Regression from Patch-Kernel

Shuicheng Yan[1], Xi Zhou[2], MingLiu[2], Mark Hasegawa-Johnson[2], and Thomas S. Huang[2]
[1] Department of Electrical and Computer Engineering, National University of Singapore, Singapore
[2] Department of Electrical and Computer Engineering, University of Illinois at Urbana-Champaign, USA

Abstract

In this paper, we present a patch-based regression framework for addressing the human age and head pose estimation problems. Firstly, each image is encoded as an ensemble of orderless coordinate patches*, the global distribution of which is described by Gaussian Mixture Models (GMM), and then each image is further expressed as a specific distribution model by Maximum a Posteriori adaptation from the global GMM. Then the* patch-kernel *is designed for characterizing the Kullback-Leibler divergence between the derived models for any two images, and its discriminating power is further enhanced by a weak learning process, called* inter-modality similarity synchronization*. Finally,* kernel regression *is employed for ultimate human age or head pose estimation. These three stages are complementary to each other, and jointly minimize the regression error. The effectiveness of this regression framework is validated by three experiments: 1) on the YAMAHA aging database, our solution brings a more than 50% reduction in age estimation error compared with the best reported results; 2) on the FG-NET aging database, our solution based on raw image features performs even better than the state-of-the-art algorithms which require fine face alignment for extracting warped appearance features; and 3) on the CHIL head pose database, our solution significantly outperforms the best one reported in the CLEAR07 evaluation.*

1. Introduction

A face image may encode many human characteristics, *e.g.*, identity, expression, gender, ethnicity, age, and pose. Age and pose information are useful in many applications, but are less well-studied than the other identifiable characteristics of an image, perhaps because both age and pose are best represented as continuous rather than discrete hidden variables [7] [8] [12]. Human age estimation can provide useful information for electronic consumer relations management and demographic data collection, and head pose estimation has broad applications including gaze detection, driving safety, and auto-mouse on large screens. Moreover, currently the main challenges for practical face recognition systems come from human age and head pose variations.

Geng *et al.* [6] [5] proposed to conduct age estimation by modeling the statistical properties of the aging pattern, namely a sequence of personal facial age images, based on the assumption that multiple images of different ages are available for each person. Recently, Yan *et al.* [19] proposed an algorithm based on semi-definite programming for age estimation, with allowance made for uncertainty in the reference age labels. The pose estimation problem has also attracted much attention [3] [12] [13] in recent years owing to its great potential in practical systems.

Most previous algorithms for these two tasks are based on holistic image features, but holistic features are sensitive to illumination variations and image occlusions. Lucey *et al.* [9] demonstrated that face verification may benefit from the free-patch based representation, which has the potential to overcome these issues. The human age and head pose problems are, however, beyond the solution from [9] for two reasons. First, the free-patch representation discards the coordinate information, which has been proven to be necessary for accurate pose estimation [4]. Second, the algorithm in [9] is limited in addressing classification problems rather than regression problems, and its discriminating power may be greatly degraded if large within-class variations exist. Thus, it is desirable to have a discriminative and robust patch-based framework for general visual regression tasks.

In this work, we present a general patch-based framework for addressing visual regression problems, *e.g.*, human age estimation and head pose estimation. First, each image is encoded as an ensemble of overlapped coordinate patches, each of which integrates coordinate information together with the features extracted by the 2D discrete cosine transform. The global distribution of these patches is modeled by a Gaussian Mixture Model (GMM). Then, the patch-kernel for measuring image similarity is derived by representing each image as a patch distribution, which is adapted from the global GMM using *Maximum a Posteriori* adaptation. To further enhance the discriminating power of patch-kernel, a weak learning process, called

978-1-4244-2243-2/08/$25.00 ©2008 IEEE

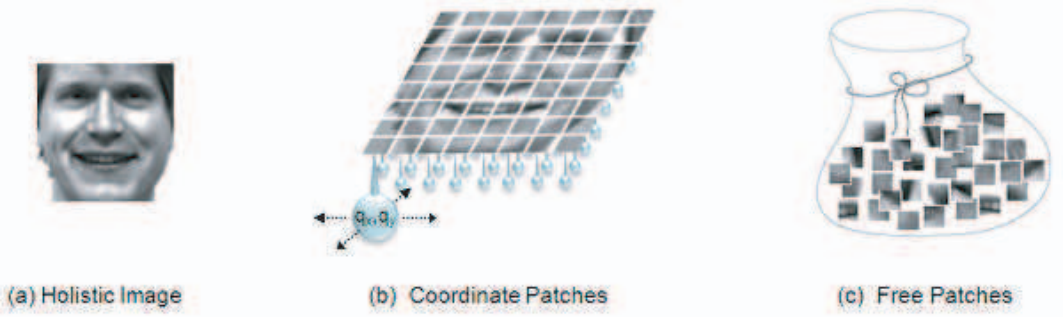


Figure 1. An illustration of the three image representations: a) holistic image, where an appearance feature is assigned for each fixed coordinate; b) coordinate patches, where certain appearance features may appear in a flexible area, and the attached ball for each local patch means that the coordinate of the patch is changeable; and c) free patches, where coordinate information is discarded entirely [9].

Inter-Modality Similarity Synchronization, is proposed by removing the kernel components with high-variability for data with similar labels. Finally, the kernel regression parameters are learned by a Least Squares Error approach based on the enhanced patch-kernel and a set of reference images or image sets.

2. Image as Ensemble of Coordinate Patches

In this work, we address the human pose and age estimation problems. The image set for model training is denoted as a matrix $X = [x_1, x_2, \ldots, x_N], x_i \in \mathbb{R}^m$, where N is the image number and m is the dimension of each feature vector. The human age or head pose label for an image x_i is denoted as l_i, where $l_i \in R$. The task is to predict the human age or head pose of any new image x based on the knowledge from the training images X and their labels. The label l_i can represent continuous (real) values, and thus the human age and head pose estimation problems are essentially regression problems rather than classification problems.

Inspired by the recent progress [14] [1] in speech/speaker recognition research, we present a novel framework for visual regression tasks. The whole framework consists of three components, namely, coordinate patch based image representation, patch-kernel design and enhancement by inter-modality similarity synchronization, and kernel regression. We shall introduce the image as an ensemble of coordinate patches in this section, and the latter two components are introduced in the next section.

2.1. Coordinate Patches

Most previous algorithms for human age and head pose estimation are based on holistic image features, and hence are sensitive to illumination variations and image occlusions. In contrast to a holistic image representation, a patch-based image representation has the potential to overcome these limitations [9] [10].

In this work, we introduce a local descriptor for image representation called the coordinate patch. Lucey *et al.* [9] proposed to encode each image as an ensemble of free patches, containing no information about patch coordinates. Unlike free patches, coordinate patches integrate both appearance information and coordinate information, in order to provide a local representation that is informative about the holistic structure of the image. For example, the combination mode of appearance and coordinate information within one coordinate patch can provide useful information for pose estimation and expression recognition. For a position within the image plane, denoted as $q = (q_x, q_y)^T$, its corresponding coordinate patch for a given image x_i is defined as

$$Q(x_i, q) = \begin{bmatrix} R(x_i, q) \\ q \end{bmatrix}, \tag{1}$$

where $R(x_i, q)$ denotes the feature vector extracted from the image x_i within the rectangle centered at the position q. In this work, to compute $R(x_i, q)$, we first remove the mean of the intensity values within the rectangle, then normalize the intensities to unit variance, and finally use the 2D discrete cosine transform to extract the final feature vector $R(x_i, q)$. Thus the coordinate patch is relatively robust to illumination variations.

Discussion: What are the advantages of encoding an image as an ensemble of coordinate patches for human age and head pose estimation? Age information is often embodied by local information, e.g., wrinkles around the eye corners. The positions of these informative areas are transformable due to the shape differences among different subjects, but they are not totally unconstrained. The proposed coordinate patch naturally has the potential to model the phenomena that patches with a certain characteristic (wrinkle-like) appear in certain geometric regions. On the other hand, for the head position estimation task, different combinations of local appearance information (nose-tip) and coordinate information can indicate head pose. Although free-patch does not have the above advantages, the coordinate patch inherits its merits in terms of robustness to image occlusions as validated in the experiment section.

2.2. Coordinate Patch Distribution Model

A coordinate patch describes the correlation existing between the local appearance information and coordinate information. Modeling the distribution of coordinate patches is equivalent to modeling the possible modes of dependency between these two modalities of information. Due to the large variations of these modes, we model the distribution of coordinate patches from each image using a Gaussian Mixture Model (GMM), because the GMM can approximate any distribution when the number of Gaussian components is sufficient. However, the number of coordinate patches from one image is small and insufficient for robustly estimating a GMM of moderate scale. Therefore, we first estimate a global GMM which does not consider the human age or head pose information and is derived based on coordinate patches from all training images. It is similar to the so-called Universal Background Model (UBM) in speech/speaker verification [14]. Then the distribution of the coordinate patches belonging to one image is adapted from the global GMM by *Maximum a Posteriori* (MAP) [16]. The philosophy of unsupervised global model followed by MAP adaptation with a small number of labeled examples can effectively overcome the small sample size issue suffered in conventional learning processes.

For ease of presentation, here we denote z as the combined local appearance information and coordinate information of a coordinate patch, namely, $z = Q(x_i, q)$. Then, the distribution of the variable z is

$$p(z;\Theta) = \sum_{k=1}^{K} w_k \mathcal{N}(z;\mu_k,\Sigma_k), \tag{2}$$

where w_k, μ_k and Σ_k are the weight, mean and covariance matrix of the kth Gaussian component, respectively, and K is the total number of Gaussian components.

The density is a weighted linear combination of K unimodal Gaussian densities, namely,

$$\mathcal{N}(z;\mu_k,\Sigma_k) = \frac{1}{(2\pi)^{\frac{d}{2}}|\Sigma_k|^{\frac{1}{2}}} e^{-\frac{1}{2}(z-\mu_k)^T\Sigma_k^{-1}(z-\mu_k)}. \tag{3}$$

Many approaches can be proposed to estimate the model parameters. Here we obtain a maximum likelihood parameter set using the Expectation-Maximization (EM) algorithm. For computational efficiency, the covariance matrices are restricted to be diagonal [1] as conventionally.

2.3. Image-Specific Model via MAP

We derive the image-specific coordinate-patch distribution model by adapting the mean vectors of the global GMM and retaining the mixture weights and covariance matrices.

Mean vectors are adapted using MAP adaptation [16] with conjugate priors [16], thus the parameters $\hat{\mu}_k$ are selected to maximize

$$\begin{aligned}\ln p(\hat{\theta}, Z) &= \sum_{k=1}^{K} \ln \mathcal{N}(\hat{\mu}_k;\mu_k,\Sigma_k/r) \\ &\quad + \sum_{i=1}^{H} \ln \sum_{k=1}^{K} w_k \mathcal{N}(z_i;\hat{\mu}_k,\Sigma_k), \end{aligned} \tag{4}$$

where $\hat{\theta} = \{\hat{\mu}_1,\ldots,\hat{\mu}_K\}$ is the set of image-dependent parameters, $\theta = \{w_1,\mu_1,\Sigma_1,\ldots\}$ are the parameters of the universal background model, and $Z = \{z_1,\ldots,z_H\}$ are the patches of the image being modeled. As shown, the conjugate prior for parameter $\hat{\mu}_k$ is itself Gaussian ($\mathcal{N}(\hat{\mu}_k;\mu_k,\Sigma_k/r)$), with a covariance matrix shrunk by smoothing parameter r. The joint distribution function $p(\hat{\theta}, Z)$ has the same form as the likelihood function $p(Z|\hat{\theta})$, and may therefore be maximized in the same way as a likelihood function, *i.e.*, using EM with the hidden variable $Pr(k|z_i)$ as the posterior probability of Gaussian component k given patch z_i [16].

So in the E-step, we compute the posterior probability:

$$Pr(k|z_i) = \frac{w_k \mathcal{N}(z_i;\mu_k,\Sigma_k)}{\sum_{j=1}^{M} w_j \mathcal{N}(z_i;\mu_j,\Sigma_j)}, \tag{5}$$

$$n_k = \sum_{i=1}^{H} Pr(k|z_i), \tag{6}$$

and then the M-step updates the mean vectors, namely

$$E_k(Z) = \frac{1}{n_k}\sum_{i=1}^{H} Pr(k|z_i) z_i, \tag{7}$$

$$\hat{\mu}_k = \alpha_k E_k(z) + (1-\alpha_k)\mu_k, \tag{8}$$

where $\alpha_k = n_k/(n_k + r)$. MAP adaptation using conjugate priors is useful because it interpolates, smoothly, between the hyper-parameters μ_k and the maximum likelihood parameters $E_k(Z)$. If a Gaussian component has a high probabilistic count, n_k, then α_k approaches 1 and the adapted parameters emphasize the new sufficient statistics; otherwise, the adapted parameters are determined by the global model. In this work, r is adjusted, empirically, depending on the total number of coordinate patches for each image.

3. Regression from Patch-Kernel

In this section, the derived GMM-based coordinate patch distribution for each image is used to construct the patch-kernel for measuring image similarity. Then, discriminating power is this kernel is further enhanced by discarding the kernel components with high-variability for data with similar labels, and consequently the similarity within the feature

space characterized by patch-kernel will be better "synchronized" with the similarity between the target labels; we call this method Inter-Modality Similarity Synchronization. Finally, the resulting similarity-synchronized patch-kernel is used for kernel regression to predict the target age or pose labels.

3.1. Patch-kernel: Image as a Distribution

Suppose we have two face images x_a and x_b, with the coordinate patch sets Z_a and Z_b respectively. Then, from the GMM adaptation process in (5-8), we can obtain two adapted GMMs for them, denoted as g_a and g_b. Consequently, each face image is represented by a specific GMM distribution model, and a natural similarity measure between them is the Kullback-Leibler divergence,

$$D(g_a||g_b) = \int g_a(z) log\left(\frac{g_a(z)}{g_b(z)}\right) dz. \quad (9)$$

The Kullback-Leibler divergence itself does not satisfy the conditions for a kernel function, but there exists an upper bound from the log-sum inequality,

$$D(g_a||g_b) \leq \sum_{k=1}^{K} w_k D(\mathcal{N}(z;\mu_k^a,\Sigma_k)||\mathcal{N}(z;\mu_k^b,\Sigma_k)),$$

where μ_k^a denotes the adapted mean of the kth component from image x_a, and likewise for μ_k^b. Based on the assumption that the covariance matrices are unchanged during the MAP adaptation process, the right side of the above inequality is equal to

$$d(x_a,x_b) = \frac{1}{2}\sum_{k=1}^{K} w_k(\mu_k^a-\mu_k^b)^T\Sigma_k^{-1}(\mu_k^a-\mu_k^b). \quad (10)$$

It is easy to prove that $d(x_a,x_b)$ is a metric function, and therefore we can define the following kernel function

$$k(x_a,x_b) = e^{-d(x_a,x_b)/\delta_1^2}, \quad (11)$$

where δ_1 is a constant for controlling the final similarity. $k(x_a,x_b)$ can be considered as a conventional Gaussian kernel defined on the so-called supervector,

$$\phi(x_a) = [\sqrt{\frac{w_1}{2}}\Sigma_1^{-\frac{1}{2}}\mu_1^a;\cdots;\sqrt{\frac{w_K}{2}}\Sigma_K^{-\frac{1}{2}}\mu_K^a], \quad (12)$$

and then $d(x_a,x_b) = \|\phi(x_a)-\phi(x_b)\|^2$. This kernel function is derived by encoding each image as an ensemble of coordinate patches, and is hence called a *patch-kernel* in this work. Note that here we can also use image-specific weight w_i^a for $\phi(x_a)$ calculation, where w_i^a is adapted from w_i by MAP [14].

3.2. Synchronize Inter-Modality Similarity

The patch-kernel is derived from the generative GMM and does not consider inter-class or intra-class relationships; hence it does not necessarily provide good discriminating power. More specifically, the supervector $\phi(x_a)$ is computed directly from the image x_a by adapting the global GMM, and hence is not ensured to be close to the supervectors computed from images with similar ages or poses.

In this subsection, we present a weak learning process for enhancing the kernel discriminating power. More specifically, we want the patch-kernel computed using a pair of images with similar ages or poses to have a large value, while the patch-kernel computed using images with greatly different ages or poses should have a small value. In this way the similarities measured in the feature modality and label modality are synchronized, hence we call this process Inter-Modality Similarity Synchronization. A natural way to achieve this goal is to remove any patch-kernel components in which supervectors corresponding to similar labels (similar ages or similar poses) are spread out over a wide range of values (high-variability). These directions are assumed in this work to be characterized by a subspace spanned by the projection matrix V. In order to identify V, we first define the label-similarity matrix W as

$$W_{ij} = e^{-||l_i-l_j||^2/\delta_2^2}, \quad (13)$$

which measures the label similarity between image x_i and image x_j, using hyper-parameter δ_2 to control the scale over which label similarities are distinguished.

The goal of inter-modality similarity synchronization is to identify the subspace, V, that has maximum inter-image distance (maximum $\|V^T\phi(x_i)-V^T\phi(x_j)\|^2$) for image pair with high label similarity (large W_{ij}). Expressing this goal in the form of an optimality criterion, we find that

$$V = \arg\max_{V^TV=I}\sum_{i\neq j}||V^T\phi(x_i)-V^T\phi(x_j)||^2W_{ij}. \quad (14)$$

Denote $\hat{X} = [\phi(x_1),\phi(x_2),\cdots,\phi(x_N)]$, then the optimal V consists of the eigenvectors corresponding to the top few largest eigenvalues of the matrix $\hat{X}(D-W)\hat{X}^T$, where D is a diagonal matrix with $D_{ii} = \sum_{j=1}^{N} W_{ij}, \forall i$.

V identifies the components in which feature similarity and label similarity are most out of sync (high label similarity corresponds to low feature similarity, and vice versa). In order to achieve inter-modality similarity synchronization, we must discard the components $V^\phi(x_i)$ prior to computing the similarity between any two images. It is possible to define a similarity-synchronized distance metric, $d(x_a,x_b)$, as

$$d(x_a,x_b) = (\phi(x_a)-\phi(x_b))^T(I-VV^T)(\phi(x_a)-\phi(x_b)), \quad (15)$$

where we have taken advantage of the equality $(I - VV^T)(I - VV^T) = (I - VV^T)$.

Note that the patch-kernel is applicable not only for image pairs. If any object can be characterized by a coordinate patch set Z, then we can adapt the global GMM to a new one by the process in (5-8); thus we can compute the kernel similarity between an image and an image set, or between two images with missing patches.

3.3. Kernel Regression

Kernel regression [11] is a non-parametric technique in statistics to estimate the conditional expectation of a random variable. In this work, we generalize this model and set the expected values of the reference points as model parameters to be determined. In kernel regression, a set of reference points is required for learning the model. We evenly divide the label field into multiple subsets, and then for each subset with training images denoted as X_m, we can compute the similarity, denoted as $k(x, X_m)$, between an image x and the image set X_m. Then the kernel regression model is expressed as

$$F(x) = \frac{\sum_{m=1}^{M} \beta_m k(x, X_m)}{\sum_{m=1}^{M} k(x, X_m)}, \tag{16}$$

where M is the number of reference subsets, and the parameters β_m can be easily derived by using the Least Squares Error method based on the training images. For a new image, its age or pose label can be directly computed from (16).

Discussion: There exist many popular algorithms for regression, *e.g.*, linear regression and neural networks [15]. In this work, we choose kernel regression because the patch-kernel itself provides reasonably good similarity measurement, and then the term $k(x, X_m)$ will have a large value if x is within the age or pose range in which X_m lies, which coincides with the philosophy of kernel regression. The result of learning is a set of kernel regression coefficients that approximately equal the label means of the reference image sets X_m.

4. Experiments

In this section, we systematically evaluate the effectiveness of the **r**egression framework from the **p**atch **k**ernel (RPK), and compare RPK with the state-of-the-art algorithms for human age and head pose estimation. The human age estimation experiments are conducted on the YAMAHA aging database[1] and the FG-NET [20] aging database. The head pose estimation experiments are conducted on the CHIL data used for CLEAR07 evaluation [21].

[1]To protect the portrait rights of the participants, sample images of the YAMAHA face database are not shown here.

Figure 2. Sample images of one person in the FG-NET database.

4.1. Data Sets and Experimental Setups

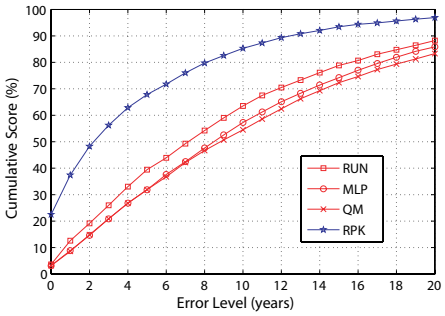
4.1.1 Human Aging Databases

The YAMAHA aging database contains 8000 Japanese facial images of 1600 persons with ages ranging from 0 to 93. Each person has 5 images and the YAMAHA database is divided into two subsets with 4000 images from 800 males and another 4000 images from 800 females. Our experiments are carried out separately on female and male subsets. For each subset, 1000 images are randomly selected for model training while the remaining 3000 samples are used for testing, and the configurations of the training and testing sets are the same as in [19]. The FG-NET aging database [20] contains 1002 face images of 82 persons with ages ranging from 0 to 69. For both databases, the image is cropped and scaled to 32-by-32 pixels, and some example images of one person from the FG-NET database are depicted in Figure 2.

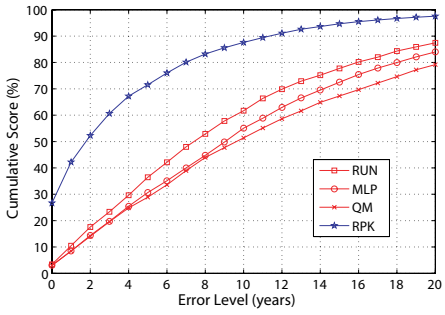
For comparison, the results from the traditional regression algorithms, Quadratic Models (QM) [8], Neural Networks [15], and the Nonlinear Regression with Uncertain Nonnegative Labels (RUN) algorithm [19], were used as baselines to evaluate the performance of our RPK framework. For the YAMAHA aging database, the latest results were obtained from the RUN algorithm as reported in [19], and for the FG-NET database the best results were reported in [18], where the evaluation protocol is Leave-One-Person-Out.

4.1.2 CHIL Head Pose Database

For the CHIL data in the CLEAR07 evaluation, each sample consists of four images captured by four cameras. In our experiments, we use the same experimental configuration as designed by the evaluation committee. For training, 10 videos are provided with the annotations of the head bounding boxes and the original ground truth information on three pose angles, namely, pan, tilt, and roll. For evaluation, 5 videos from 5 subjects are provided. In total, the training set contains 5348 samples (each consists of four images), and the testing set contains 2402 samples. Each image is cropped and scaled to the size of 24-by-24 pixels for our experiments.



(a) Female@Yamaha.



(b) Male@Yamaha.

Figure 3. Cumulative scores of human age estimation results for QM, MLP, RUN [19], and RPK algorithms at error levels from 0 to 20 years on the two configurations of the YAMAHA aging database. Note that the results for the first three algorithms are obtained from [19], which achieved the best reported results on this database.

Table 1. MAEs (year) of different algorithms on the YAMAHA database.

Female@Yamaha					Male@Yamaha				
Range	RPK	RUN [19]	QM [19]	MLP [19]	Range	RPK	RUN [19]	QM [19]	MLP [19]
0-9	1.83	11.21	11.97	14.33	0-9	1.61	9.86	13.42	14.08
10-19	3.93	6.23	9.58	8.85	10-19	3.62	7.52	10.33	9.46
20-29	5.27	7.95	9.29	9.70	20-29	4.65	8.85	10.21	9.35
30-39	6.73	8.17	9.85	9.66	30-39	5.62	7.76	9.35	8.60
40-49	6.73	8.64	10.45	8.78	40-49	5.06	8.67	11.71	9.10
50-59	5.37	9.43	10.15	9.53	50-59	5.12	11.10	13.38	10.08
60-69	4.39	11.12	13.49	10.88	60-69	3.66	12.49	15.99	13.44
70-93	5.22	15.56	19.66	16.52	70-93	5.73	16.60	20.44	19.69
Average	**4.94**	9.79	11.80	11.03	Average	**4.38**	10.36	13.10	11.72

Table 2. MAEs (year) of different algorithms on the FG-NET aging database. Note that *BM* below signifies the bilinear model used in [18].

Range	RPK	RUN [19]	QM [19]	MLP [19]
0-9	2.30	2.51	6.26	11.63
10-19	4.86	3.76	5.85	3.33
20-29	4.02	6.38	7.10	8.81
30-39	7.32	12.51	11.56	18.46
40-49	15.24	20.09	14.80	27.98
50-59	22.20	28.07	24.27	37.20
60-69	33.15	42.50	37.38	49.13
Average	**4.95**	5.78	7.57	10.39
BM [18]: 5.33 AGES [5]: 6.77 WAS [6]: 8.06				

4.1.3 Experimental Setups

In this work, we used two measures to evaluate algorithmic performance. The first one is the Mean Absolute Error (MAE) criterion as used in [6] [8] . MAE is defined as the average of the absolute errors between the estimated labels and ground truth labels, *i.e.*, $MAE = \sum_{i=1}^{N_t} |\hat{a}_i - a_i|/N_t$, where $\hat{a}_i$ is the estimated age or pose for the ith testing sample, a_i is the corresponding ground truth, and N_t is the total number of the testing samples. Another popular measure is the cumulative score [6] defined as: $CumScore(\theta) = N_{e\leq\theta}/N_t \times 100\%$, where $N_{e\leq\theta}$ is the number of samples on which the absolute errors are not higher than θ.

For these two aging databases, the patch size is set as 6-by-6 pixels, and for each image, the patches are densely sampled pixel by pixel within the image plane. The GMM contains 512 Gaussian components.

For the CHIL head pose database, there exist four images for each sample, hence the CHIL database is larger than the other two databases. To speed up the process, we train four GMMs for these four images respectively, and finally combine them to compute the supervectors. The patch size is set as 5-by-5 pixels, and each GMM contains 256 components. For all the experiments, the column number of matrix V, the number of reference image sets (M), the parameters δ_1 and δ_2 are fixed empirically.

4.2. Human Age Estimation Results

4.2.1 YAMAHA Aging Database

Figure 3 depicts the cumulative scores from RPK and the other three comparison algorithms, and Table 1 lists the detailed MAEs of these algorithms. Notice that:

1. The MAE of human age estimation is substantially reduced from 9.78 years (best reported result [19]) to 4.94 years for the female subset, and 10.36 years (best reported result [19]) to 4.38 years for the male subset. On average, an MAE reduction of more than 50% is achieved compared with the best results ever reported.

2. Our proposed patch-kernel based regression framework performs perfectly in the age range of $[0, 9]$. This is quite different from the behavior of the other

Table 3. MAEs (degree) of the algorithms PCA, LEA, SSE and RPK on the CHIL data from the CLEAR07 evaluation.

Pan Angle	Subject-1	Subject-2	Subject-3	Subject-4	Subject-5	Total Average
PCA	8.54	8.19	6.91	4.53	4.78	6.94
LEA	7.60	8.77	6.33	4.50	4.511	6.72
SSE	8.45	7.27	6.22	4.33	3.94	6.60
RPK	7.08	4.80	4.89	3.95	3.23	**4.96**
Tilt Angle	Subject-1	Subject-2	Subject-3	Subject-4	Subject-5	Total Average
PCA	8.49	5.97	11.59	5.25	12.53	10.86
LEA	7.88	5.74	12.29	5.29	12.23	10.87
SSE	8.61	6.28	9.08	4.92	9.64	8.25
RPK	6.14	4.99	7.72	4.08	14.09	**6.66**
Roll Angle	Subject-1	Subject-2	Subject-3	Subject-4	Subject-5	Total Average
PCA	4.66	2.59	4.20	2.86	3.30	4.01
LEA	5.41	2.59	4.06	2.90	2.91	4.07
SSE	5.55	2.22	3.72	2.38	2.34	3.42
RPK	4.51	2.38	3.24	2.16	2.57	**3.02**

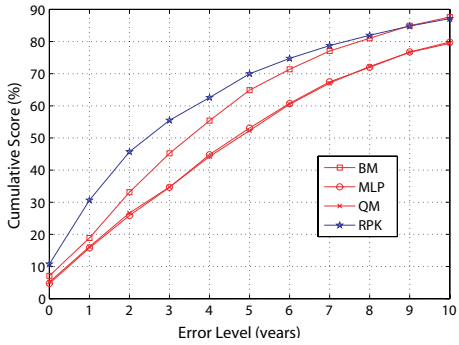


Figure 4. Cumulative scores of age estimation for the QM, MLP, BM [18], and RPK algorithms at error levels from 0 to 10 years on the FG-NET aging database. Note that the results for the first three algorithms are obtained from [18], which achieved the best reported results on this database.

three comparison algorithms, which perform particularly poorly in this age range.

4.2.2 FG-NET Aging Database

On the FG-NET database, all the conventional algorithms are based on warped appearance features [5]. First, 68 key facial points are labeled for each image, and then the shape, texture, and appearance models are trained based on all the samples. Finally the first 200 appearance parameters [5] from the appearance model are used to represent each face image. For detailed information on shape, texture, and appearance models, please refer to [2]. In practical systems, face alignment is still a tough problem, especially for the cases with pose and expression variations as in Figure 2.

RPK works directly on original raw image features without the requirement of face alignment. Figure 4 depicts the cumulative scores from RPK and the other comparison algorithms, and Table 2 lists the detailed MAEs of these algorithms. From these results, we make the following observations: 1) even without face alignment, RPK still outperforms state-of-the-art algorithms that require precise face alignment; and 2) the age range of $[0, 9]$ is again the one in which RPK has its best age estimation accuracy.

4.3. Head Pose Estimation Results

For comparison, we implemented Principal Components Analysis (PCA), Locally Embedded Analysis (LEA) [3], and Submanifold Synchronized Embedding (SSE) [17] which produced the best results as reported in the CLEAR07 evaluation. The detailed results on the three angles of head pose are listed in Table 3, from which we can observe that the RPK framework performs the best among all the algorithms evaluated. Note that we carefully tuned the parameters for SSE, and hence the results reported here for SSE are a little better than those originally reported in the CLEAR07 evaluation.

4.4. Algorithmic Analysis

In this subsection, we give an in-depth analysis of the effectiveness of the three components of the RPK framework, namely coordinate patch representation, inter-modality similarity synchronization, and kernel regression. Then we evaluate the algorithm's robustness to image occlusions.

4.4.1 Effectiveness of individual components of RPK

In this subsection, we evaluate the effectiveness of the individual components of RPK on the YAMAHA-Female subset. For each experiment, we remove one component of RPK, and conduct the regression based on the other two components. More specifically, when the coordinate patch is not used, we use the free-patch instead; and when the kernel regression component is removed, we predict the label of a new datum as the label mean of the nearest X_m. Detailed comparison results are listed in Figure 5 as confusion matrices, from which we can observe that: 1) the removal of any component degrades the overall performance; and 2) the inter-modality similarity synchronization component proves to be the most important in the whole framework.

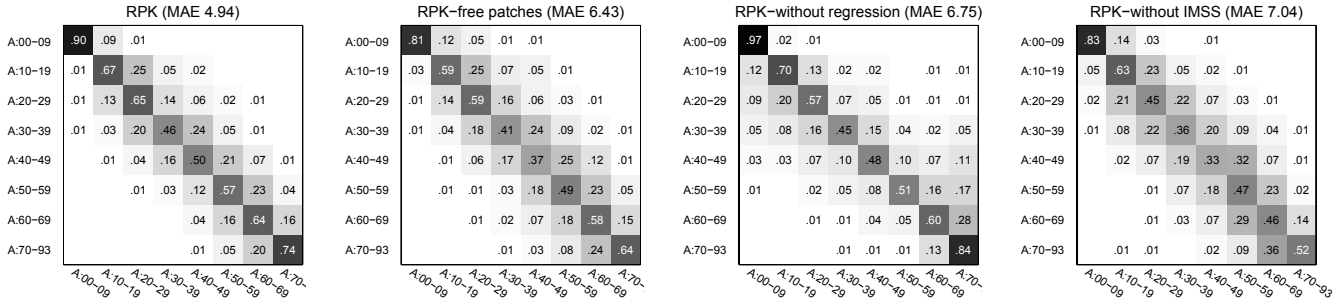


Figure 5. Comparison confusion matrices on the YAMAHA-Female subset for the original RPK, RPK with free-patches, RPK without kernel regression, and RPK without inter-modality similarity synchronization (IMSS). For better viewing, please see the pdf file.

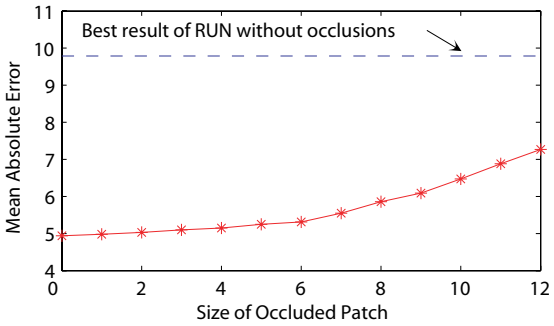


Figure 6. MAEs (year) of RPK with occlusions of different sizes. The blue dashed line denotes the best result reported in [19] without image occlusions.

Note that RPK without kernel regression achieves better group-based classification accuracy, but at the expense of higher variance of the regression output, hence the overall regression performance is much worse than that of RPK. Apparently the kernel regression component smooths the outputs from different reference sets.

4.4.2 Robustness to image occlusions

To demonstrate the algorithm's robustness to image occlusions, we systematically evaluate the performance of RPK on testing images with occluded patches of different sizes superimposed at random positions. Results are depicted in Figure 6. When the size of the occluded patch is not larger than 6-by-6 pixels, RPK is almost insensitive to image occlusions. When the patch size is 12-by-12 pixels, the performance of RPK is still much better than the best result reported in [19] for images without occlusion.

5. Conclusions and Future Directions

In this paper, we proposed a novel patch-based framework for visual regression problems and the whole framework consists of three mutually complementary components: coordinate patch representation, inter-modality similarity synchronization, and kernel regression. Some interesting future directions of this work include: 1) to extend this framework for other visual classification tasks, *e.g.*, gender recognition, expression recognition, and face recognition; 2) to adopt coordinate cube, instead of coordinate patch, for video event detection and recognition; and 3) to use the philosophy of UBM for multi-task learning.

Acknowledgment

This work was supported in part by US Government VACE Program and in part by AcRF Tier 1 Grant of R-263-000-464-112/133, Singapore.

References

[1] W. Campbell, D. Sturim, D. Reynolds, and A. Solomonoff. SVM Based Speaker Verification using a GMM Supervector Kernel and NAP Variability Compensation. *ICASSP*, vol. 1, pp. 97-100, 2006. 2, 3

[2] T. Cootes, G. Edwards, and C. Taylor. Active appearance models. *PAMI*, vol. 23, no, 6, pp. 681–685, 2001. 7

[3] Y. Fu, and T. Huang. Graph Embedded Analysis for Head Pose Estimation. *AFGR*, pp. 3-8, 2006. 1, 7

[4] A. Gee and R. Cipolla. Determining the gaze of faces in images. *Image and Vision Computing*, vol. 12, no. 10, pp. 639–647, 1994. 1

[5] X. Geng, Z. Zhou, and K. Smith-Miles. Automatic age estimation based on facial aging patterns. *PAMI*, vol.29, no.12, pp. 2234-2240, 2007. 1, 6, 7

[6] X. Geng, Z. Zhou, Y. Zhng, G. Li, and H. Dai. Learning from facial aging patterns for automatic age estimation. *ACM MM*, pp. 307–316, 2006. 1, 6

[7] Y. Kwon and N. Lobo. Age classification from facial images. *CVIU*, vol. 74, no. 1, pp. 1–21, 1999. 1

[8] A. Lanitis, C. Draganova, and C. Christodoulou. Comparing different classifiers for automatic age estimation. IEEE Transactions on Systems, Man and Cybernetics, Part B, vol. 34, no. 1, pp. 621–628, 2004. 1, 5, 6

[9] S. Lucey and T. Chen. A GMM Parts Based Face Representation for Improved Verification through Relevance Adaptation. *CVPR*, pp. 855-861, 2004. 1, 2

[10] S. Lucey and T. Chen, Learning Patch Dependencies for Improved Pose Mismatched Face Verification. *CVPR*, vol. 1, pp. 909-915, 2006. 2

[11] H. Takeda, S. Farsiu, and P. Milanfar. Kernel Regression for Image Processing and Reconstruction. *TIP*, vol. 16, no. 2, pp. 349-366, 2007. 5

[12] V. Nallure, J. Ye, and S. Panchanathan. Biased Manifold Embedding: A Framework for Person-Independent Head Pose Estimation. *CVPR*, 2007. 1

[13] B. Raytchev, I. Yoda, and K. Sakaue. Head pose estimation by nonlinear manifold learning. *ICPR*, pp. 23-26, 2004. 1

[14] D. Reynolds, T. Quatieri, and R. Dunn. Speaker Verification using Adapted Gaussian Mixture Models. *Digital Signal Processing*, vol. 10, no. 1-3, pp. 19-41, 2000. 2, 3, 4

[15] D. Rumelhart, G. Hinton and R. Williams. Learning representations by back-propagating errors. *Nature*, vol. 323, pp. 533–536, 1986. 5

[16] Chin-Hui Lee, Chih-Heng Lin and Bing-Hwang Juang. A study on speaker adaptation of the parameters of continuous density hidden Markov models. *tsap*, vol. 39, no. 4, pp. 806-814, 1991. 3

[17] S. Yan, Q. Zhen, Y. Fu, Y. Hu, J. Tu, and T. Huang. Learning a Person-Independent Representation for 3D Pose Estimation. *CLEAR07*, 2007. 7

[18] S. Yan, H. Wang, T. Huang, Q. Yang, and X. Tang. Xiaoou. Ranking with Uncertain Labels. *ICME*, pp. 96-99, 2007. 5, 6, 7

[19] S. Yan, H. Wang, T. Huang, and X. Tang. Auto-Structured Regressor from Uncertain Labels. *ICCV*, 2007. 1, 5, 6, 8

[20] FG-NET database:http://sting.cycollege.ac.cy/∼alan-itis/fgnetaging/. 5

[21] http://isl.ira.uka.de/clear07/?The_Evaluation. 5