Discriminating Projections for Estimating Face Age in Wild Images

Ryan Tokola, David Bolme,
Christopher Boehnen
Oak Ridge National Laboratory
tokolara, bolmeds, boehnencb@ornl.gov

Del Barstow
obarstow@gmail.com

Karl Ricanek
University of North Carolina
at Wilmington
ricanekk@uncw.edu

Abstract

Despite the fundamental variability of human appearance, the last several years have seen considerable advances in age estimation from images of faces. Many of these advances have been made possible by artificially removing external sources of variability—they focus on highly constrained images from datasets such as the MORPH face database and FG-NET. We introduce a novel approach to estimating age from a single "wild" image, where pose, illumination, expression, face size, and face occlusions are not managed. Our method is able to reduce the effects of variations that already exist within in image.

Using pose-specific projections, we map image features into a latent space that is pose-insensitive and age-discriminative. Age estimation is then performed using a multi-class SVM. We show that our approach outperforms other published results on the Images of Groups dataset [9], which is the only age-related dataset with a non-trivial number of off-axis "wild" face images. We also show results that are competitive with recent age estimation algorithms on the mostly-frontal FG-NET dataset, and we experimentally demonstrate that our feature projections introduce insensitivity to pose.

1. Introduction

There have been dramatic advances in face detection and recognition in the last decade, as demonstrated by Phillips et al. [21]. Unfortunately, progress achieved in recognition has not directly led to improvements in facial analytics.

Over the last several years there has been significant interest in systems that can predict biographical information from a single face image. There are innumerable environmental, genetic factors, and medical conditions that alter the appearance of a person's face, which make automatic techniques for determining biographical details such as gender, race, and age extremely challenging, even in controlled environments.

The problem worsens when wild images are used. Features are much less consistent across images when there are significant variations in pose and illumination, and the generally unpredictable nature of wild images adds yet another level of difficulty ("is that person's face wrinkle-free or is it a low resolution photo?").

Most age estimation algorithms don't explicitly account for the variations found in wild data, even algorithms designed to operate on uncontrolled images. Some introduce insensitivity to variables such as face shape and lighting, but the assumption of a frontal image is virtually universal. In this paper we propose an approach that mitigates the effect of image variations while preserving age discrimination. Our current implementation focuses on pose variation, but our overall framework can be extended to account for other sources of variation, such as illumination, gender, expression, and race. In our approach pose-specific projections are applied to image features. These projections map the features to a latent space that is insensitive to pose.

Our algorithm has three primary distinguishing characteristics. (1) Instead of one classifier that takes features from the entire face, we build independent sets of features from several specific regions of the face. This in itself improves robustness to differences in pose, face shape, or other sources of spatial variability. By considering the regions independently of each other, local confusers such as occlusions or sunglasses will have a reduced impact on the overall classification. (2) For each region, we learn projections to a latent space that is discriminative with respect to age. There are different projections for different poses, but they all map to the same latent space. This allows direct comparison between images of faces with dramatically different poses. (3) Instead of projecting from image space we project from a feature space. There are a few advantages to this. First, our feature space has a lower dimensionality than images. Second, the image features introduce some degree of robustness to variations, which results in a more stable projection. For example, a change in lighting may have unpredictable effects on an image-based projection, but will have no effect on a projection from features that are themselves lighting invariant.

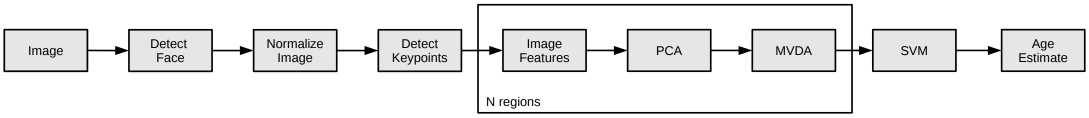

Figure 1. Age estimation pipeline.

2. Related works

2.1. Age estimation

There are several approaches to representing the human face for age estimation. Although some of the works discussed below have some insensitivity to pose, only one directly addresses the problem of pose variation.

• **Anthropomorphic models.** Kwon and da Vitoria Lobo [17] use ratios of distances between facial landmarks to distinguish between babies and adults. Ramananathan and Chellappa [22] learn a continuous progression in face age with similar ratios.

• **Active appearance models (AAM).** Luu et al. [19] use AAMs to encode a representation of the face. The locations of the AAM's landmarks are used to warp the image to align with a "neutral" set of landmarks and image features are extracted from the warped image. Image features are mapped to age estimates using support vector regression. It is important to note that the image warping helps provide some amount of robustness to pose variation. Chen et al. [5] explore various methods of model selection to accomplish discriminative dimensionality reduction.

• **Age manifolds.** A number of works learn mappings between face images and a low dimensional manifold [8], [7]. A regression function is used to estimate age given the projection of an image onto the manifold. These methods are sensitive to variations in pose and image normalization because raw images are used for the projection. Furthermore, the use of raw images necessitates a very large training set due to their high dimensionality.

• **Biologically-inspired models.** Taking a cue from studies of human perception and recent efforts at object category recognition, some works apply alternating layers of simple and complex processing to a face image that has been filtered with a set of Gabor filters [12]. Dimensionality reduction is applied to the features and then a regressor or classifier can be learned.

• **Compositional models.** A complex model of the human face, based on the compositional model from Xu et al. [26], is proposed by Suo et al. [25].

Suo et al. [25] add an age-based dynamic element to the model, with particular attention paid to how wrinkles develop with age.

• **Patch-based models.** Several works follow the grid-based approach introduced in Ahonen et al. [1], which extracts image features from a grid of image patches. Shan [23] uses features from local binary pattern (LBP) and Gabor wavelet features. Others use completed local binary patterns, which are a variant of LBP that include contrast information [32], [31]. Alnajar et al.[2] use soft assignments of codewords in an attempt to make the encoding less sensitive to image noise and illumination changes.

Li et al. [18] cluster patches from training images into age- and pose-based code groups. At test time a feature vector is created that indicates the distance between the query image and each of the code groups. This approach assumes that relatively simple distances are capable of providing good discrimination between code groups. The authors demonstrate that their framework performs better than image-based classifiers, but there is no comparison to other published methods.

Yan et al. [29] characterize faces with a mixture of spatially-flexible image patches. As with the AAMs, the ability of the image patches to move within an image will help compensate for small differences in pose.

• **Hybrid approaches.** Choi et al. [6] combine global AAM features with local image features. Facial landmarks are used to define regions in which to measure the responses of Gabor filters (for wrinkle detection) and LBP features (for skin texture description). All features are then combined and a coarse age bin classification is followed by specific age estimation.

• **Contextual models.** Gallagher and Chen [9] attempt to exploit the spatial relationships between people in groups. Analysis of a large dataset of groups of people revealed correlations between the age of people and their relative locations in an image.

2.2. Multi-view face recognition

Pose variation can dramatically complicate face recognition. To address this, some works have explored projecting face images into a pose-neutral latent space. Sharma et al. [24] propose a general multi-view approach called generalized multiview analysis (GMA), which is a supervised extension of canonical correlation analysis. Kan et al. [15] introduce a related method called multi-view discriminant analysis (MvDA). Both of these methods attempt to learn a set of pose-specific projections that minimize the distance between instances of one class (in this case, class corresponds to the identity of the person) and maximizes the distance between classes.

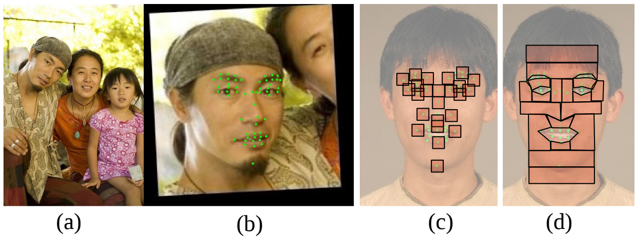

Figure 2. (a) Example from the Images of Groups dataset. (b) After normalization and landmark detection. (c,d) Image regions.

3. Robust age estimation in wild images

A general overview of our age estimation pipeline can be seen in Figure 1. To briefly summarize our algorithm: face detection is used to spatially normalize the image and locate facial keypoints. The keypoints are used to define a set of regions and within each region a set of features is extracted. The dimensionality of the features is reduced with PCA, and then the resulting feature vector is projected into a latent space using MvDA. The projected features from each face region are concatenated together and fed into a SVM-based classifier to provide the final age estimate.

3.1. Preprocessing

We use the PittPatt [20] face detection software to find faces in an image and locate major facial landmarks. Given the location of two eyes (or if both eyes are not found, an eye and a nose), we use an affine transform to rotate and scale the image so that the eyes in all images are aligned. All images are also cropped to a uniform size.

PittPatt also provides a partial pose estimate (in terms of roll and yaw). Images with a negative yaw are flipped.

We use an extension of [3] that was provided by Kriegman-Belhumeur Vision Technologies (KBVT) to locate 55 facial landmarks in each normalized image. An example of image normalization and keypoint detection can be seen in Figure 2.

3.2. Defining regions

Given the facial landmarks, we define image regions in two ways. First, we take square patches around 25 of the landmarks, as seen in Figure 2(c). These patches are intended to capture shape characteristics of parts of the face (eyes, nose, etc.). Second, we use the landmarks as reference points for 18 polygonal patches of the face, as seen in Figure 2(d). These are intended to capture characteristics of the skin, such as surface texture and wrinkles.

3.3. Features

Each of the N image regions defined in Section 3.2 is used to generate a feature vector: $f = [f_1, ...f_N]^\top$. In our specific implementation, $N = 43$.

Each feature vector f_n begins with the concatenation of four feature types, which will be discussed in detail below.

$$\hat{f}_n = [f_n^{Ga}, f_n^L, f_n^{GL}, A_n f_n^I]^\top \tag{1}$$

As we will discuss in section 3.5, the feature vector $\hat{f}_n$ is first projected into a subspace of lower dimensionality with basis vectors B_n. Next, the resulting vector is projected into the latent age-encoded space through the pose-dependent linear transformations W_n^p, where p is the estimated pose of the face.

$$f_n = W_n^p B_n \hat{f}_n, \tag{2}$$

- **Gabor filters** (f^{Ga})Finding wrinkles is either implicitly or explicitly at the core of many age estimation algorithms. Gabor filters are excellent features for detecting the fine lines that characterize wrinkles, and are consequently frequently used [30], [6]. It is common to either use the entire response of each filter or to use simple statistics (such as maximum value and standard deviation) in portions of the image. We take a somewhat different approach and build a set of 8-bin histograms of filter responses for each of our 43 image regions. Following Yang et al. [30] we use 5 scales and 8 orientations of Gabor filters.
- **Local Binary Patterns** (f^L)We again use histograms for LBP features, which are commonly used as a textural descriptor. We first calculate LBP features for the entire image and then build a histogram over the 256 LBP encodings in each image region.
- **GLCM statistics** (f^{GL})The gray-level co-occurrence matrix (GLCM) is another way of describing texture and is similar in spirit to LBP features. However, instead of directly using the encoded relationships between pixels, it is most common to use statistics from the GLCM matrix [13]. We use the four GLCM properties available in MATLAB's Image Processing Toolbox (contrast, correlation, homogeneity, and energy). GLCM matrices are calculated with single-pixel offsets at 0°, 45°, 90°, and 135°.
- **Image PCA** (f^I)Some face parts, such as the eyes and nose, have shapes that change with age. We learn PCA basis vectors from each of the image patches. We keep the first 50 basis vectors, which collectively capture approximately 90% of the variance in our training set. We use these features for the square image patches only, as the irregular shape of the polygonal patches make image-based PCA methods difficult. In Equation 1, f_n^I represents the set of image pixels in the n^{th} image patch, and A_n is the associated set of PCA basis vectors.

3.4. Dimensionality reduction

Altogether, for each polygonal image region we have 5×8 8-bin Gabor feature histograms (f^{Ga}), a 256-bin LBP histogram (f^L), and 4×4 GLCM features (f^{GL}), resulting in a $320 + 256 + 16 = 592$ dimensional feature vector. The same features are used for the square image regions, but we also include the first 50 image intensity basis vectors (f^I), resulting in a 642 dimensional feature vector.

These feature vectors are relatively large, and not all of their elements are particularly useful. For example, there are some LBP encodings that never appear in the training set. To improve the overall quality of the feature vectors, as well as to reduce the complexity of future calculations, we reduce the dimensionality of the feature vectors through PCA. This is done independently for each image patch. We keep enough basis vectors to account for 95% of the variance in the training set.

3.5. MvDA projection

At this point it would be perfectly valid to train classifiers or regressors on each $B_n\hat{f}_n$, but we are interested in enhancing our model's robustness and flexibility, particularly with respect to pose. To this end, we partition the training data into P pose bins and jointly learn MvDA projections for each pose bin.

MvDA is a method for jointly learning linear projections from different "views" to one common and discriminative latent space. It is easiest to think of views in terms of camera angles, but a "view" can apply to many different representational modalities. For example, one could learn MvDA projections in which a photograph is one considered to be from one view, a sketch from a second view, and a painting from a third. Here, we simply associate each of our P pose bins with a different view.

MvDA operates, much like Fisher Discriminant Analysis (FDA), by maximizing the inter-class variance and minimizing the intra-class variance. The nature of the class depends upon the desired application. When the goal is face recognition, MvDA vectors with be learned with all photos of a specific subject belonging to one class. We, however, are interested in a person's age, so we use age bins as classes. This is the critical moment at which we cripple our ability to estimate an exact age—the MvDA projections only care about age bins and an ideal set of MvDA vectors will project features from everybody in one age bin onto the same point in latent space.

What we lose in fine-grained accuracy (age estimation to the year), however, we gain in robustness and flexibility. A good MvDA projection will have several benefits. (1) The dimensionality of the feature vector will be further reduced. (2) Any variation in the training data will be reflected in the projection. For example, a projection that was trained on images with varying light sources will have a natural insensitivity to lighting. This also applies to pose. Even if only one pose label is used ($P = 1$), the MvDA projection will find the linear projection that is most robust to the pose variations in the training data. (3) Most importantly, the projections of images from different views can be directly compared to each other.

It is important to note that our use of relatively coarse age bins for class labels is an entirely practical matter. Given sufficient training data, it would be reasonable to learn projections with a separate class label for each year of age. With a limited number of images, it is necessary to strike a balance between a large number of class labels (and consequently a more fine-grained classification) and projections of high quality.

The use of MvDA projections has an interesting effect on the need for training data. Clearly, many training examples are required for every combination of view and class in order to learn successful projections. Once the projections have been made, training is dramatically simplified. In fact, a successful classifier can be learned even if training data is only available for one of the views. This is particularly important in the case of face age estimation. We wish to learn classifiers that can be applied to wild images over a wide range of poses, but the overwhelming majority of the training data consists of frontal images. We explore this issue in Section 4.4.2 and Figure 4.

We wish to point out one important difference between the MvDA projections that we learn, and the projections used in [15],[8], and others. Those works learn projections from image space, whereas we project from a feature space. Not only does our feature space have an inherently smaller dimensionality than an image (thereby reducing the number of required training images), but it is a smarter space. After all, the features themselves have been specifically designed to extract useful information from the image, and it can be expected that they would result in a more stable projection.

In Equation 2, W_n^v represents the set of direction vectors used to project $B_n\hat{f}_n$ into the latent age-encoded space when the face in the image has been estimated to have pose p. Please refer to [15] for details regarding the construction of MvDA projections.

3.6. Classification

We use the MATLAB interface to LIBSVM [4] to learn a C-SVC-type multi-class support vector machine (SVM) with a radial basis function (RBF) kernel for age group classification. Parameters for the SVM were learned with five fold cross-validation on training images.

4. Experiments

To properly assess our model's insensitivity to pose, we naturally must use a dataset with sufficient off-axis training data. Unfortunately, we are unaware of any datasets of people with known ages that have significant diversity in both age and camera angle. In the absence of an ideal dataset we give results on two datasets, each of which lacks one kind of diversity.

MLP	WAS	QM	AGES	RUN	BM	RPK	BIF		HIE		Ours	Ours	Ours
[28]	[11]	[28]	[10]	[28]	[27]	[29]	[12]	[6]	[19]	[5]	(ID)	(age)	(no proj.)
10.39	8.06	7.57	6.22	5.78	5.33	4.95	4.77	4.66	4.37	4.04	6.08	5.41	5.38

Table 1. Mean absolute error (MAE) on the FG-NET dataset. "Ours (ID)" uses identity-based projections that were learned from the PubFig dataset. "Ours (age)" uses age-based projections from the Images of Groups dataset. "Ours (no proj.)" does not use projections.

4.1. Datasets

The FG-NET Aging database is among the most commonly used datasets for evaluating contemporary age estimation algorithms. Although the database is small (with 1,002 total images) and mostly frontal (our algorithm is therefore unable to benefit from pose-specific projections) we present results on FG-NET to show how our framework compares to the best contemporary algorithms.

To better understand our algorithm's robustness to pose variation, we also present results on the Images of Groups (IoG) dataset [9]. The dataset consists of 5,080 images with a total of 28,231 labeled faces. The images were acquired through searches on the website Flickr. Ground truth annotations were not available, so Gallagher et al. manually assigned each face to one of seven age bins: 0-2, 3-7, 8-12, 13-19, 20-36, 37-65, and 66+. As the images were collected from searches, there is an extremely uneven distribution of images across age and pose. See Figure 3(a) for an illustration of the image distribution. In this figure, the cyan bars are used to show the number of images in the 20° to 30° pose bin. There are, for example, only 33 images of people in the 66+ age bin with a pose greater than 20°. A negligible number of images have a pose greater than 30°, and these are simply included in pose bins that are described at extending to 30°.

In the discussion below, by "pose" we refer to the yaw estimate given by PittPatt. We ignore pitch and compensate for roll through our image normalization.

4.2. Training

The FG-NET dataset contains images of 82 subjects, and testing is performed with a leave-one-person-out approach. Classifiers are trained on images of 81 of the subjects and evaluated on the remaining subject. This is repeated for each of the subjects. Results are reported by the mean absolute error (MAE) over all test samples.

FG-NET does not have a sufficient number of images to learn robust projections, so we must learn them elsewhere. We show results for three different approaches to projection: (1) We learn age-based projections from the IoG dataset, (2) we learn identity-based projections from the PubFig dataset [16], and (3) we bypass projections altogether and simply use raw features (after PCA dimensionality reduction). Because age-based projections that are learned from another dataset introduce age discrimination beyond that of FG-NET, they do not lead to an entirely fair comparison to other methods that only use age information from FG-NET itself. For this reason, we also learn projections from a different dataset that are based on identity, and not age. These projections do not carry any age information, leading to a more fair comparison to baseline methods. Finally, since the FG-NET is almost entirely frontal, pose-specific projections will be of limited value in the first place. For this reason, we show results without any projections at all. In all three of these cases, the final classifiers were learned using only data from the FG-NET dataset. Because the dataset is evaluated in terms of MAE, we learn a support vector regressor (SVR) instead of a multiclass SVM.

Evaluation on the IoG dataset typically relies on the random selection of training and testing images. We followed the same procedure as in [9]: training uses 3500 images that are randomly selected with the constraint that an equal number of images fall in each age bin, and testing is performed on 1050 independent images that are also uniformly distributed. Because of the relatively small number of images in any given test set, we performed independent random experiments 10 times and averaged the results.

To learn the image intensity PCA basis vectors we used images from the Labeled Faces in the Wild dataset [14]. This avoids the extra computational burden of learning the PCA basis vectors and applying the projections with every train/test fold. Please note that the image intensity PCA basis vectors are agnostic with respect to age, so any advantage from using data external to the training set is negligible. We relearn feature PCA basis vectors and MvDA projections for each of the 10 evaluation trials.

4.3. FG-NET results

Results on the FG-NET dataset are shown in Table 1. Our performance does not exceed every previous method, but this is not the best platform for evaluating our algorithm as we emphasize robustness to pose variation, which is largely absent from this dataset. Our performance is best when no projections are used, but this should not be surprising given the nature of the dataset. Recall that the projections are designed to put features from radically different poses in the same latent space. When the dataset mostly contains one pose, the projections are essentially superfluous. Furthermore, the projections work better with classification than regression because they are trained to discriminate on a class-by-class basis. For example, if a projection is trained with a 13-19 year old class, features from a thirteen year old may end up closer to features from a nineteen year old than from those of a twelve year old. This works well when the same age classes are used for testing via classification, but it is not ideally suited for regression.

4.4. Images of Groups results

We present results for two sets of experiments on the IoG dataset. The first set of experiments compares our results to the best published results on the full IoG dataset. The second set of experiments demonstrate that our projections introduce pose-insensitivity to the age estimation process.

For each of the experiments, we evaluate three variations of our proposed model. The first variation uses only one pose bin for MvDA projections. In this case, MvDA reduces to FDA. The second variation uses two pose bins (one for poses between 0° and 15°, and the other for poses greater than 15°). The third variation takes the projections from the first two and concatenates them together, to make a double-length feature vector. Unfortunately, the limited number of off-angle images makes it impossible to learn successful projections with more pose bins.

4.4.1 Primary results

Results for all three variations of our model are shown in Table 2 with other published results. A confusion matrix for results with 1&2 pose bins is shown in Table 3. Note that we report a 7% improvement over the best published results [32]. The results are shown in more detail in Figure 3. Figures 3(b)-(d) report results for the portion of the test set with images in the 0° to 10°, 10° to 20°, and 20° to 30° pose bins, respectively.

There are a number of observations we wish to make about this figure. First, notice that the performance for all three variations of our model perform similarly for test images in the 0° to 10° pose bin (Figure 3(b)). In the 10° to 20° pose bin (Figure 3(c)) the "2 poses" variation lags a bit behind the other two variations, but the difference is not dramatic. Things get interesting in the 20° to 30° (Figure 3(d)). Here, we see the variations with multiple pose views with a clear advantage in four of the seven age bins, while the "1 pose" variation dominates in three bins. It is perhaps unsurprising that two of these three age bins are those with the fewest off-angle images to train upon. It seems safe to assume that a more balanced dataset would see reliably superior results from models with multiple pose views.

Approach	rank 1	rank 2
Appearance [9]	38.3%	71.3%
Appearance + Context [9]	42.9%	78.1%
Gabor + Adaboost [23]	43.7%	80.7%
LBP + Adaboost [23]	44.9%	83.0%
boosted Gabor + SVM [23]	48.4%	84.4%
boosted LBP + SVM [23]	50.3%	87.1%
LBP variants [32]	51.7%	88.7%
Ours, 1 pose bin	58.3%	92.5%
Ours, 2 pose bins	57.1%	91.5%
Ours, 1&2 pose bins	**58.7%**	**92.6%**

Table 2. Results on the Images of Groups dataset. Training and testing images were randomly selected from dataset following the procedure described in [9]. Results are an average of 10 trials. "rank 1" results are for correct age group classification. "rank 2" results allow for an error of one age group.

		Algorithm Estimate						
		0-2	4-7	8-12	13-19	20-36	37-65	66+
Human Estimate	0-2	82.7	15.6	0.7	0.3	0.3	0.3	0.0
	4-7	13.1	60.9	20.7	3.9	0.6	0.6	0.2
	8-12	0.7	27.0	46.6	16.9	5.4	2.8	0.7
	13-19	0.2	2.9	16.5	50.7	22.6	5.9	1.1
	20-36	0.3	0.7	3.5	23.5	48.1	22.1	1.9
	37-65	0.1	0.7	2.6	8.0	22.9	46.2	19.5
	66+	0.1	0.5	0.8	1.1	2.7	16.0	78.7

Table 3. Confusion matrix of main results (1&2 pose bins).

4.4.2 Results with restricted training

To demonstrate that MvDA projections allow our model to better generalize across pose, we restrict the training set to images with a pose of 3° or less. For the three model variations we have discussed thus far, it is still necessary to learn MvDA projections using training data across the entire range of poses, but the final classifier is only shown features from nearly frontal training images. We also report results for a new variation of our model that does not have *any* exposure to age information for images with off-angle poses, as described below.

Figure 4(a) shows results for our model with two additional variations. "1 pose (ID)" only uses one projection, as with "1 pose (age)," but the MvDA projection was learned differently. The projection was learned from the Labeled Faces in the Wild, MUCT, and PubFig datasets using subject ID as a class. None of the images in the IoG dataset were used to learn the projection, and the projection is completely agnostic with respect to age (the three datasets don't even have age annotations available). The final variation does not use any MvDA projection at all, and simply uses the feature vectors from the PCA described in Section 3.4.

In Figure 4(a), we see that our first three model variations perform best on test images with pose less than 10°, but the "1 pose (ID)" variation is the highest performer on test images with pose greater than 30°. This may be somewhat surprising, since the first three model variations use projections that were learned using age-binned data at all angles, but it is actually just a reminder that the dataset does not provide enough off-angle images to provide reliable age-based projections at the more extreme angles.

We attempt to mitigate the effects of the dataset's imbalance by eliminating the two most poorly represented age bins (8-12 and 66+) from the testing set, with results shown in Figure 4(b). Only the testing set is altered—the classifiers are the same as in Figure 4(a). Here, we see the variations with multiple projections performing the best, while accuracy of the variation with no projection declines significantly as the test images get further away from frontal.

To more clearly show how the performance of each variation changes from more frontal images to more off-angle images, Table 4 shows the percentage decrease in performance from the 0° - 10° pose bin to the 20° - 30° pose bin. The first column of results corresponds to Figure 4(a). The "1 pose bin (ID)" variation shows the smallest performance drop of less than 20%. The second column of results corresponds to Figure 4(b). Notice that the variation with no projection shows a decrease in accuracy of nearly 50%, while both the variations with projections based on more than one pose bin show a decline of less than 20%. We were somewhat surprised by the difference between the two columns for the "1 pose bin (ID)" variation, since its projections do not rely on the IoG dataset. It may be due to a bias in the classifier.

	% decrease from 0°-10° to 20°-30°	
Model	test all ages	without 8-12 and 66+
1 pose bin (age)	23.6%	24.2%
2 pose bins (age)	30.9%	**18.9**%
1&2 pose bins (age)	26.4%	19.5%
1 pose bin (ID)	**19.3**%	29.6%
no projection	32.0%	47.4%

Table 4. Decrease in performance from frontal pose to off-angle when trained only on the 0° to 3° pose range. Smaller is better.

5. Conclusions

We have introduced a novel approach to human age estimation from a single uncontrolled image. By projecting image features into a pose-invariant latent space, we introduce insensitivity to camera angles, which leads to increased accuracy on wild images. We have shown that our algorithm outperforms the best published results on a challenging dataset, is competitive on another dataset, and experimentally verified the pose-insensitivity of our model. Our model is by no means constrained to age estimation. In our next effort we will apply the same model to estimating gender, ethnicity, and expression in wild images.

Notice: This manuscript has been authored by UT-Battelle, LLC, under Contract No. DE-AC05-00OR22725 with the U.S. Department of Energy. The United States Government retains and the publisher, by accepting the article for publication, acknowledges that the United States Government retains a non-exclusive, paid-up, irrevocable, world-wide license to publish or reproduce the published form of this manuscript, or allow others to do so, for United States Government purposes.

References

[1] T. Ahonen, A. Hadid, and M. Pietikainen. Face description with local binary patterns: Application to face recognition. *PAMI*, 28(12), 2006.

[2] F. Alnajar, C. Shan, T. Gevers, and J.-M. Geusebroek. Learning-based encoding with soft assignment for age estimation under unconstrained imaging conditions. *Image and Vision Computing*, 30(12):946–953, 2012.

[3] P. N. Belhumeur, D. W. Jacobs, D. J. Kriegman, and N. Kumar. Localizing parts of faces using a consensus of exemplars. In *CVPR*, 2011.

[4] C.-C. Chang and C.-J. Lin. LIBSVM: A library for support vector machines. *ACM Transactions on Intelligent Systems and Technology*, 2, 2011. Software available at `http://www.csie.ntu.edu.tw/~cjlin/libsvm`.

[5] C. Chen, Y. Chang, K. Ricanek, and Y. Wang. Face age estimation using model selection. In *CVPR*, 2010.

[6] S. E. Choi, Y. J. Lee, S. J. Lee, K. R. Park, and J. Kim. Age estimation using a hierarchical classifier based on global and local facial features. *Pattern Recognition*, 44(6), 2011.

[7] Y. Fu and T. S. Huang. Human age estimation with regression on discriminative aging manifold. *Multimedia, IEEE Transactions on*, 10(4), 2008.

[8] Y. Fu, Y. Xu, and T. S. Huang. Estimating human age by manifold analysis of face pictures and regression on aging features. In *Multimedia and Expo*, 2007.

[9] A. C. Gallagher and T. Chen. Understanding images of groups of people. In *CVPR*, 2009.

[10] X. Geng, Z.-H. Zhou, and K. Smith-Miles. Automatic age estimation based on facial aging patterns. *PAMI*, 2007.

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

[12] G. Guo, G. Mu, Y. Fu, and T. S. Huang. Human age estimation using bio-inspired features. In *CVPR*, 2009.

[13] R. M. Haralick, K. Shanmugam, and I. H. Dinstein. Textural features for image classification. *Systems, Man and Cybernetics, IEEE Transactions on*, (6), 1973.

[14] G. B. Huang, M. Mattar, T. Berg, and E. Learned-Miller. E.: Labeled faces in the wild: A database for studying face recognition in unconstrained environments. 2007.

[15] M. Kan, S. Shan, H. Zhang, S. Lao, and X. Chen. Multi-view discriminant analysis. In *ECCV*. 2012.

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

[17] Y. H. Kwon and N. da Vitoria Lobo. Age classification from facial images. *CVIU*, 74(1), 1999.

[18] Z. Li, Y. Fu, and T. S. Huang. A robust framework for multiview age estimation. In *CVPRW*, pages 9–16. IEEE, 2010.

[19] K. Luu, K. Ricanek, T. D. Bui, and C. Y. Suen. Age estimation using active appearance models and support vector machine regression. In *BTAS*, 2009.

[20] M. C. Nechyba, L. Brandy, and H. Schneiderman. Pittpatt face detection and tracking for the clear 2007 evaluation. In *Multimodal Technologies for Perception of Humans*. Springer, 2008.

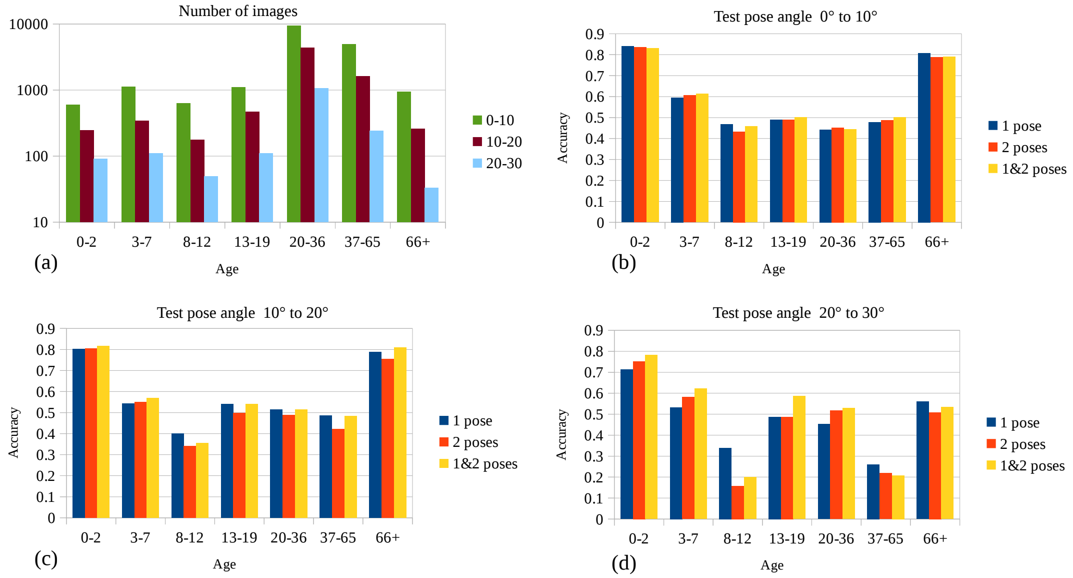

Figure 3. (a) Number of images in the Images of Groups dataset by age and pose. Notice that there are very few images in the 8-12 and 66+ bins for poses between 20° and 30°. (b) Results for test images with a head pose between 0° and 10°. (c) Results for test images between 10° and 20°. (d) Results for test images between 20° and 30°. Note that the models with more than one pose tend to perform more poorly than the model with one pose in the 8-12 and 66+ age bins.

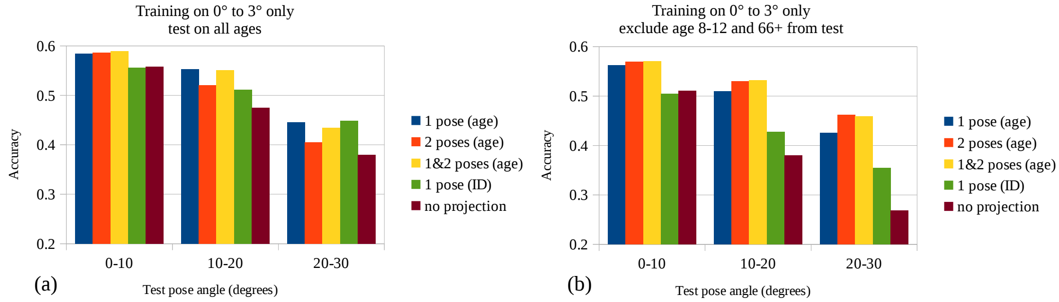

Figure 4. (a) Results when trained only on images with a pose of 0° to 3°. Models include the same models shown in Figure 3, with two additional models as described in Section 4.4.2. (b) Results for the same experiment, but with the age bins 8-12 and 66+ removed from the test set. These are the bins with the fewest off-angle examples in the Images of Groups dataset.

[21] P. J. Phillips, W. T. Scruggs, A. J. OToole, P. J. Flynn, K. W. Bowyer, C. L. Schott, and M. Sharpe. Frvt 2006 and ice 2006 large-scale results. *NISTIR*, 7408, 2007.

[22] N. Ramanathan and R. Chellappa. Modeling age progression in young faces. In *CVPR*, 2006.

[23] C. Shan. Learning local features for age estimation on real-life faces. In *ACM international workshop on Multimodal pervasive video analysis*, 2010.

[24] A. Sharma, A. Kumar, H. Daume, and D. W. Jacobs. Generalized multiview analysis: A discriminative latent space. In *CVPR*, 2012.

[25] J. Suo, S.-C. Zhu, S. Shan, and X. Chen. A compositional and dynamic model for face aging. *PAMI*, 2010.

[26] Z. Xu, H. Chen, S.-C. Zhu, and J. Luo. A hierarchical compositional model for face representation and sketching. *PAMI*, 2008.

[27] S. Yan, H. Wang, T. S. Huang, Q. Yang, and X. Tang. Ranking with uncertain labels. In *Multimedia and Expo*, 2007.

[28] S. Yan, H. Wang, X. Tang, and T. S. Huang. Learning auto-structured regressor from uncertain nonnegative labels. In *ICCV*, 2007.

[29] S. Yan, X. Zhou, M. Liu, M. Hasegawa-Johnson, and T. S. Huang. Regression from patch-kernel. In *CVPR*, 2008.

[30] W. Yang, C. Chen, K. Ricanek, and C. Sun. Ensemble of global and local features for face age estimation. In *Advances in Neural Networks*. Springer, 2011.

[31] J. Ylioinas, A. Hadid, X. Hong, and M. Pietikäinen. Age estimation using local binary pattern kernel density estimate. In *ICIAP*. 2013.

[32] J. Ylioinas, A. Hadid, and M. Pietikainen. Age classification in unconstrained conditions using lbp variants. In *ICPR*, 2012.