Abstract

In this paper we propose a novel Contourlet Appearance Model (CAM) that is more accurate and faster at localizing facial landmarks than Active Appearance Models (AAMs). Our CAM also has the ability to not only extract holistic texture information, as AAMs do, but can also extract local texture information using the Nonsubsampled Contourlet Transform (NSCT). We demonstrate the efficiency of our method by applying it to the problem of facial age estimation. Compared to previously published age estimation techniques, our approach yields more accurate results when tested on various face aging databases.

1. Introduction

Face based age estimation studies have become prominent in recent years due to their potential role in enhancing facial recognition systems, in assisting human law enforcement etc. Human face age progression can be divided physically into two categories, i.e. growth and development (childhood to pre-adulthood), which primarily affects the physical structure of the craniofacial complex in the form of lengthening and widening of bony-tissue, and adult aging, where the primary impacts are in the formation of wrinkles, lines and soft tissue morphology. Active Appearance Models (AAMs) [3] are a commonly used tool in face age estimation [19] because they can successfully represent both shape that is prominent in the growth and development period, and texture that is prominent in the adult aging period.

1.1. Motivation

There are two main driving reasons for carrying out this work. Firstly, prior work on AAM implementations indicates that their ability to generalize to automatically fit facial landmarks on unseen images that exhibit illumination, pose or expression variation is quite poor [4]. Active Shape Models (ASMs) [5] have undergone several changes since the classical ASM was proposed in [6] and have recently become more robust [14, 20, 32] and hence more popular for accomplishing this task as they are not only faster, but also more accurate when dealing with unseen variations that are not present in training images.

Secondly, although AAMs are an excellent model based approach for face related problems [19, 24], they only represent holistic and not local information. This concern is something that local based approaches such as Local Binary Patterns (LBPs) [1], Gabor features etc., have addressed and hence their use in many applications has increased tremendously over the last few years. Compared to LBPs and Gabor features, the Nonsubsampled Contourlet Transform (NSCT) [8] can extract both local and directional information. Additionally, compared to LBPs that deal poorly with noisy images, NSCT can distinguish between the noise and the high frequency components, which contain edges and directional information, in an image.

This paper proposes a new Contourlet Appearance Model (CAM) that combines the robust facial landmarking capability of the Modified Active Shape Model (MASM) implementation [25] and NSCT features. Once the landmarking task is accomplished, both gray-scale pixels that represent the holistic features, prominent in the period of childhood, and Contourlet features that represent local features (wrinkles), that are outstanding especially in the adult period when faces are aging, are used to extract facial features from given images. These combined features are used as the discriminative facial inputs to a face aging function and are found to yield promising results on unseen images across two databases.

The rest of this paper is organized as follows. In section 2, we review prior work carried out on AAMs and age estimation. In section 3, we describe the Contourlet Transform and the NSCT. In section 4, we describe how our Contourlet Appearance Model is used to extract shape and texture information from facial images and go on to describe how we setup face aging functions for age estimation in section 5. Finally, we compare our approach against existing age estimation methods in section 6 and offer some conclusions on this work in section 7.
Figure 1. Sample ASM fitting results. The images in the first row are the initialization provided to the ASMs while the images in the second row show the corresponding fitting results under such initialization conditions. (a) An example of poor initialization, (b) Accurate initialization provided to the classical ASM implementation (c) Accurate initialization provided to MASM, (d) Fitting results produced by MASM under poor initialization conditions, (e) Fitting results produced by classical ASM under accurate initialization conditions, (f) Fitting results produced by MASM under accurate initialization conditions.

2. Background

In this section, we will first review prior work carried out on AAMs and will then describe previous studies carried out on age estimation.

2.1. Prior work on AAMs

AAMs were first presented by Cootes et al. in 1998 [3]. Since then, there have been numerous studies on possible improvements to them which have mainly focused on either improving their texture modeling or fitting stage.

In the texture modeling stage, it is necessary to deal with high-dimensional texture representations. When representing high resolution images, the twin problems of storage and computational requirements are often encountered. Therefore, two solutions were suggested to reduce redundancy in representation, i.e. image compression and local textures. In the first approach, wavelet based approaches were proposed to compress the texture to eliminate redundant pixels with small variances [29]. However, the compression ratio is still limited at the fitting stage. In the second approach, due to local textures around feature points being more important than others at the model fitting stage, Cristinacce and Cootes [7] suggested replacing the global textures by local ones to avoid dealing with high-dimensional texture. Kahraman et al. [17] used two independent subspaces, illumination and identity, to combat variations in illumination. Christoudias and Darrell [2] presented a nonlinear modeling approach combining a Gaussian Mixture Model (GMM) and a Nearest Neighbor Model (NNM) to model objects in nonlinear subspaces.

At the fitting stage, searching algorithms can suffer from local minima problems due to poor initial model parameters and as shown in Figure 1, initialization is extremely important to obtaining accurate fitting results. Therefore, the initial parameters have to be within a small radius of convergence. To solve this problem, Wimmer et al. [28] used ASMs to initialize the AAM fitting stage more precisely. Their approach yields both a large radius of convergence and accurate fitting. However, their method is still an AAM based fitting algorithm and as we have outlined in section 1.1, if ASM fitting results are good, we only need to locate facial landmarks using an ASM and then apply Principal Component Analysis (PCA) on the region bounded by their convex hull to determine the best AAM parameters without actually needing to use an AAM for fitting the image. This is a key idea in our approach, in which we use the MASM implementation, which has been shown to have a reasonably high fitting accuracy, to obtain the landmark locations. This is highlighted by Figure 1, which shows that MASM obtains more accurate fitting results than classical ASM under identical initialization conditions.

2.2. Prior work on age estimation

Age estimation techniques can be broadly classified under two headings which are described in the following paragraphs.

In holistic features based approaches, facial features are globally extracted from the entire facial region. Lanitis et al. [18] used AAM features with a Weighted Aging Function (WAS method) setup by a genetic algorithm to address the age estimation problem. Ricianek et al. [24] developed a multi-ethnic face age determination system to deal with race variations. In this implementation, each facial image is annotated with 161 landmarks. After extracting AAM features, Least Angle Regression (LAR) is used to determine the best features for age prediction. Luu et al. [19] used AAMs and Support Vector Regression (SVR) to build an age estimation framework.

In local features based approaches, facial features are locally extracted from windows around key facial landmarks, which often encode age information (such as wrinkles around the corners of the eyes etc.). Yan et al. [31] used Spatially Flexible Patches (SFPs) and GMMs to setup a method known as Regression from Patch Kernels (RPK). Each SFP integrates local features with coordinate informa-
tion to overcome problems such as occlusions, illumination effects or slight head pose variations. Suo et al. [26] designed a multi-resolution hierarchical graphical face model for age estimation. Facial images are first decomposed into detailed components from coarse to fine resolutions. AAMs are then applied to the overall face and to each type of facial component. Moreover, skin zones such as wrinkles and blobs are also considered in this graph. Exploiting the strength of Gabor features, Gao and Ai [11] used them for facial image representation and used a Linear Discriminant Analysis (LDA) for building an age classifier. They also proposed using fuzzy LDA to cope with age ambiguity. Their method was tested on a dataset containing 4787 consumer images with four age groups: babies, children, adults and the elderly and was found to exhibit a classification rate of around 91%. Recently Guo et al. [16] investigated Biologically Inspired Features (BIFs) derived from a feed-forward model of the primate visual object recognition pathway. The advantage of this approach is that small translations, rotations and scale changes can be well handled by the new set of features. Guo et al. [15] then improved on this method by integrating an age manifold into the aging process. Instead of dealing with each image independently, Geng et al. [12] introduced an Aging Pattern Subspace (AGES) that uses all facial images of each subject. Their idea was to model the aging pattern, a sequence of an individual’s facial images sorted chronologically, by constructing a representative subspace. A detailed literature review on age estimation methods can be found in [10].

3. Contourlet Transform and Nonsubsampled Contourlet Transform

This section will review the Contourlet Transform and the Nonsubsampled Contourlet Transform, which is an extension of the Contourlet Transform that allows redundancy to archive better decomposed images.

3.1. Contourlet Transform

Do and Vetterli introduced the Contourlet Transform [9] that has the ability to represent images with smooth contours. Unlike other wavelet transforms originally developed in continuous domain and then applied to discretized signals, the Contourlet Transform is designed specially to discover the 2D geometry of any given digital image, and not only captures the main features of wavelets, e.g multi-resolution, localization and critical sampling, but also offers a high degree of directionality and anisotropy. The transform is based on an efficient 2D multi-scale and multi-direction image expansion using Directional Filter Banks (DFBs).

The Contourlet Transform is constructed by using a double filter bank to decompose an image into several directional sub-bands at multiple scales. This double filter bank is a combination of the Laplacian Pyramid (LP) with a DFB at each scale. The LP is first used to decompose the image into low frequency and high frequency sub-bands, which capture point discontinuities. Then the DFB is applied on the high frequency sub-bands to link point discontinuities into linear structures resulting in several directional sub-bands. Repeating these processes on the low frequency sub-bands results in a multi-scale and directional decomposition of images. The schematic diagram of the Contourlet scheme is shown in Figure 2.

3.2. Nonsubsampled Contourlet Transform (NSCT)

The Nonsubsampled Contourlet Transform [8] is a fully shift-invariant, multi-scale, and multi-directional transform that can separate strong edges, weak edges and noise from a given image. It is constructed using a Nonsubsampled Pyramid (NSP), responsible for the multi-scale information, and the Nonsubsampled Directional Filter Bank (NSDFB), which is responsible for conducting the directional information. Given an input image \( I \) represented at the \( i \)-th level, \( 1 \leq i \leq J \), where \( J \) is the highest maximum level of the decomposition, \( I \) is divided into a low-pass sub-band \( I^{(L)} \) and a high-pass sub-band \( I^{(H)} \) by using the NSP equipped with a low-pass filter \( H^{(L)} \) and a high-pass filter \( H^{(H)} \) as shown in (1) and (2), in which \( * \) is the convolution operator.

\[
I^{(L)} = H^{(L)} * I_i \tag{1}
\]

\[
I^{(H)} = H^{(H)} * I_i \tag{2}
\]

The high-pass sub-band \( I^{(H)} \) is then decomposed into directional sub-bands by using the NSDFB, which is constructed in cascades comprising of parallelogram filters and two-channel fan filter banks. The low-pass sub-band \( I^{(L)} \) is in turn used for the next decomposition. Consequently, the directional sub-bands \( I_{i,k} \) can be computed by using the equivalent filter \( F^{eq}_k \) using (3) in which \( 2^i \) is the number of directional sub-bands at the \( i \)-th level.

\[
I_{i,k} = F^{eq}_k * I_{i}^{(H)}, \quad k = 1, \ldots, 2^i \tag{3}
\]

The procedure is in turn applied to the low-pass sub-band to obtain the next level decomposition. Generally, the decomposition procedure of NSCT is formulated as shown in (4).
Figure 3. Features extracted by different texture extraction methods (a) Original facial image, (b) Noisy image with std. dev. of noise $\sigma$ set to 0.1, (c) - (f) Low-pass, strong edge, weak edge and noise components respectively obtained after applying LNSCT on the noisy image, (g) LBP map of the noisy image, (h) Noisy image filtered using a Gabor filter.

in which $I$ is the given image, $I_j^{(0)}$ is the low-pass sub-band and $(I_{i,k})_{i,k}$ are directional sub-band coefficients.

$$\text{NSCT}(I) = (I_j^{(0)}, (I_{i,k})_{i,k})$$ (4)

Both noise and weak edges have low magnitude coefficients in the frequency domain and because of this most image enhancement methods cannot separate weak edges from noise. However, weak edges have geometric structure while noise does not. NSCT can therefore be used to discriminate between them. This is illustrated by Figure 3, which shows how NSCT is able to separate out these three components from a noisy image which can in turn enable the extraction of more potentially discriminative feature vectors compared to conventional texture extraction methods such as LBPs or Gabor filters.

The strong edges in an image are the pixels with the largest magnitude coefficients in all sub-bands. Weak edges correspond to those pixels having both large and small magnitude coefficients in directional sub-bands within the same scale. Noise corresponds to those pixels that have small magnitude coefficients in all sub-bands. To enhance the range of values of a processed image, the image is converted from the spatial domain into the logarithmic domain before applying the NSCT [30]. The modified Logarithmic NSCT (LNSCT) coefficients $x$ are nonlinearly mapped by $v(x)$ using (5) in which $\bar{m}$ and $m$ are respectively the mean and the maximum magnitude of the coefficients for each pixel across the directional sub-band, $\sigma$ is the standard deviation of noise of the sub-bands at a specific pyramidal level and $p$ and $c$ are constants with $1 \leq c \leq 5$.

$$v(x) = \begin{cases} x & \text{if } (\sigma \leq \bar{m}) \\ \max((\frac{x}{\sigma})^p, 1) & \text{if } (\sigma > \bar{m}, \sigma \leq m) \\ \max((\frac{x}{m})^p, 1) & \text{if } (\sigma > \bar{m}, \sigma > m) \\ 0 & \text{(weak-edge pixels)} \\ \text{(noise)} \end{cases}$$ (5)

The steps involved in extracting Contourlet features are as follows:

1. Given an image $I$, calculate its logarithm: $L = \log(I)$
2. Apply NSCT on $L$ for $N$ levels
3. Estimate the noise standard deviation
4. For each level of the pyramid:
   - (a) Estimate the noise variance
   - (b) At each pixel, compute $\bar{m}$ and $m$ of the corresponding coefficients in all directional sub-bands at this level, and apply (5)
5. Apply the inverse NSCT to get the processed image

4. Contourlet Appearance Model (CAM)

This section outlines our proposed Contourlet Appearance Model for extraction of features to be used later by face aging functions for age estimation. The three key stages in building CAM are shape modeling, texture modeling and finally combining the resultant shape and texture parameters into a form that can be used for regression based age estimation.

4.1. CAM shape modeling

Our CAM first requires the automatic localization of several facial landmarks before the texture information from within their convex hull is extracted. This process, referred to as the shape modeling stage, is accomplished by using an ASM. Once an ASM is trained on images with a set of manually annotated landmarks (that lie along the facial boundary, eyes, nose, etc.), it can exhibit a fairly high degree of accuracy in locating the same landmarks in unseen faces.
During the training stage an ASM builds a point distribution model to capture the variation in shape exhibited by several faces with manually positioned landmarks. The shapes in the training set are aligned using Generalized Procrustes Analysis \[13\] before the mean shape \( \mathbf{x} \) is calculated. PCA is then applied on the aligned shapes to capture the main axes of shape variation. It now becomes possible to approximate any shape \( \mathbf{x} \) by the shape model equation given by (6), in which \( \mathbf{P}_s \) is a matrix of the principal eigenvectors of the covariance matrix of the aligned shapes and \( \mathbf{b}_s \) is a vector of the PCA coefficients of \( \mathbf{x} \).

\[
\mathbf{x} \approx \mathbf{x} + \mathbf{P}_s \mathbf{b}_s
\]  

(6)

The final stage in the training process involves building statistical models of the local pixel variation in regions around each landmark across all training images. This information is represented by the mean \( \mathbf{g} \) and covariance matrix \( \mathbf{S}_g \) of the profiles around a landmark which are constructed by sampling the image intensities along lines perpendicular to the direction of the shape boundary at a point (1D profiles) or in square regions around it (2D profiles) at multiple resolutions of the face in an image pyramid.

When an unseen face is to be fitted by an ASM, a face detector is first used to initialize its parameters and roughly align the mean shape over the face. The landmarks are then iteratively moved at the lowest resolution image in the image pyramid into locations whose surrounding profiles \( \mathbf{g} \) bear the lowest squared Mahalanobis distance \( f(\mathbf{g}) \) to the mean profile for the respective landmark, as shown in (7).

\[
f(\mathbf{g}) = (\mathbf{g} - \mathbf{g}_0)^\top \mathbf{S}_g^{-1} (\mathbf{g} - \mathbf{g}_0)
\]  

(7)

The PCA coefficients for the shape in question are then calculated and constrained, so that the shape generated is representative of a typical human face. The process is repeated until the landmark locations converge at the lowest resolution and are then scaled and again iteratively updated at the higher image resolutions. Thus, a course to fine adjustment of the landmark positions is made possible and once convergence is reached at the highest resolution, the final set of landmark coordinates is obtained.

4.1.1 Modified Active Shape Model (MASM)

The ASM implementation used in our CAM is the Modified Active Shape Model (MASM) proposed in \[25\]. MASM uses only 2D profiles and 2D search regions in order to determine the best location for each landmark. The use of 2D profiles results in a better modeling of the pixel intensities in regions around each landmark compared to the 1D profiles used by the classical ASM described by Cootes et al. in \[6\]. MASM also goes one step further than classical ASM by building a PCA subspace of the pixel variation around each landmark. At the test stage, a 2D profile around a potential landmark location is projected onto this subspace to obtain a vector of projection coefficients and is then reconstructed with these coefficients. The reconstruction error between this reconstructed profile and the original profile is calculated and the candidate location with the lowest reconstruction error is the one that is used as the new location for the landmark in question. The use of these improvements contribute to the fairly high fitting accuracy of MASM and enable it to outperform classical ASM as demonstrated in \[25\].

4.2. CAM texture modeling

After finding the facial region bounded by the convex hull of the landmarks determined by MASM, CAM texture parameters must be extracted from this region. These CAM texture parameters are extracted using methods similar to those used by AAMs. During the training stage, a PCA subspace \( \mathbf{P}_t \) is built on the “shape-free” and illumination normalized gray-level texture patches enclosed by the convex hulls of several training faces. The phrase “shape-free” is used as all texture patches are warped to a common mean shape thus ensuring that the resultant patches are all of the same shape and contain the same number of pixels. In addition to this gray-level texture subspace, we also build a PCA subspace \( \mathbf{P}_w \) of the warped weak edges texture vectors obtained after processing the images using the NSCT as outlined in subsection 3.2. When a test image has been landmarked, we obtain a “shape-free” gray-level texture vector \( \mathbf{t} \) and a “shape-free” weak edges texture vector.
w which are normalized using the same process followed during training. These texture vectors are then projected onto their respective subspaces to obtain two sets of texture coefficients $b_t$ and $b_w$ using (8) and (9) in which $\bar{t}$ and $\bar{w}$ are the mean texture vector obtained from the original training images and the mean texture vector obtained from the processed images respectively.

$$b_t = P^T_t (t - \bar{t}) \tag{8}$$

$$b_w = P^T_w (w - \bar{w}) \tag{9}$$

4.3. Combining CAM shape and texture parameters

The three sets of parameters that we obtain from the CAM shape and texture modeling are combined and serve as our input to a regression based age estimator function. This is where our implementation differs from the process followed by conventional AAM based aged estimators. During their training stages, AAM based methods extract the shape parameters $b_s$ and texture $b_t$ parameters from landmarked images, combine them into a single vector $b$ and then go on to build a third PCA subspace $P_c$ for this set of combined features so that they can now be represented by (10) in which $c$ is the set of PCA coefficients of the combined appearance model.

$$b = P_c c \tag{10}$$

When fitting an unseen image, the search is driven by updates to $c$ and it is the final set of combined appearance parameters that is used for regression based age analysis. This is where the frailties of AAM based methods emerge. Since the search is driven by the combined appearance parameters and not the texture parameters themselves, the reconstructed “shape-free” patches are not as close to the original gray-level textures as obtained by our CAM. This is clearly illustrated in Figure 4. It is here that our major contribution lies. By not relying on the combined appearance parameters to drive the landmark search, we obtain better estimates of the texture parameters as well as better texture reconstructions compared to AAM based methods. As a result of this, our age estimates also improve as the coefficients used in our age estimation functions are more accurate. This is borne out by the results we obtain in section 6.

5. Using CAM for age estimation

To use our CAM for age estimation, we construct two aging functions, one for youths, using training images with subjects aged between 0 and 20 and another for adults, using training images in which the subjects are aged between 21 and 69. The face aging functions are built using Support Vector Regression (SVR). Given an input image, the combined features are first extracted using our CAM as described in section 4. We then use a Support Vector Machine (SVM) with a Gaussian (RBF) kernel on the hybrid feature vectors, to decide whether the facial image is that of a child’s or an adult’s. Based on the output from the binary SVM classifier, the appropriate aging function is used to determine the age of the face.

6. Experimental Results

For our experiments, we use a landmarking scheme consisting of 68 facial landmarks [20] as shown in Figure 5. MASM is trained on a set of 500 images of 115 subjects manually annotated with these 68 landmarks and drawn from the National Institute of Standards and Technology (NIST) Multiple Biometric Grand Challenge - 2008 (MBGC-2008) still face challenge database [22, 23], which consists of 34,729 frontal images of 810 subjects. At the fitting stage, the Viola-Jones face detector [27] is used to detect faces in test images and initialize MASM.

To test the Combined Appearance Model on age estimation we used two face aging datasets, i.e. FG-NET [18] and PAL [21]. The FG-NET database contains 1002 face images of 82 subjects with ages ranging from 0 to 69. PAL consists of 443 neutral facial images of adults with ages ranging from 18 to 91. Sample images from the two databases are shown in Figure 6. The Mean Absolute Error (MAE), defined as the average of the absolute error between the es-
<table>
<thead>
<tr>
<th>Methods</th>
<th>MAE (years)</th>
</tr>
</thead>
<tbody>
<tr>
<td>CAM</td>
<td>4.12</td>
</tr>
<tr>
<td>AAM</td>
<td>6.77</td>
</tr>
<tr>
<td>AAM + LBP</td>
<td>6.17</td>
</tr>
<tr>
<td>BIFs [16]</td>
<td>4.77</td>
</tr>
<tr>
<td>AGES [12]</td>
<td>6.77</td>
</tr>
</tbody>
</table>

Table 1. Comparison of MAE values (in years) obtained by different age estimation methods when the LOPO scheme was used on the FG-NET database.

<table>
<thead>
<tr>
<th>Methods</th>
<th>MAE (years)</th>
</tr>
</thead>
<tbody>
<tr>
<td>AAM + LBP</td>
<td>7.17</td>
</tr>
<tr>
<td>CAM</td>
<td>6.0</td>
</tr>
</tbody>
</table>

Table 2. MAE values (in years) obtained by using different age estimation methods on the PAL database.

The estimated ages and the ground truths, is computed using (11) and is used as our performance metric to compare the different age estimation techniques.

\[
\text{MAE} = \frac{\sum_{i=1}^{N_t} |\hat{x}_i - x_i|}{N_t} (11)
\]

In (11), \( \hat{x}_i \) is the estimated age for the \( i^{th} \) testing sample, \( x_i \) is the corresponding ground truth age (the true age), and \( N_t \) is the total number of testing samples.

In our experiment, the Leave One Person Out (LOPO) scheme is set up to experimentally evaluate the different methods on the FG-NET database. In this scheme, all facial images of one subject are used as the testing set and the remaining images are used as the training set for training the CAM texture model as well as the face aging functions. This process is applied in turn to all 82 subjects in the FG-NET database. Finally, the overall MAE is computed. Table 1 lists the MAE values obtained by several age estimation methods when the LOPO scheme was used on the FG-NET database while Table 2 compares the MAEs obtained by two additional age estimation techniques when the same testing scheme was used on the PAL database.

Some previously published results used the Cumulative Score (CS) rather than MAE to evaluate their estimation results. It is computed using (12), in which \( N_c \leq \theta \) is the number of testing samples with MAE values less than or equal to \( \theta \). Figure 7 compares the cumulative scores obtained by our method against such approaches when tested on the FG-NET database.

\[
\text{CumScore}(\theta) = \frac{N_c \leq \theta}{N_t} \times 100 (12)
\]

From Tables 1 and 2 and Figure 7 it is clear that our CAM based approach obtains more accurate age estimates than several existing methods.

7. Conclusions

We have proposed a novel Contourlet Appearance Model to localize facial landmarks and extract texture based features from the convex hull bounded by them. Our CAM is shown to be more accurate at reconstructing unseen texture compared to conventional AAMs because of its use of a Modified Active Shape Model for locating the facial landmarks. By decoupling the fitting of facial landmarks from the extraction of texture based features and by the use of the Nonsubsampled Contourlet Transform we are able to obtain discriminative features for age estimation using regression based aging functions. The results of using our CAM based age estimation technique on databases such as FG-NET and PAL are quite promising and highlight the improved accuracy of our method over several existing age estimation techniques.

Acknowledgements

This research was supported by CyLab at Carnegie Mellon under grants DAAD19-02-1-0389 and W911NF-09-1-0273 from the Army Research Office.

References


