve our understanding of images of people. To this end, we exploit the large volume of anthropometric measurements gathered by the medical and health communities.

1.1. Related Work

A large amount of research is directed at understanding images of humans, addressing issues such as recognizing an individual, recognizing age and gender from facial appearance, and determining the structure of the human body. Most age and gender classification algorithms construct feature vectors solely from the face region [2, 11, 12, 14]. In

fact, the vast majority of classification work related to images of people treats each face as an independent problem and relies solely on information gleaned from images from which classifiers are constructed. However, there are some notable exceptions where information external to the image is used as context for classification. In [4], names from news captions are associated with faces from images or video in a mutually exclusive manner. Similar constraints are employed in research devoted to solving the face recognition problem for consumer image collections. In [10], the popularity trends of first names provide context in conjunction with facial appearance to infer age and gender.

Regarding height estimation, several researchers either estimate height, or use broad height distributions with pedestrian detection to understand scenes. The position of people in an image provides clues about the geometry of the scene. As shown in [20], camera calibration can be achieved from a video of a walking human, under some reasonable assumptions (that the person walks on the ground plane and head and feet are visible). In [18], the problem is reversed, and the height of a person with visible feet and head is estimated from a calibrated camera. Criminisi *et al*. [7], Hoiem *et al*. [16], and Lalonde *et al*. [19] describe the measurement of various objects (including people) rooted on the ground plane. However, all of these papers require that the intersection of the object (i.e. the feet) and the floor be visible. Our method relies on anthropometric face measurements and requires instead that the face be visible.

Our work uses information from anthropology and medicine as context for demographic inference in computer vision. In anthropology, the relationships between various body measurements has been exploited to estimate an individual's height from a single recovered bone [8]. Perhaps the closest work on human height measurement from images is BenAbdelkader and Yacoob [3] where anthropometric data is used in combination with manually identified key points and apriori knowledge of age and gender. We build on this work by automatically locating facial anthropometric features and introducing a model that naturally incorporates the uncertainty over gender and age. As a result, gender, age and height can also be inferred from our model.

Our contributions are the following: We propose a model for measuring the height of a person while jointly estimating age, gender and facial feature points, based on a calibrated camera and anthropometric data. We introduce the idea of combining height estimation with appearance features for demographic recognition, and show that estimating height improves the recognition of demographic quantities. Further, by performing inference over age, gender, and height simultaneously with our model, we improve the accuracy of height estimation. Finally, we demonstrate the effectiveness of our model on a test set of 127 individuals to achieve height estimates with good accuracy.

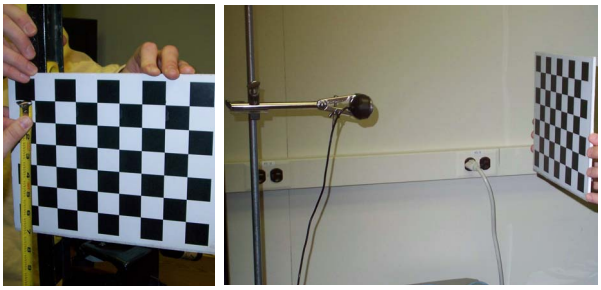


Figure 2. **Left**: During calibration, for one image a level is used to position the calibration target to be perpendicular with the floor. The distance between the floor and the world coordinate system origin is measured by hand. **Right:** Our camera is a standard webcamera with VGA resolution.

In Section 2, we introduce human height estimation with a calibrated camera. In Section 3, we describe data related to anthropometric features. Section 4 describes our model of the relationship between height, gender, age and anthropomorphic data. Section 5 contains experimental results.

2. Calibrated Camera Height Estimation

As is well known, a camera can be modeled as a projective pinhole [15] to map world points $\mathbf{X}$ to image points $\mathbf{x}$ according to the following relationship:

$$\mathbf{x} \equiv \mathtt{P}\mathbf{X} \equiv \begin{bmatrix}\mathtt{A} & \mathtt{b}\end{bmatrix}\mathbf{X} \tag{1}$$

where the calibration matrix $\mathtt{P}$ is composed of internal camera parameters $\mathtt{K}$ and extrinsic parameters including a coordinate rotation matrix $\mathtt{R}$, and translation $\mathbf{t}$ as follows: $\mathtt{P} = \mathtt{K}\begin{bmatrix}\mathtt{R} & \mathbf{t}\end{bmatrix}$. In the form shown in (1), the 3×3 matrix $\mathtt{A} = \mathtt{KR}$, and the 3×1 matrix $\mathtt{b} = \mathtt{K}\mathbf{t}$. The matrix $\mathtt{P}$ essentially captures the relationship between image and scene points, and allows one to extract metric information from image coordinates. Each point in the image corresponds with a world line passing through the camera center.

2.1. Camera Calibration

We perform camera calibration using a checkerboard target according to the method of [24], and shown in Figure 2. The target defines the world coordinate system. As such, we ensure that for one image, the target is held perpendicular to the ground. Consequently, the world coordinate system axes are aligned with the physical ground plane (the y-axis is perpendicular to the ground plane, and the x- and z-axes are parallel to the ground plane). In addition, for this image, the distance h_y from the coordinate origin $\begin{bmatrix}0 & 0 & 0\end{bmatrix}^T$ is measured by hand, as shown in Figure 2. The floor has the equation $y = -h_y$ in the world coordinate frame. Calibration errors affect the quality of the algorithm results.

2.2. Estimating Subject Distance and Height

Our key idea is illustrated by Figure 1: Multiple feature points on a face image corresponding to pairwise anthropometric features define multiple rays in the world. The distribution of possible distances between the camera and the subject is functionally related to the distribution of the size of these anthropometric features. As the uncertainty in the anthropometric feature distribution is reduced (e.g. by concluding that the subject is an adult male), a corresponding reduction in the uncertainty of the distance to the camera is achieved. Furthermore, because the camera is calibrated, an improvement in our confidence about the distance to the subject is directly related to improvements in the determination of the height above the ground plane of each facial feature point.

Estimating Subject Distance: We consider pairwise anthropometric features, defined as the distance between two feature points on the human body. In world coordinates, the pairwise anthropometric feature F is described by a Gaussian distribution $N(\mu_F, \sigma_F^2)$ over a measurement metric. Each feature F has a corresponding pair of image points $\mathbf{f} = \begin{Bmatrix} \mathbf{x}_i & \mathbf{x}_j \end{Bmatrix}$.

A world line $\mathbf{L}$ passing through a particular image feature point $\mathbf{x}_i$ has the equation $\mathbf{L}_i = \mathbf{\Omega} + t\omega_i$ where the camera center is $\Omega = -\mathtt{A}^{-1}\mathtt{b}$ and the vector pointing from Ω to the feature point $\mathbf{x}_i$ is $\omega_i = \mathtt{A}^{-1}\mathbf{x}_i$.

The angle ϕ between two feature lines $\mathbf{L}_i$ and $\mathbf{L}_j$ is:

$$\phi = \cos^{-1}\left(\frac{\omega_i^T \omega_j}{|\omega_i||\omega_j|}\right) \tag{2}$$

and the distance d in world coordinates from the camera center Ω to the midpoint of two feature points on the human body having separation distance d_F is:

$$d = d_F \frac{1}{2\tan(\phi/2)} \tag{3}$$

The distribution of the distance d is represented as a Gaussian $N(\mu_d, \sigma_d^2)$ where the parameters are found by considering that (3) is a linear function of random variable F. Consequently, $\mu_d = \mu_F \frac{1}{2\tan(\theta/2)}$ and $\sigma_d = \sigma_F \frac{1}{2\tan(\theta/2)}$.

In summary, our knowledge about the distributions of pairwise anthropometric features is exploited to estimate the distance between the subject and the calibrated camera.

Estimating Subject Height: From a subject to camera distance estimate d_i, the feature point can be approximately located (assuming the pair of feature points is parallel to the image plane) in the world coordinate frame as:

$$\hat{X}_i = \Omega + d_i \frac{\omega_i}{|\omega_i|} \tag{4}$$

Because our world coordinate frame is axis-aligned with the physical world (the xz-plane is parallel with the ground), the height of a point h_i above the ground is simply:

$$h_i = \begin{bmatrix} 0 & 1 & 0 \end{bmatrix} \hat{X}_i + h_y \tag{5}$$

The estimate for the subject's stature is based on the pairwise anthropometric feature of the eye centers F_e. The stature of a person is the height of the eyes above the ground, plus the distance from the eyes to the top of the head $F_{v,en}$, as reported in [9]. Note that this dimension $F_{v,en}$ has a distribution over gender and age and in practice, the expected value of this distribution is used.

$$h = \begin{bmatrix} 0 & 1 & 0 \end{bmatrix} \hat{X}_i + h_y + F_{v,en} \tag{6}$$

As with distance, the distribution of height h is represented with a Gaussian, where the parameters are derived by considering h as a function of the random distribution over distance d.

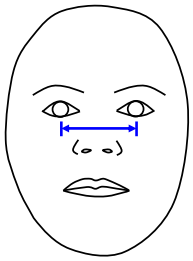
3. Age, Gender, and Anthropomorphic Data

There exists a great amount of data describing the distribution of measurements of the human body [9, 13, 21]. Our goal is to use pairwise anthropometric features to infer subject to camera distance, height, age and gender. Ideal anthropometric features are those that markedly change in size with age and gender. We have the additional practical requirement that the corresponding image of each feature point can be reliably located in the image automatically with an Active Shape Model [5].

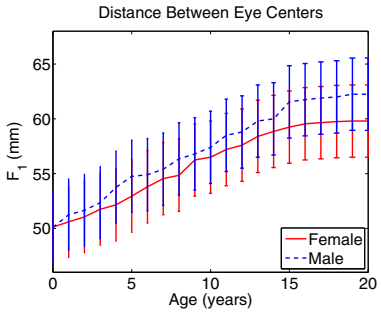
We consider two pairwise anthropometric features, illustrated in Figure 3. The size distributions as functions of age and gender for each of these pairwise anthropometric features is derived by smoothing data from [9]. The first feature F_1 is the distance between eye centers, and the second F_2 is the distance between the mouth and the nasion (i.e. the intersection of the nose and forehead). Our automatic detection of the associated feature points on several images is shown in Figure 4.

4. Anthropometric and Demographic Model

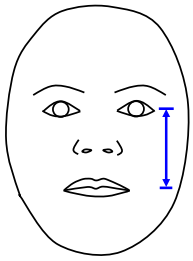
We would like to represent the relationships between a person's age, gender, height and appearance in the image. Of course, our degree of uncertainty about one attribute affects our belief about others. For example, if we are confident that a subject is tall (e.g. 190 cm), then it is more likely that the subject is an adult male than an adult female. However, it is intractable to learn the relationship between all quantities simultaneously. Our model incorporates conditional independence assumptions to make inference tractable and allows inference over all quantities in a unified manner.



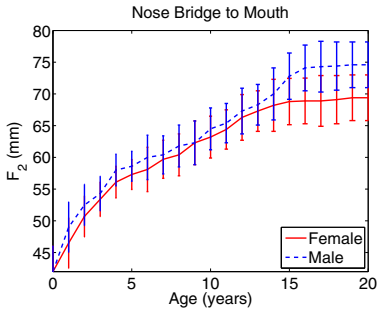
(a) Eye Centers



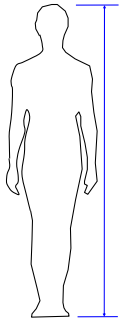
(b) The distribution over age and gender.



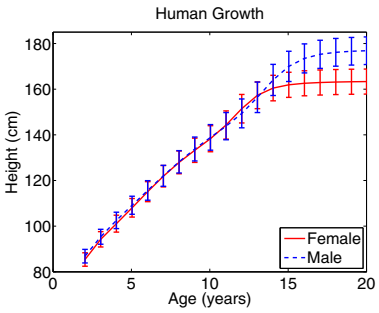
(c) Nasion to Mouth



(d) The distribution over age and gender.



(e) Height



(f) Growth curves over age and gender.

Figure 3. The two pairwise anthropometric features we use in this paper are the distance between eye centers (Top), and the distance between the mouth and nasion (the point between the eyes where the nose bridge meets the frontal bone of the skull) (Middle), which have known distributions with respect to age and gender. The relationship between gender, age, and height is also shown (Bottom). Error bars represent one standard deviation.

Figure 5 shows a graphical representation of our model. We represent the demographic and anthropometric quantities as random variables in the model. Each subject has an age A, gender G, height H, and distance from the camera D. The true value of the subject's i^{th} pairwise anthropomorphic feature is denoted by the variable F_i and the set of all such features is $\mathbf{F}$. Observed evidence includes a set of image points for each pairwise anthropometric feature $\mathbf{f}$, the camera calibration parameters $\mathtt{P}$, and appearance features extracted from the pixel values of the face region cor-

Figure 4. Example images with automatically recovered key points corresponding to two pairwise anthropometric features. The eye center distance is related to the distance between the circles, and the mouth to nasion feature points are marked with the symbol '+'.

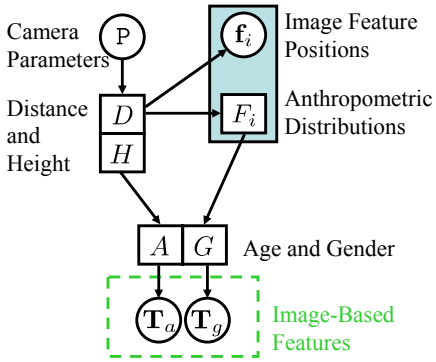


Figure 5. Our graphical model to infer over age A, gender G, height H, and camera to subject distance D, based on the evidence that includes the camera parameters $\mathtt{P}$, the extracted feature points $\mathbf{f}_i$, the anthropometric feature distributions F_i and the appearance features $\mathbf{T}_a$ and $\mathbf{T}_g$ related to age and gender respectively. Hidden variables are squares, with adjacent squares representing joint variables, and observed variables are circles.

responding the age $\mathbf{T}_a$ and gender $\mathbf{T}_g$. Our model includes simplifying conditional independence assumptions. For example, we assume that once age and gender are known, the facial appearance is independent of the height of the subject. Further, once the subject height and pairwise anthropometric measurements are known, the calibration parameters provide no further insight regarding the subject's demographic information. The structure of the Bayes Network is selected to exploit known relationships documented with publicly available statistics as well as known relationships from perspective geometry.

The model represents joint distribution of the variables as a product of conditional probability terms:

$$P(A,G,H,\mathbf{F}|\mathtt{P},\mathbf{f},\mathbf{T}_a,\mathbf{T}_g) \propto P(H|D)P(D|\mathtt{P})P(A|\mathbf{T}_a)P(G|\mathbf{T}_g) \prod_i P(A,G|H,F_i)P(F_i|D)P(D|f_i) \quad (7)$$

Gaussians are used to represent the distributions over variables related to distance (D, H and $\mathbf{F}$). Gender G is a bi-

nary variable $G \in \{\texttt{male}, \texttt{female}\}$. Age A is a discrete variable with a set of 125 possible states corresponding to the ages 0 to 124 years. In the following sections, we describe the terms of our model and inference with the model.

4.1. Estimating Age and Gender from Appearance

Our model employs appearance-based age and gender classifiers. These content-based classifiers provide probability estimates $P(G|\mathbf{T}_g)$ and $P(A|\mathbf{T}_a)$ that the face has a particular gender and age category, given the corresponding visual appearance features.

Our gender and age classifiers were motivated by the works of [11, 14] where a low dimension manifold for the age data. An independent set of 4550 faces is used for training. The age and gender of each person was labeled manually. To establish age ground truth, we labeled each face as being in one of seven age categories: 0-2, 3-7, 8-12, 13-19, 20-36, 37-65, and 66+, roughly corresponding to different life stages. Using cropped and scaled faces (61×49 pixels, with the scaling so the eye centers are 24 pixels apart) from the age training set, two linear projections ($\mathbf{W}_a$ for age and $\mathbf{W}_g$ for gender) are learned. Each column of $\mathbf{W}_a$ is a vector learned by finding the projection that maximizes the ratio of interclass to intraclass variation (by linear discriminate analysis) for each pair of age categories, resulting in 21 columns for $\mathbf{W}_a$. A similar approach is used to learn the gender subspace $\mathbf{W}_g$. A set of seven projections is found by learning a single projection that maximizes gender separability for each age range.

The distance d_{ij} between two faces is measured as:

$$d_{ij} = (\mathbf{T}_i - \mathbf{T}_j)\mathbf{W}\mathbf{W}^T(\mathbf{T}_i - \mathbf{T}_j)^T \qquad (8)$$

For classification for both age and gender, the nearest N training samples (we use $N = 101$) are found in the space defined by $\mathbf{W}_a$ for age or $\mathbf{W}_g$ for gender. The class labels of the neighbors are used to estimate $P(A|\mathbf{T}_a)$ and $P(G|\mathbf{T}_g)$ by MLE counts. One benefit to this approach is that a common algorithm and training set are used for both tasks, only the class labels and the discriminative projections are modified.

4.2. Anthropometrics from Age and Gender

By assuming conditional independence between height and anthropometric features when age and gender are known, we can show that the term $P(A, G|H, F_i)$ required by the model is proportional to $P(A, G|F_i)$ and $P(A, G|H)$. For $P(A, G|F_i)$, the conditional distribution of age and gender given a particular pairwise anthropometric feature F_i. This term is provided by the statistical data of [9], illustrated in Figure 3 for the two anthropometric features we consider.

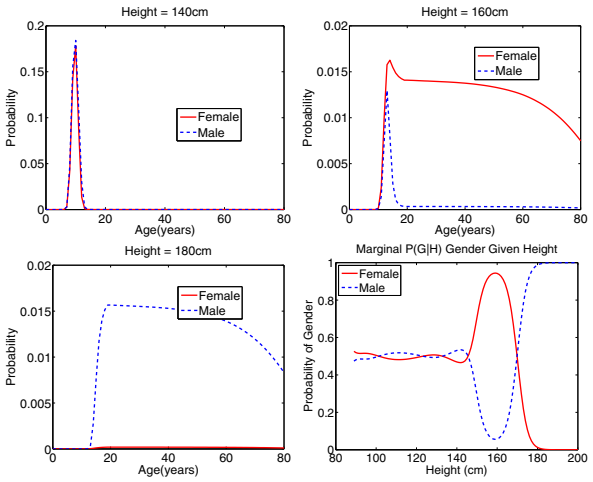


Figure 6. Illustrations of $P(A, G|H = h)$, the joint distributions over age and gender given height, for several different heights h. **Top Left:** When $h = 140$ cm, the subject age distribution is centered at 10 years with nearly equal likelihood of each gender. **Top Right:** There are two reasonable explanations when height is $h = 160$ cm. Either the subject is an adult female, or an adolescent male in the process of "growing through" that height. **Bottom Left:** A person with a height $h = 180$ cm is most likely a male. **Bottom Right:** The marginal distribution of gender given height. Note the peaks at heights common for adult women and men.

4.3. Distance and Height

The relationship between the camera parameters $\mathtt{P}$, the pairwise demographic features $\mathbf{F}$, the corresponding features $\mathbf{f}_i$ in the image, and distance to the subject D is a deterministic function of random variables, described in Section 2.2. Therefore, the term $P(D|F_i, \mathbf{f}_i, \mathtt{p})$ is simply a function of a random variable, where the distribution of F_i is related to the the distribution of D. Likewise, the term $P(H|D)$ is also a deterministic function of the random variable distance D (4)-(6).

4.4. Height, Age, and Gender

Our model requires the term $P(A, G|H = h)$, the conditional distribution of age and gender given height. This term is provided by the statistical data of [21] and is illustrated in Figure 6. The conditional probability of age and gender given height is found with $P(A, G|H) \propto P(H|A, G)P(A)P(G)$, with the simplifying assumption that age and gender are independent. The gender prior $P(G)$ is assumed to be equal for each gender ($P(G = \texttt{male}) = 0.5$), and the prior for age $P(A)$ is based on life expectancy from a standard actuarial table [1].

We make several observations. First, the conditional distribution $P(A, G|H)$ is not well-modeled with a Gaussian distribution because of the rapid growth in the adolescent years, justifying our decision to represent age as a discrete

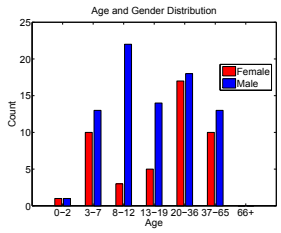


Figure 7. The distribution of the 127 subjects used in our study.

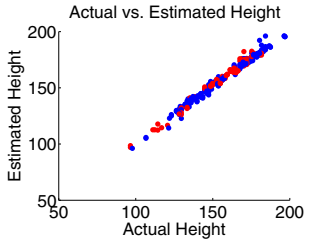


Figure 8. A scatter plot of the estimated and actual height (cm) of subjects in our study.

variable. Second, we note that for adults aged 20 or greater, 170 cm represents the optimal decision boundary to classify gender when height is the only available information. Finally, we mention that our model does not consider the phenomena of stature loss among the elderly, but this effect could be added if the relevant statistical data are available.

4.5. Inference with Expectation Maximization

We perform inference on our model to consider all the evidence from an image of a subject captured with the calibrated camera, and find the distribution over age, gender, height and distance to the camera. Final classifications are based on the maximum likelihood aposterior distributions for age $\hat{a}$, gender $\hat{g}$, height $\hat{h}$, and distance $\hat{d}$ from the camera. For each variable, our final estimate is the assignment that maximizes its marginal distribution obtained by marginalizing over all other variables.

For computational efficiency, we do not perform exact inference over the entire model. Instead, similar to [23], we use Expectation Maximization to simplify inference. In the E-step, we fix the distribution over F_i as a unidimensional Gaussian and perform inference on the model. In the M-step, the distribution over each anthropometric feature F_i is updated using the winner-take-all variant of EM [22] based on the most likely estimate of age a^* and gender g^* as $P(F_i|A = a^*, G = g^*)$. In our case, the winner-take-all variant has the advantage that, in inference, each anthropometric distribution remains a Gaussian. After convergence, the most likely assignment of each variable is found. With this inference, our algorithm is fast enough for real-time application at video frame rates.

	Height		Distance		Age		Gender
	MAD	STD	MAD	STD	MAD	STD	Error
Height	30.5	40.9	182	201	-	-	-
Model+$\mathbf{T}_a$, $\mathbf{T}_g$	-	-	-	-	8.5	12.3	32.8%
Model+P, $\mathbf{f}_i$	**24.1**	**26.7**	142	**167**	7.0	10.6	35.3%
Full Model	**24.1**	**26.7**	**136**	171	**5.4**	**9.7**	**28.1%**

Table 1. By reasoning about gender, age and height with our full model we achieve the best overall results for predicting age and gender. Errors (mean absolute and standard deviation) are shown for height and distance. Age errors are in years, and gender classification error rate is shown. Results are shown for height alone (no modeling of age or gender), using the model but observing only appearance features, using the model but observing only height (no appearance), and using the full model.

5. Experiments

Our model was tested on images of 127 subjects ranging in age from 2 to 56 with a total of 81 male and 46 female subjects. To sample from a wide variety of demographics, subjects were recruited in several different venues (a science museum, a research laboratory, and an educational institution) on four different occasions. The gender and age distribution of subjects is reported in Figure 7. Most subjects are Caucasian, but a wide variety of ethnicities participated. Each subject reported his or her age (binned into one of 14 bins) and gender, and a stadiometer was used to measure each subject's height. Subjects were photographed looking toward the camera as our model currently assumes a frontal facial pose, making the eyes and mouth coplanar with the image plane (3). The camera height is about 160 cm off the ground, but this varied at each session. Two pieces of tape were placed on the floor at different distances from the camera, one near (ranging from 0.91 m to 1.63 m) and one far (ranging from 1.80 m to 2.69 m). Each subject was photographed at the two distances marked by the tape. The camera has VGA resolution (480×640 pixels). The entire procedure requires about five minutes for each subject. A total of 237 images are used in our experiments (two images for most subjects; 17 subjects have only one image).

For detecting faces, we use a commercial package that implements a cascade face detector similar to [17]. An active shape model [5] is applied to recognize key points on the face, as illustrated in Figure 4. Finally, for each subject image, inference is performed with our model in Figure 5 to obtain maximum likelihood aposterior estimates for age $\hat{a}$, gender $\hat{g}$, height $\hat{h}$, and distance $\hat{d}$ from the camera.

5.1. Height and Distance Accuracy

Table 1 reports the accuracy of the model on our test set for height, distance to the subject, age, and gender. We compare height estimation with the baseline approach where age and gender are not in the model, and the anthropometric dis-

	Height		Age		Gender
	MAD	STD	MAD	STD	Error
Single-frame	25.9	24.1	**5.9**	10.1	27.3%
Multi-frame	**22.4**	**22.1**	6.2	**9.8**	**24.5**%

Table 2. Compared with performing inference on each frame individually (first row), using evidence from multiple frames (second row) improves the accuracy.

	Stature		Distance		Age		Gender
	MAD	STD	MAD	STD	MAD	STD	Error
Children(0-16)	22.4	22.3	106	116	**0.5**	**0.9**	29.6%
Adults(17+)	22.3	22.9	120	130	11.6	11.9	**19.6%**

Table 3. Age classification is an easier problem for children, and gender classification is easier for adults. Height estimation performs well across age. These results include using evidence from two images of the same subject, when available.

tributions are from the entire population, marginalizing over age and gender. Overall, the complete model estimates human height with an accuracy of 26.7 mm in standard deviation, reducing the error of the baseline approach by 34.7% (from 40.9 mm). Figure 8 shows a scatter plot of the true and estimated statures of the subjects.

This result is believed to be the most accurate automatic result achieved on a large dataset. In [3], estimation error of about 50 mm in standard deviation is reported on a test set of 27 adults, where the model has full knowledge of gender and feature points are manually labeled. In [6], a reference length from the scene is required, and the result on a single subject is within 2 cm. Finally, in [18], height is estimated by a calibrated camera detecting the full silhouette of the subject. On three subjects, this achieves an estimation error with standard deviation of 43 mm.

We estimate the distance between the subject and the camera with an accuracy of 171 mm in standard deviation. In reporting this result, it is noted that the distance to the subject is somewhat variable as each subject's interpretation of "standing on the tape" varied. Therefore, we expect that our reported results represents an upper (i.e. pessimistic) bound on the achievable distance accuracy. Further, it should be noted that height accuracy is positively correlated with both calibration accuracy and distance from the camera. The height estimation error ranged from 15.2 mm for the most accurate calibration to 32.8 mm when the calibration was poorest.

5.2. Combining Multiple Observations

Evidence from multiple observations is combined to estimate the age, gender, and height of a person using a Naïve Bayes model with an assumed uniform prior over the variable in question. For example, when estimating height from multiple images:

$$P(H|\mathbf{e}_1, \ldots, \mathbf{e}_N) = \prod_{n=1}^{N} P(H|\mathbf{e}_n) \tag{9}$$

where $\mathbf{e}_n$ represents all the available evidence associated with the $n^{\mathtt{th}}$ image capture.

Table 2 reports the result of consolidating evidence from multiple frames (both the near and far image captures) for each subject. Overall, more accurate height estimates and gender classifications are achieved, but age estimation improvements are inconclusive based on which critera is examined. Note that the 17 subjects having only one image are omitted from this analysis.

5.3. Gender and Age Accuracy

By using our model to infer gender and age using both appearance and height, we achieve better accuracy than using either one alone, as reported in Table 1. Our appearance classifier achieves 67.2% gender accuracy by itself. This is lower than the results reported for this task using facial appearance (e.g. [2]), but our test set includes a large number of children who have yet to develop gender-specific facial features. Combining height with appearance by our model improves the gender classification accuracy to 71.9%.

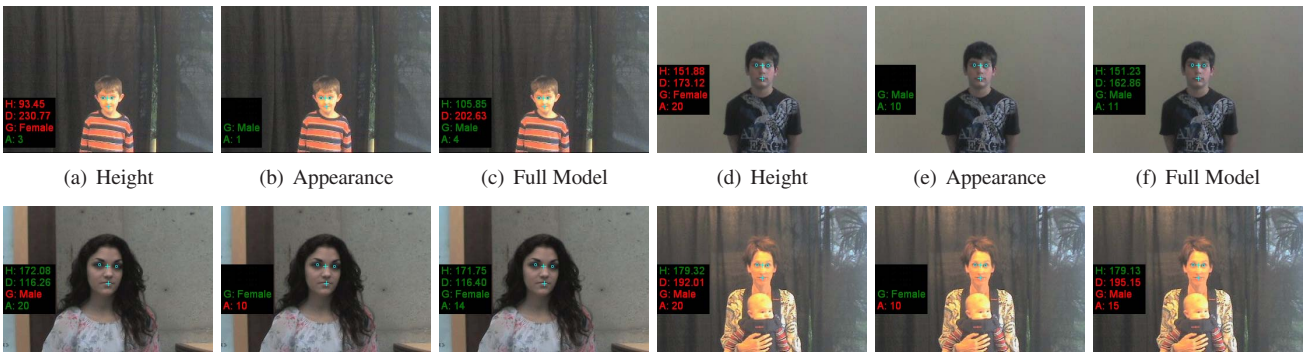
Each subject self-reported his or her age as belonging to one of 14 age bins. Using our model, we find the most likely aposterior age $\hat{a}$, and compare this with the ground truth age bin for the subject. When $\hat{a}$ falls within the bounds of the age bin, the age error is zero, otherwise the age error is the number of years between the estimated age $\hat{a}$ and the closest bound on the true age bin. Again, by inferring age with combined appearance and height features, we achieve better age estimation than using either feature type alone.

More insight is gleaned by examining the performance on children (ages 0-16) and adults (17+). Table 3 shows that age is easier to estimate for children, and gender classification is more accurate in adults. This result is explained by considering our pairwise anthropometric features, as shown in Figure 3. For age estimation, the gradient of each feature with respect to age is greatest during childhood. However, the greatest separation between the genders for the distributions for any of the anthropometric features given age occurs when adulthood is reached.

Figure 9 discusses the height, age, and gender estimates for several images from our dataset.

6. Conclusion

We introduce a model to unify inference over demographic quantities and anthropometric features using a calibrated camera. Instead of considering demographic classification and height estimation as separate problems to be



(a) Height (b) Appearance (c) Full Model (d) Height (e) Appearance (f) Full Model

(g) Height (h) Appearance (i) Full Model (j) Height (k) Appearance (l) Full Model

Figure 9. Height, age, and gender classification improve through our model that reasons over variables related to appearance, height, demographics and pairwise anthropometric features. In each group of images, the model outputs are shown when height is observed (no appearance features), appearance is considered (height is not estimated), and the full model is used. Accurate results are shown in green text and poor results are in red text. The facial appearance in (b) allows the mistaken gender from height alone (a) to be corrected in the full model (c). In (d), the subject's height is similar to an adult woman, but appearance recognized the subject as a young male (e), and the full model finds the most probable explanation is that the subject is an adolescent male (f). The incorrect age classification from appearance alone (h) is corrected by height estimation in (g) to produce the reasonable estimates in (i). A failure is shown in (j)-(l). The subject is a tall female, and the correct gender from appearance (k) is not strong enough to override the fact that few females are 179 cm in height from (j), and in the final result (l), the demographic classification is worse than from appearance only (k). Best viewed electronically.

solved independently, our model merges these problems and allows influence to flow throughout the variables.

We provide evidence our model's effectiveness by testing on images from 127 subjects spanning a wide age range to achieve accurate automatic height estimation. We show that when height provides context and is considered along with facial appearance, the age and gender estimates improve versus using appearance alone. Likewise, height estimation improves with our model which reasons about age and gender as hidden variables. Our model is extensible in that additional pairwise demographic features can be added, assuming the corresponding feature points can be located.

References

[1] E. Arias. United States life tables, 2003. Technical report, National Center for Health Statistics, 2006.

[2] S. Baluja and H. Rowley. Boosting sex identification performance. In *IJCV*, 2007.

[3] C. BenAbdelkader and Y. Yacoob. *Statistical Estimation of Human Anthropometry from a Single Uncalibrated Image*. Springer Press, 2008.

[4] T. Berg, A. Berg, J. Edwards, M. Maire, R. White, Y.-W. Teh, E. Learned-Miller, and D. Forsyth. Names and faces in the news. In *Proc. CVPR*, 2004.

[5] T. Cootes, C. Taylor, D. Cooper, and J. Graham. Active shape models-their training and application. *CVIU*, 1995.

[6] A. Criminisi. *Accurate Visual Metrology from Single and Multiple Uncalibrated Images*. PhD thesis, University of Oxford, 1999.

[7] A. Criminisi, I. Reid, and A. Zisserman. Single view metrology. *International Journal of Computer Vision*, 40:2000, 1999.

[8] C. Dupertuis and J. Hadden. On the reconstruction of stature from long bones. *American Journal of Physical Anthropology*, 1951.

[9] L. Farkas. *Anthropometric facial proportions in medicine*. Raven Press, New York, 1994.

[10] A. Gallagher and T. Chen. Estimating age, gender, and identity using first name priors. In *Proc. CVPR*, 2008.

[11] X. Geng, Z.-H. Zhou, Y. Zhang, G. Li, and H. Dai. Learning from facial aging patterns for automatic age estimation. In *ACM MULTIMEDIA*, 2006.

[12] B. Golomb, D. Lawrence, and T. Sejnowski. Sexnet: A neural network identifies sex from human faces. In *Proc. NIPS*, 1990.

[13] C. Gordon, B. Bradtmiller, T. Churchill, C. Clauser, J. McConville, I. Tebbetts, and R. Walker. 1988 anthropometric survey of US army personnel: Methods and summary statistics. *Technical Report NATICK/TR-89/044, AD A225 094*, 1988.

[14] G. Guo, Y. Fu, C. Dyer, and T. Huang. Image-based human age estimation by manifold learning and locally adjusted robust regression. In *IEEE Trans. on Image Proc.*, 2008.

[15] R. Hartley and A. Zisserman. *Multiple View Geometry in Computer Vision*. Cambridge University Press, ISBN: 0521540518, 2004.

[16] D. Hoiem, A. Efros, and M. Hebert. Putting objects in perspective. In *Proc. CVPR*, 2006.

[17] M. Jones and P. Viola. Fast multiview face detector. In *Proc. CVPR*, 2003.

[18] I. Kispál and E. Jeges. Human height estimation using a calibrated camera. In *Proc. CVPR*, 2008.

[19] J.-F. Lalonde, D. Hoiem, A. A. Efros, C. Rother, J. Winn, and A. Criminisi. Photo clip art. *ACM Trans. SIGGRAPH*, 2007.

[20] M.-F. Lv, M.-T. Zhao, and F.-R. Nevatia. Camera calibration from video of a walking human. *IEEE Trans. Pattern Anal. Mach. Intell.*, 28(9):1513–1518, 2006.

[21] National Center for Health Statistics. CDC growth charts, United States. http://www.cdc.gov/nchs/data/nhanes/growthcharts/zscore/statage.xls, 2007.

[22] R. Neal and G. Hinton. A view of the em algorithm that justifies incremental, sparse, and other variants. In *Learning in Graphical Models*, 1998.

[23] E. B. Sudderth, A. Torralba, W. T. Freeman, and A. S. Willsky. Learning hierarchical models of scenes, objects, and parts. In *Proc. CVPR*, 2005.

[24] Z. Zhang. A flexible new technique for camera calibration. *IEEE PAMI*, 2001.