Automatic age and gaze estimation under uncontrolled conditions

Alnajar, F.

Publication date
2016
Document Version
Final published version

Automatic Age and Gaze Estimation under Uncontrolled Conditions

Fares Alnajar

ISBN 978-94-6182-648-0

Automatic Age and Gaze Estimation under Uncontrolled Conditions

Fares Alnajar

Automatic Age and Gaze Estimation under Uncontrolled Conditions

Fares Alnajar

This book was typeset by the author using LaTeX 2_ε.

Printing: Off Page, Amsterdam

ISBN 978-94-6182-648-0

Automatic Age and Gaze Estimation under Uncontrolled Conditions

ACADEMISCH PROEFSCHRIFT

ter verkrijging van de graad van doctor
aan de Universiteit van Amsterdam
op gezag van de Rector Magnificus
prof. dr. D. C. van den Boom
ten overstaan van een door het college voor promoties
ingestelde commissie,
in het openbaar te verdedigen in de Agnietenkapel
op donderdag 21 januari 2016, te 12:00 uur

door

Fares Alnajar

geboren te Daraa, Syrië

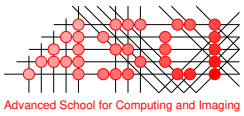
Promotiecommissie:

Promotor:	Prof. dr. T. Gevers
Co-promotor:	Prof. dr. ir. A.W.M. Smeulders
Overige leden:	Prof. dr. ir. F.C.A. Groen Prof. dr. A. Hanjalic Prof. dr. N. Sebe Prof. dr. M. Worring Dr. C.G.M. Snoek Dr. J.C. Van Gemert

Faculteit der Natuurwetenschappen, Wiskunde en Informatica

UNIVERSITEIT VAN AMSTERDAM

This research is supported by the Dutch national program COMMIT. This work was carried out in the ASCI graduate school, at the Intelligent Systems Lab Amsterdam of the University of Amsterdam.
ASCI dissertation series number 344.



COMMIT/

To My Parents: Ibrahim and Mariam.

Contents

Publications

This work is composed of the following publications:

- **Chapter 2**:
  - Fares Alnajar, Caifeng Shan, Theo Gevers and Jan-Mark Geusebroek. *Learning-based Encoding with Soft Assignment for Age Estimation under Unconstrained Imaging Conditions*, Image and Vision Computing, Volume 30 Issue 12, p. 946-953, December 2012.
- **Chapter 3**:
  - Fares Alnajar, Zhongyu Lou, Jose Alvarez, and Theo Gevers. *Expression-Invariant Age Estimation*, British Machine Vision Conference, Nottingham, 2014.
- **Chapter 4**:
  - Hamdi Dibeklioglu, Fares Alnajar, Albert Ali Salah, and Theo Gevers. *Combining Facial Dynamics with Appearance for Age Estimation*, IEEE Transactions on Image Processing, , Volume 24 Issue 6, p. 1928-1943, June 2015.
- **Chapter 5**:
  - Fares Alnajar, Theo Gevers, Jose Alvarez. *Deep Aging Features*, pending submission in Computer Vision and Image Understanding.
- **Chapter 6**:
  - Fares Alnajar, Theo Gevers, and Sezer Karaoglu. *Age Estimation under Changes in Image Quality: An Experimental Study*, International Conference on Image Processing, Quebec, 2015.

- **Chapter 7**:
  - Fares Alnajar, Theo Gevers, Roberto Valenti, and Sennay Ghebreab. *Calibration-Free Gaze Estimation Using Human Gaze Patterns*, pending submission to the International Journal of Computer Vision.

  Ideas previously appeared in:

  - Fares Alnajar, Theo Gevers, Roberto Valenti, and Sennay Ghebreab. *Calibration-Free Gaze Estimation Using Human Gaze Patterns*, International Conference on Computer Vision, Sydney, 2013.

1

Introduction

Since ancient times, the human face has been a medium for revealing many of people's aspects, personality, and characteristics. The ancient Greeks attributed facial features to certain characteristics [61]. Basic face-inferred information (e.g. age, visual attention, emotions, and identity) is well-known and humans have been able to recognize it with ease.

It comes as no surprise that this information plays a role in our life from early on. The knowledge, which comes along with perceiving this information over time, influences how people interact and converse with each another. In addition to having old roots and playing a key role in social interaction, this face-inferred information is culturally independent [32]. Estimating people's age and visual attention along with recognizing identity and emotions is universally applicable and part of everyday life everywhere.

Other face-inferred information is more subtle. The colorization of the face can reveal some social or biological cues. Increasing red or yellow of face's color is positively correlated with perceived health of black African and white Caucasians [106, 107]. Furthermore, the shape of the face is correlated with more subtle information. Various studies [16, 60, 108] have concluded a relationship between the facial width-to-height ratio to aggression, unethical behavior, deception, dominance and even politically-relevant characters [69].

The abundance of information derived from the human face inspired early research in computer vision and artificial intelligence to automatically extract this information [11, 66, 136]. Automatic face recognition has been a success story in computer vision. The line of research initially consisted of basic face-related tasks like age estimation [38], gender [79] and ethnicity recognition [37], as well as gaze estimation [58]. Later, more subtle tasks were added such as recognizing genuine smiles [25, 125] and mirco-

expression recognition [88].

As the research performance continued to improve, additional general face-related challenges were being considered. Changes in the lighting conditions and digital noise hinder the recording of faces [117]. Occlusion and changes in head pose obscure parts of the face which may contain informative cues [10]. Furthermore, task-specific face-related challenges were being addressed. Facial expressions change the dynamics of the facial muscles and make it more difficult to estimate a person's age [54]. Omitting user active engagement from calibrating gaze estimators, as required in some practical situations, proposed the challenge of automatically calibrate the gaze estimators [58].

Such challenges reflect realistic instantiations of the tasks at hand. Finding solutions under these challenges bridges the semantic gap in automatic face analysis, and pushes for efficient realization of these solutions for everyday life problems.

This thesis addresses two face-analysis tasks: automatic age estimation and automatic eye gaze estimation. Different from previous approaches [38], the focus is to provide (practical) solutions for such tasks under variant conditions and setups.

1.1 Age Estimation

Consciously and subconsciously, people estimate the age of others on a daily basis. The social interaction between two people is often influenced by the difference in perceived age. Perceived physical maturity usually correlates with mental maturity, which directs our communications with others. In a social event, people in their thirties would likely converse with people in their age group differently than with others in different age groups. Estimating the age of a person is essential for daily social communication and, therefore, has become the focus of various studies [9, 35, 92].

Humans perceive age primarily through the changes in the facial skin. During aging, the human face loses collagen beneath the skin leading to thinner, darker, and more leathery skin [38]. Age-induced facial wrinkles become more distinct as a result of repeated activation of facial muscles and they start to appear in different directions depending on these muscles [21].

With the recent surge in the fields of Human-Computer Interaction (HCI) and automatic human behavior analysis, automatic age estimation has become an urgent topic of research. In Electronic Customer Relation Management (ECRM) – a company's strategy to interact and communicate with current and potential customers – the age of customers is a key knowledge for efficient marketing. A company which sells anti-baldness cures would, reasonably, rule out teenagers from its target groups. Early works in automatic

age estimation focused on the changes in head size during early aging [67]. For later age stages, facial skin changes are chiefly employed to automatically tell the age. Many research works [41, 53, 130, 133] designed aging features with the aim to characterize the wrinkles and other skin-related changes. These features are then mapped, by machine learning techniques, to the estimated age.

Age estimation may be hindered by independent factors like facial expressions, which induce the movement of some facial muscles that overlap with age-related ones. Moreover, poor imaging conditions introduce noise, which, affects the estimated age. Aging cues, like wrinkles and fine skin details, are particularity prone to these changes. Such challenges suggest the necessity to go beyond the standard scenarios to address real-life cases.

1.2 Eye Gaze Estimation

Understanding where a person is looking at is key for social communication. It is actually argued that understanding the gaze is essential for early learning of infants [13, 14]. Knowing where people gaze reveals their areas of attention and what their interests are. The duration, fixations, and the temporal-spatial patterns reflect the behavior of the person and their interaction with the surroundings.

Successes in automatic gaze estimation have triggered applications in various domains. Early applications were designed to assist disabled users in interacting via eye movements [63]. Another application is eye typing [76] where text is generated by looking at keys on the screen. Automatic gaze estimation has been integrated in monitoring and surveillance systems; e.g. monitoring behavior in plane cockpits or driver fatigue detection [8, 64]. The more HCI systems are integrated in everyday life, the more automatic human behavior analysis and, consequently, gaze estimation are needed. Recently, in marketing research and ECRM, the behavior and interests of (potential) customers are monitored around the clock.

Typical gaze estimation systems require explicit calibration. One type of calibration aims to estimate the relationships between the user, the gaze plan (e.g. screen), and the camera [58]. The user is usually asked to follow explicit instructions. While this procedure is feasible for some applications such as eye typing, marketing research will restrict to little active interaction from the user and ECRM will permit no active user contribution at all. In the latter cases, estimating the gaze points should be carried out completely passively.

1.3 Research Questions

In this thesis, we aim to address the following research questions:

Question 1: For poor-quality images, does loose interpretation of aging features alleviate the negative influence of noise on age estimation?

When presented with poor-quality images, the facial skin complexion is often distorted or obscured. While this is a challenge in automatic face analysis in general, it is particularity problematic for automatic age estimation since it relies primarily on the fine skin changes. The skin details become less lucid and the attribution of certain skin features (aging cue) to an estimated age becomes less certain. To overcome this, in chapter 2, we assign multiple interpretations to the aging features. In other words, the single 'ambiguous' age cue can suggest more than one age at a time. Technically speaking, the aging features are assigned to pre-learned visual words in a soft manner which results in more noise-robust features and, hence, better performance.

Question 2: How can we alleviate the negative influence of facial expressions when predicting age?

Facial expressions induce changes in facial muscles, which distort the aging cues. A facial expression is described by a combination of these changes in the face, which are called Action Units [33]. A problem in age estimation is that expression-related muscles overlap with aging-induced facial changes. For example, smiling involves the activation of some facial muscles leading to raising the cheeks and pulling the lip corners. This influences the aging wrinkles around the mouth and near the eyes. Consequently, the changes to aging cues caused by expressions show the necessity of addressing the influence of expression when estimating age. To this end, in chapter 3, age and expression are jointly learnt to model their relationship. The aim is to achieve expression-invariant age estimation. More specifically, we introduce a new graphical model, which contains a latent layer between the age/expression labels and the facial features. This layer captures the relationship between age and expression and, consequently, leads to better age prediction.

Question 3: Can we use the movements of the facial muscles to infer further details about age and hence achieve better age estimation?

During aging, the face experiences multiple changes in muscle tones and fat tissues. While such changes induce age-related wrinkles, it also influences the way facial expressions are being displayed. In chapter 4, we aim to make use of some dynamic facial features to boost the accuracy of age estimation. Dynamic features such as speed, acceleration, and amplitude are extracted from facial landmarks such as eyelids, lip corners, and cheeks. When combined with skin appearance features, the dynamic features

produce more discriminative aging descriptors and hence, better age estimation.

Question 4: Can we automatically design region-specific, efficient and robust aging features?

Facial aging cues differ from one area to another. Several aging features [41, 53, 122] are aimed to capture the details of the wrinkles on the face (wrinkle features) while others [7, 83, 130] measure aging changes of skin texture (skin texture features). These features are typically applied exhaustively to cover the entire face. Other works [21] combine both features for more discriminative aging descriptors. However, the design (or choice) of the feature types and the corresponding face regions are handcrafted. This suggests designing a feature extraction scheme that is automatically adapted for each facial area. To this end, in chapter 5, region-specific aging filters are learnt for each of the different face regions. More specifically, a convolutional neural network is assigned to each facial part. The filters are designed to fit the corresponding face regions and, hence, produce effective, yet robust aging descriptors.

Question 5: What is the influence of different types and levels of digital noise on the performance of aging features?

The influence of image quality on the performance has long been a challenge in face-related image processing tasks. It is particularly important to address this challenge in age estimation since the aim is to capture subtle aging cues such as skin texture and wrinkle. Such cues are sensitive to small changes in image quality. Moreover, face images in real-life scenarios are taken using various capturing devices and are prone to noise due to digital transmission and compression. This makes it important to study the performance of aging features with varying image quality degradation. In chapter 6, we introduce a scheme to explore the influence of image quality on the performance of appearance aging features and we propose a basic framework to automatically assign the best aging features based on the quality of the face image.

Question 6: Can we automatically calibrate gaze estimation systems without any active engagement from the user side?

A prerequisite for current gaze estimation systems is user-involved calibration. This process is needed to set values of a number of parameters. For instance, geometric-calibration involves determining the relative locations and orientations between the elements of the setup (e.g. the user, the plane, and the camera) [58]. For some applications like analyzing customers' gazing behavior in shopping mall, the calibration should be done entirely passively on the user side. In chapter 7, we exploit the gaze patterns of others to auto-calibrate the gaze system (i.e. without manual calibration). We make use of the observation that the gaze patterns of people are indicative of where a new user will look in that same scene [65]. The calibration process can be seen as adding con-

straints to a set of variables (parameters) and solving the system. In most of the current gaze estimators, the constraints are set by eye (image) measurements while fixating at predefined points on the gaze plane. In our work, the geometrical constraints defined by the human gaze patterns (for the same stimulus) serve to obtain calibration-free gaze estimation, which is sufficient to trace the user's attention.

2

Soft Encoding for Age Estimation under Uncontrolled Conditions

2.1 Introduction

Automatic age estimation of a person is an interesting and challenging task, with many important applications in human-computer interaction, market intelligence and visual surveillance. Since human faces provide most information to perceive the age, most previous research efforts have focused on age estimation from face images [38].

Constructing a proper face image representation is a key component for successful face age estimation systems. Typically two kinds of features are extracted from face images: appearance features (e.g. wrinkles, skin roughness) and geometric features (e.g. shapes, ratios of distances between facial landmarks). For applications where images acquired in unconstrained settings, it is difficult to automatically detect a sufficient number of fiducial landmarks to compute the geometrical features of the face.

As reviewed in [38], many approaches have been exploited to represent and model faces from images such as anthropometric models, age subspace or manifold, and active appearance models. However, each representation has its limitations and strengths. For example, the anthropometric model is useful for young ages, but not appropriate for adults; for age manifold learning, a large number of training samples is needed. The facial representation should not only be discriminative but also robust to appearance variations and noise. In recent years, local descriptor based approaches have been proven to be effective for face image analysis [23, 71, 75, 83]. Traditionally, Gabor-wavelets have widely been exploited to model local facial appearance [42, 75]. Recently, the his-

togram of Local Binary Patterns [83] have been adopted to describe the micro-structures of the face [3, 104, 133]. Tolerance against monotonic illumination changes and computational simplicity are the most important properties of LBP features. Scale-Invariant Feature Transform (SIFT) [71] and Histogram of Oriented Gradients (HOG) [23] are other types of local descriptors that have shown good performance in face analysis [48] and object recognition.

More recently, Cao et al. [15] argued that these local descriptors use manually designed encodings, and it is difficult to get an optimal encoding method. As shown in [15], the existing handcrafted codes are unevenly distributed, and some codes may rarely appear in face images. This means that the resulting code histogram is less informative and less compact. They presented a learning-based encoding method, which adopts unsupervised learning methods to encode the local micro-structures of the face into a set of discrete codes. With Principal Component Analysis (PCA) and normalization, their learning-based descriptor achieves superior performance on face verification. Instead of face verification, in this paper, we consider learning-based encoding in the context of age estimation.

We adopt the learning-based encoding method for age estimation and propose a principled approach of extracting robust and discriminative facial features and encoding. First, instead of learning a codebook from the entire face, we extract and learn multiple codebooks for individual face patches. The intuition behind this is that the features histogram is computed for each patch. Second, the encoding is done by a weighting scheme in which each pixel is softly assigned to multiple candidate codes. This is to alleviate ambiguity especially in noisy real-life images. Aging effects are mainly observed as textural variations in faces such as wrinkles and other skin artifacts. Therefore, we investigate the use of orientation histogram of local gradients to describe faces for age estimation.

The rest of the paper is organized as the following. In Section 2.2 we provide an overview on related work. Section 2.3 describes learning-based encoding method. We outline our adaptations in Section 2.4. Experiments are presented in Section 2.5. Section 2.6 concludes the paper.

2.2 Related Work

In the last few years, many research efforts have been invested on age estimation from face images. A thorough survey of the state of the art can be found in [38].

Geng et al. [45] introduces the Aging Pattern Subspace for age estimation, where an aging pattern is defined as a sequence of face images from the same person, sorted in the

temporal order. This approach is evaluated on the FG-NET aging database, achieving a Mean Absolute Error (MAE) of 6.77 years. However, in general, it is difficult to collect multiple face images of the same person at different ages. Instead of learning a specific aging pattern for each individual, a common aging pattern could be learned from face images of multiple people [39]. Manifold learning techniques are adopted to embed face images into a low-dimensional aging manifold. The age manifold based regression [51] produces a MAE of 5.07 years on the FG-NET aging database.

Further, Yan et al. [130, 132] propose to use Spatially Flexible Patches as face representation. This technique considers local patches and information about the position. Modeled by a Gaussian mixture model, their approach achieves a MAE of 4.95 years on the FG-NET database. Guo et al. [53] introduces the Biologically Inspired Features for age estimation. Combined with SVM, the proposed features produce a MAE of 4.77 years on the FG-NET database. Recently Ni et al. [81] collected a large web image database, and built a universal age estimator based on multi-instance regression.

Yang and Ai [133] consider LBP features for age estimation. They achieve the error rate of 7.88% on the FERET database and 12.5% on the PIE database. Further, Gao and Ai [42] study the problem of age estimation in consumer images. In their approach, Gabor features are extracted and used with Linear Discriminant Analysis (LDA). They consider four age categories: baby (0-1), child (2-16), adult (17-50), and old (50+). Trained on 5,408 faces, their age estimator achieved an accuracy of 91% on 978 testing images. Gabor features are demonstrated to be more effective than LBP features and pixel intensities in their study. More recently, Shan [103] applies Adaboost to learn local features, both LBP and Gabor features, for age estimation on real-life faces acquired in unconstrained conditions.

Cao et al. [15] presents a learning-based encoding method, which adopts unsupervised learning methods to encode the local micro-structures of the face into a set of discrete codes. The method achieves high accuracy for face verification. In the next section, we extend this method to age estimation. The extension consists of three points: the codes are assigned in soft manner, different codebooks for different face patches, and using features more related to estimating the age.

2.3 Learning-based Encoding

In this section, we briefly describe the learning-based encoding method [15]. At each pixel, its neighboring pixels intensities are sampled in a ring-based pattern to form a low-level feature vector. $r*8$ values are sampled at even intervals on the ring of radius r. The authors extensively varied the parameters (e.g. ring number, ring radius, sampling

number of each ring), and found the differences among patterns are not of influence on the face database they used. Following [15], we use the second sampling method with two rings ($r = 1$, $r = 2$, with center), that is, 25 values (8 from the first ring, 16 from the second ring, and the center value). After sampling, the sampled feature vector is normalized into unit length, to make the feature vector invariant to local illumination changes.

Then, the encoder is learned by applying unsupervised learning to a set of training face images. The feature vectors are extracted at each pixel. Different unsupervised learning methods are considered. In [15], three methods are examined: K-means, PCA tree, and random-projection tree [36]. Their experiments show that the difference among these learning schemes is small. In this paper, the PCA tree [36] is adopted. The largest principal component for the vectors at each node is first computed. After projecting the vectors onto that principal component, the vectors are split from the median value and two children nodes are created; the principal component and the median value are stored in the parent node. These children nodes are further split until the leaf number is equal to the code number, where each leaf represents one code. With the learned encoder, the input face image is encoded. Similar to LBP features, the encoded face image is divided into a grid of patches (7×5 patches used in [15]), and the code histogram computed at each patch is concatenated to form the descriptor of the whole face image.

2.4 Our Approach

In this section, the learning-based encoding method is transformed to face age estimation.

2.4.1 Patch-based Code Learning

In Cao et al. [15], the code set is learned using the sampled vectors from the whole face. However, the histograms are derived at the level of regions (patches). The histogram is constructed from the sampled vectors in each patch. These histograms are concatenated later to form the global descriptor.

There are variations among different face patches. Each individual patch may have different codes or code distributions, e.g. some codes may appear frequently in one patch while they are rare for another patch. To illustrate this point we build two code sets from 2080 training images (used in Section 2.5). One code set is learned from the sampled vectors extracted from the whole face, and the other is learned from the sampled vectors extracted from one face patch (the upper left). Later, we extracted the sampled vectors

from the upper left patch in 664 testing images (also used in Section 2.5), then we encoded the vectors using the two code sets and constructed the frequency histograms. Figure 2.1 shows the two histograms. As can be observed, for this face patch, the codes learned from the whole face are unevenly distributed (i.e., some codes rarely appear), while the codes learned from the face patch are more uniformly distributed (i.e., they are used more efficiently). Therefore, with different code set for each individual patch, the code histogram is much more informative and compact. However, learning multiple code sets introduces increase in both time and memory complexities.

2.4.2 Soft Encoding

When encoding the input image with the learned codebook, each sampled vector (at each pixel) is assigned to the closest code. We call this hard encoding. However, for face images (especially real-world images), ambiguities always exist. That is, for a given sampled vector, there are multiple candidate codes. Assigning to the closest code makes the encoding sensitive to image noise and varying conditions (e.g. illumination). These factors can distort the sampled vector, resulting in different code assignments. We use soft encoding assigning the given sample vector to multiple codes with weights. Soft encoding is used in image classification [126].

When deriving the codes with the PCA tree, after dividing the training samples using the median value, a Gaussian distribution model is estimated for each branch. For soft assignment, the probability that it is from either branch is estimated using the Gaussian model. This is used as the weight for that branch. The weight is multiplied with the weight coming from the parent node. The new weight is passed to the children. In this way, each code (leaf) is assigned with a weight $\mathrm{leaf}_c(r_i)$, where c is the code, r_i is the feature vector i. The encoding is started with weight of 1 at the tree root. The weights of all the codes are normalized. Thus the histogram bins are computed as follows:

$$Bin(c) = \sum_{i=1}^{n} \frac{\mathrm{leaf}_c(r_i)}{S_i} \tag{2.1}$$

$$S_i = \sum_{c=1}^{C} leaf_c(r_i) \tag{2.2}$$

where C is the number of codes, n is the number of sampled vectors, and S_i is a normalization factor, i.e., the sum of weights of all codes (for the given sampled vector).

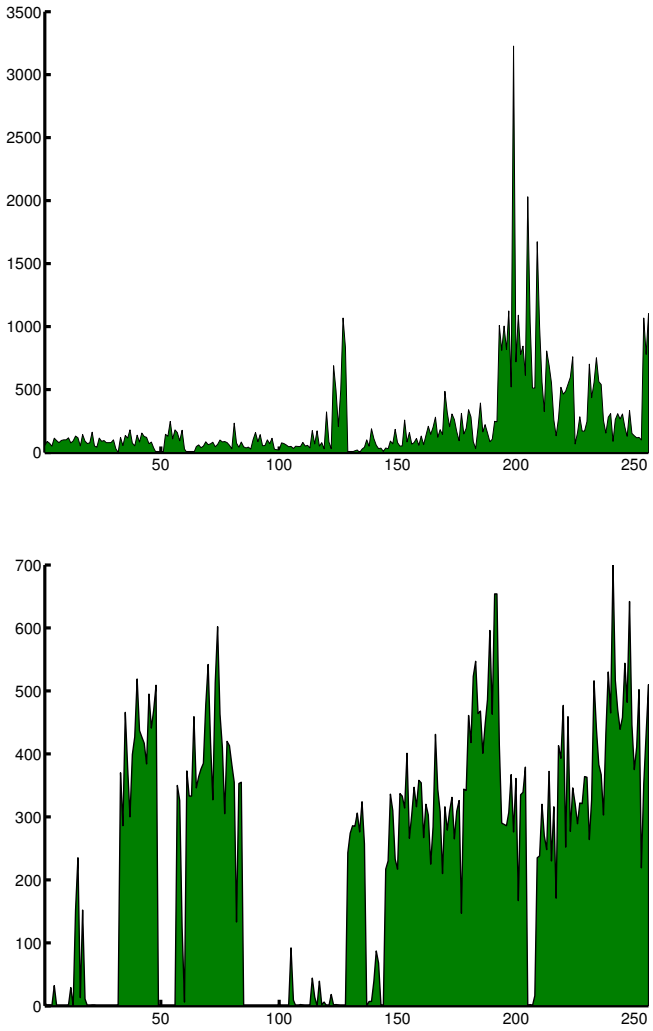


Fig. 2.1: The codes frequency histograms in one face patch of 664 face images using two different code sets; one learned from the whole face (*Top*) and the other learned from the corresponding face patch (*Bottom*). 2,080 face images are used for learning both code sets.

2.4.3 Orientation Histogram of Local Gradients

For each pixel, neighboring pixels are sampled in the ring-based pattern to form a low-level feature vector. However, the extracted local features are sensitive to image noise and illumination variations. Furthermore, as aging effects in faces are mainly observed as texture variations such as wrinkles and other skin artifacts, local gradients (or edge

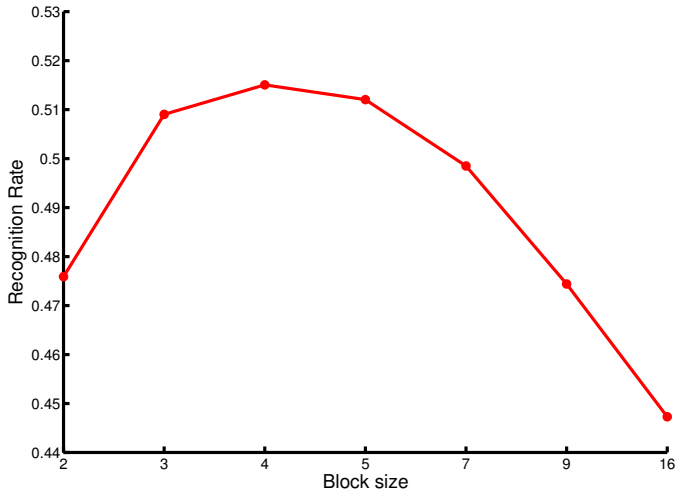


Fig. 2.2: The performance over different block sizes.

responses) may be more effective. Following HOG [23], we extract the orientation histogram of local gradients in neighborhood as the low-level feature vector for code learning.

Therefore, we use the following approach. Given a pixel, local gradients in the neighborhood (i.e., local block) are computed, and a 1-D histogram of gradient directions is accumulated over the pixels in the block. The orientation bins are evenly spaced over $0° - 360°$. Each gradient contributes to one or more bins, where the vote is weighted by the magnitude of the gradient; the magnitude is added to the corresponding bin. There are some parameters to choose in the implementation, including block size, gradient computing, and orientation binning. Therefore, we aim to study the influence of the various on the learning-based encoding. We use the dataset detailed in Section 2.5, where 1,000 face images are used for code learning. 2,080 training images and 664 testing images are used for age group classification n using linear SVM. All faces have a resolution of 61x49 pixels. Throughout this section, results are obtained with the following default setting: 5x5 block size, 8 orientation bins (i.e., each bin covers angle of $45°$), gradient computing Sobel-1D [-1,0,1].

Block Size — We test the block size of 2×2, 3×3, 4×4, 5×5, 7×7, 9×9, 16×16. Figure 2.2 shows the results of different block sizes when using 256 codes. It seems the block size of 4x4 or 5x5 are the best choice for the dataset we use.

Gradient Computation — We test different gradient filters, namely: Sobel-1D [-1,0,1],

Sobel-2D [-1,-2,-1; 0,0,0; 1,2,1], cubic [1,-8,0,8,-1], diagonal [-1,0; 0,1], Prewitt [1,1,1; 0,0,0; -1,-1,-1] and Gaussian derivatives with different sigma values. The best performance using 256 codes is achieved using Gaussian derivatives with $\sigma = 0.75$. It seems that the smoothness of Gaussian helps, and fine scale derivatives perform better for this task.

Gradient	Result(%)	Gradient	Result(%)
Sobel1-D	51.2	Gaussian (0.5)	51.8
Sobel2-D	50.9	Gaussian (0.75)	52.8
Cubic	46.1	Gaussian (1)	52.3
Diagonal	51.7	Gaussian (3)	47.9
Prewitt	49.2	Gaussian (5)	41.9

Tab. 2.1: The performance with different gradient filters.

Orientation Binning — We test different bin numbers (2,3,4,6,8,12,16) with Gaussian and Sobel-1D gradients using 256 codes. The Gaussian derivative consistently outperforms Sobel-1D for all bin numbers. The best results are achieved using 6, 8 or 12 bins.

2.5 Experiments

2.5.1 Dataset and Experimental Settings

In most of the existing studies, face images with limited variations are considered. Images are usually high-quality frontal faces, occlusion-free, with clean background and limited facial expressions. However, in real-world applications (e.g. collecting demographic statistics in shops), age estimation needs to perform on real-life face images captured in unconstrained environments. There are appearance variations in real-life faces, which include facial expressions, illumination changes, head pose variations, occlusion or make-up, and poor image quality. Therefore, age estimation on real-life face images is much more challenging.

The FG-NET dataset is used in many studies. It contains face images with 68 facial landmarks. These landmarks are manually detected and often used by other methods to extract shape information that helps estimating the age [45, 51, 131]. However, under unconstrained conditions these landmarks cannot be accurately detected automatically. And using manually-annotated landmarks is not plausible in real-life applications. So comparing our method with other methods applied on FG-NET dataset is not feasible.

To analysis the contribution of the manually annotated landmarks, Choi et al. [21] compared the performance of their method using manually and automatically obtained landmarks on FG-NET dataset. The MAE error increased around 20%.

Therefore, in this paper, we conduct experiments on real-life faces using a face image set[1] collected recently [40]. The dataset consists of 28,231 faces from 5,080 Flickr images, 86% of which were detected by a face detector, and others were manually added. Each face was labeled with the gender and age category. Seven age categories were considered: 0-2, 3-7, 8-12, 13-19, 20-36, 37-65, and 66+, roughly corresponding to different life stages. Example faces in the dataset are shown in Figure 2.3.

The dataset contains large diversity in race, pose, illumination conditions, and facial expressions. Many faces in the dataset have low resolution: the median face has only 18.5 pixels between eye centers, and 25% of the faces have under 12.5 pixels. To study age estimation on faces with reasonable resolution, Shan [103] considered only faces with the eye distance more than 24 pixels. This results in a collection of 12,080 faces. The author selected 2,080 faces as the training set, and 644 faces as the testing set. The gender in the training/testing data sets is evenly distributed. In our experiments, we select another 1,000 face images that are excluded from the training/testing sets for code learning, and perform age group classification using the training/testing sets. All face images are normalized to 61×49 pixels based on eye centers. Linear SVM is used as the classifier for simplicity. We used LIBSVM [2] for training and testing.

2.5.2 Experimental Results

Code Learning: Image vs Patches — We first examine the learning-based encoding method for age estimation. Figure 2.4 shows the results. It is shown that the recognition performance increases when the code number increases for most of code numbers. The performance decreases a bit when the code number is higher than 512. This might be due to overfitting when learning the codebook for large number of codes. The best performance of 56.2% is obtained using 512 codes. Then we compare this default image-based learning with the patch-based learning. The patch-based learning provides comparable or better performance than the image-based learning for most of the code numbers. The best performance is 56.5% with 128 codes. This suggests that code learning at the regional level leads to more informative code histogram.

Soft Encoding — We apply soft encoding for face image encoding. The results are shown in Figure 2.5. It is evident that soft encoding achieves better results than hard

[1] chenlab.ece.cornell.edu/people/Andy/ImagesOfGroups.html

[2] www.csie.ntu.edu.tw/ cjlin/libsvm

Fig. 2.3: Example faces in the dataset [40].

encoding. This illustrate that soft encoding leads to a more robust code histogram.

Orientation Histogram of Local Gradients (OHLG) — We conduct experiments on code learning using the OHLG feature extraction. Based on the study in Section 2.4,

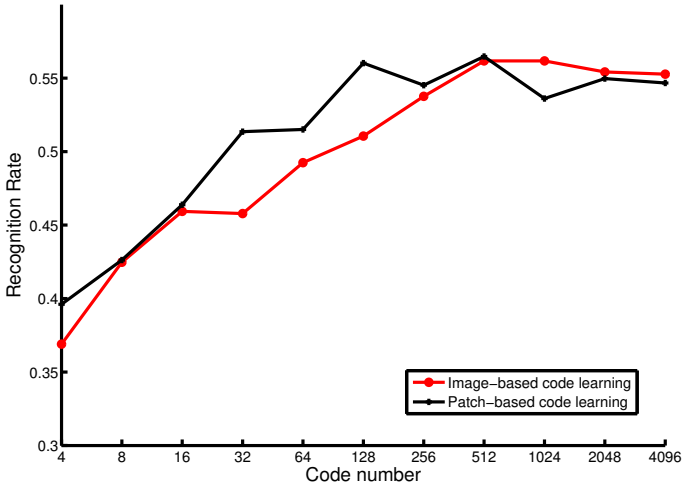


Fig. 2.4: The performance of image-based learning vs patch-based learning over different code numbers.

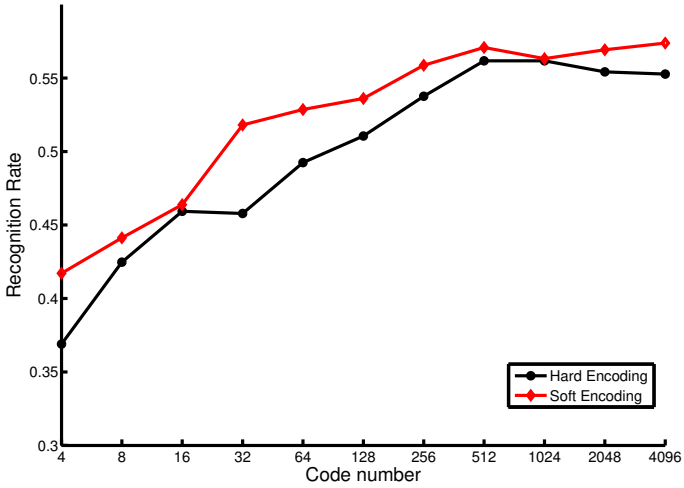


Fig. 2.5: The performance of soft encoding vs hard encoding over different code numbers.

we select the following setting: 5x5 block size, Gaussian derivative, and 8 orientation bins. Figure 2.6 compares the results of OHLG with the sampling method. It is shown that the OHLG feature extraction produces comparable performance as the ring-based sampling. It does not outperform the sampling method. This might be due to the poor quality of the images for which the textural patterns (e.g. wrinkles) are not obvious. To verify this, we further conduct experiments on the dataset with better quality face images.

We conduct experiments on the FG-NET database [1] and MORPH database [93], both

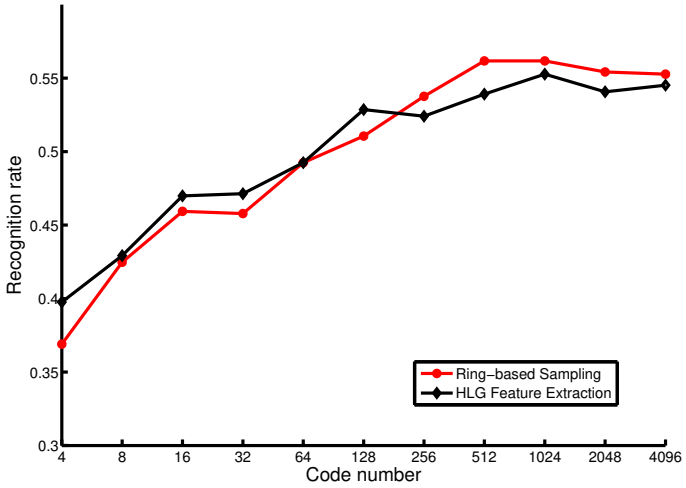


Fig. 2.6: The performance of code learning using the OHLG feature extraction over different code numbers.

of which have better quality faces. FG-NET contains 1,002 face images from Caucasian people, with the ages range from 0 to 69 years. MORPH contains 1,690 images from different ethnicities (433 Caucasian-descendant faces), with the age ranging from 15 to 68. We use the FG-NET data with ages between 15 and 68 as the training set, and use the 433 Caucasian images from MORPH as the testing set. The code learning is done using the remaining non-Caucasian faces in MORPH. Since we have exact ages instead of categories, we use the Mean Absolute Error (MAE) as the criterion. Figure 2.7 shows the results. It can be derived that the OHLG feature extraction outperforms the sampling method in most of code numbers. We further test soft encoding with the OHLG feature extraction on the MORPH and FG-NET dataset. The results are shown in Figure 2.8. Soft encoding reduces the MAE when using the OHLG feature extraction for most of the codes, especially for larger codes. Overall, soft encoding with OHLG feature extraction outperforms the ring-based sampling for all code numbers. This illustrates the effectiveness of our improvement.

Codebook discriminative power — Since the codebook is learned from a separate set, the discriminative power of the images in this set and how much they reflect the differences between the age categories may affect the discriminative power of the codebook. In the following experiment, we test different sets for learning the codebook. Sets with sizes 500, 750, 1000, 1250, and 1500 are taken. The larger sets contain the smaller ones. For each set we ran the experiment using soft encoding over different code numbers. The performance is evaluated by a 2-fold cross-validation over the training set. This is to ensure that the learning code set does not fit the test set. The results are shown in Figure 2.9. We noticed that the 750-image set gave the best results. This suggests

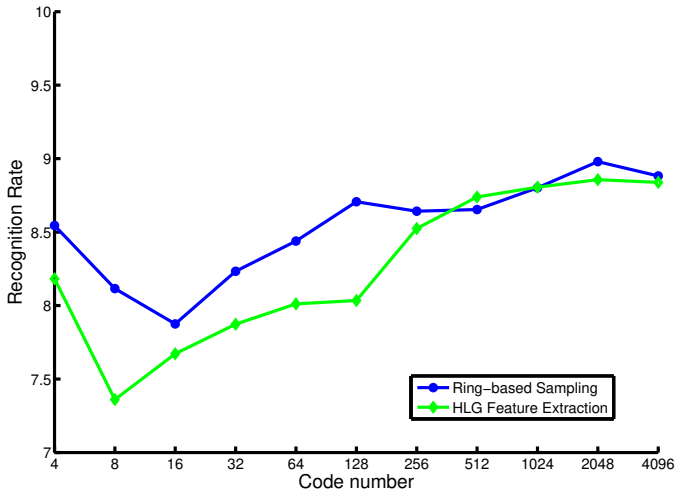


Fig. 2.7: The MAE on the MORPH and FG-NET dataset of code learning using the OHLG feature extraction.

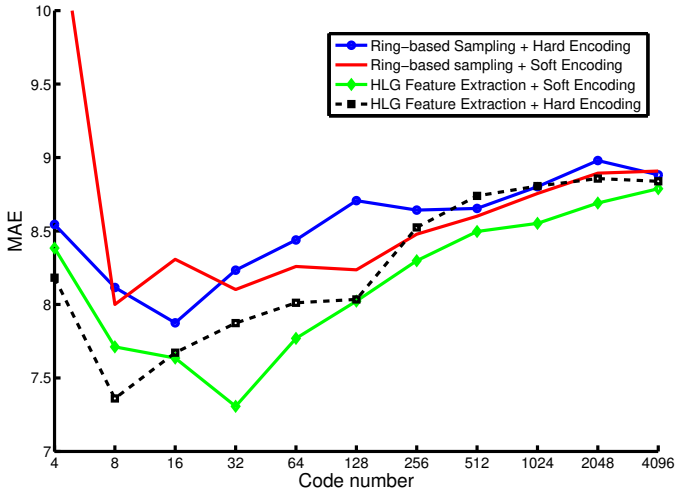


Fig. 2.8: The MAE on the MORPH and FG-NET dataset using soft encoding with the OHLG feature extraction.

that the corresponding codebook is the most discriminative. The codebooks learned from the larger sets result in lower performances. It is possible that images outside the 750-image set may contain noise negatively affecting the discriminative power of the codebook. We reran the experiment using the codebook learned from the 750-image set. Following the setup in [103], we train the descriptors over all the training set images and reported the results on the testing set in Figure 2.10. The highest recognition rate 59.5% achieved using 1024 codes. This is 3.6 point higher than the last reported

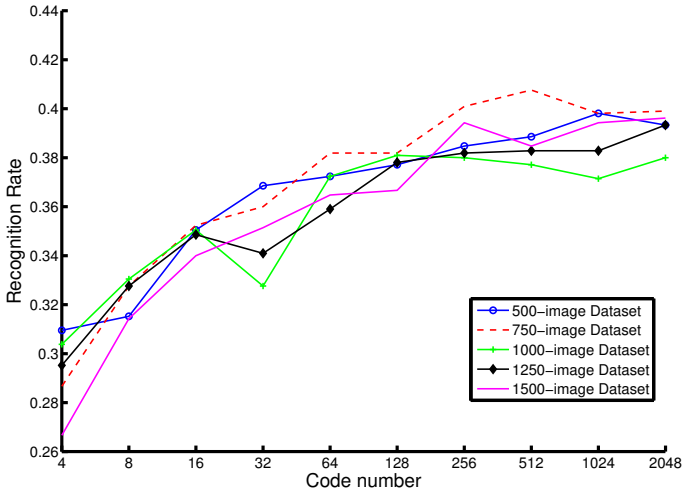


Fig. 2.9: Soft encoding using different learning code sets. The results were computed using two-fold cross-validation on the training set.

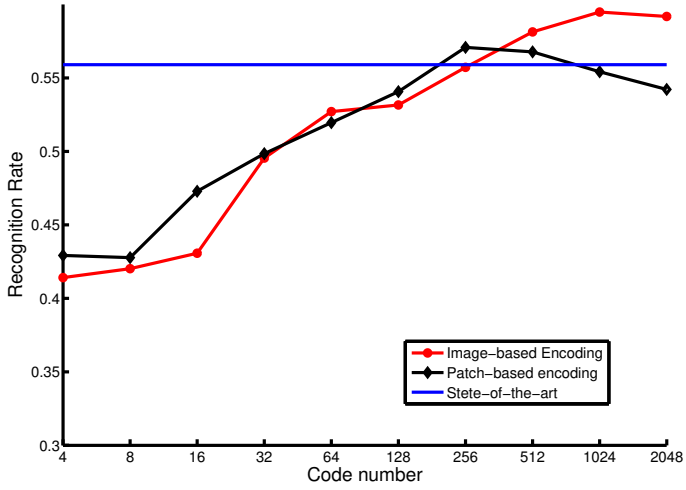


Fig. 2.10: Soft encoding using 750-image learning code set. The red curve represents Image-based encoding results while the black one represents patch-base encoding results.

result in [103], were the recognition rate was 55.9%.

Face Verification — We apply soft encoding to the face verification problem. The LFW benchmark [62] is used. The LFW test set consist of 10 subsets each contains 300 same-person pairs and 300 different-persons pairs. The evaluation is reported using 10 fold cross-validation. At each fold, one subset is used for testing and other 9 are used for training. The final results are the average of the 10 folds results. Another 1000 images are used for learning the codebook. The face size is 96x84. As in [15], we apply a DoG preprocessing step and the codes are learned once for all the 10-folds. The 1000 images identities, used for learning the codebook, never appear in the 10 sets. Figure 2.11 shows the results.

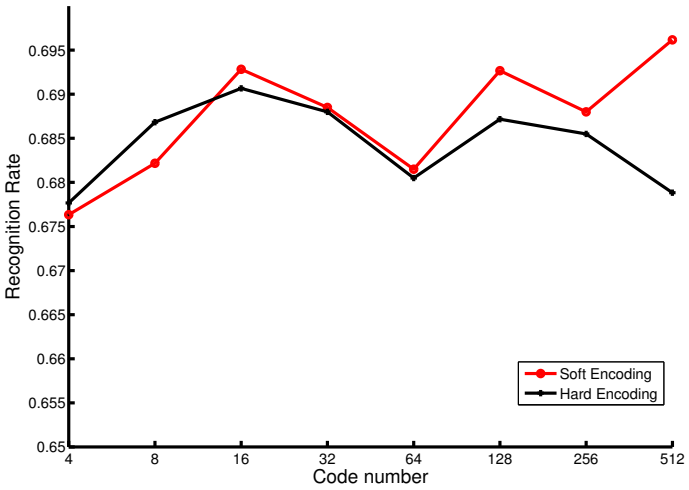


Fig. 2.11: The performance of soft encoding vs hard encoding over different code numbers on LFW face verification dataset.

Soft encoding achieves higher results than hard encoding for most of the code numbers. This suggests that our method can be directed to other face-related problems. The reported results in [15] are around 5% higher than our results. Cao et al. used another commercial software for face alignment [3]. This might explain the difference of the results. In their paper, Cao et al. applied further dimensionality reduction and normalization steps. Here we compare with the raw feature vectors.

[3] After personal communication with the first author of the paper (Cao).

2.6 Conclusions

In this paper, we adopted the learning-based encoding method for age estimation. Instead of learning a set of codes from the entire face, we extracted and learned multiple codebooks for individual face patches. Soft encoding has been used. Orientation histogram of local gradients in neighborhood has been introduced as feature vector for code learning.

Experiments showed that our extensions produced better or comparable performance for most of the cases. Using discriminative codebook, our method outperforms the best performance reported on Gallagher dataset [103]. We extend our method to face verification and show improvements which suggests that our method can be directed to other face-related problems.

3

Expression-Invariant Age Estimation

3.1 Introduction

Automatic age estimation is an important research field in the area of computer vision and has many applications such as human-computer interaction, security, and surveillance. In general, the human age is derived from facial aging cues. The aging of adults is primarily perceived via skin changes [38]. During aging, the human face loses collagen beneath the skin leading to thinner, darker, and more leathery skin [38]. Age-induced facial wrinkles become more distinct as a result of repeated activation of facial muscles and they start to appear in different directions depending on these muscles [21]. For example, vertical wrinkles intensify between the eyebrows while horizontal wrinkles become more apparent close to the eye corners.

External factors like facial expressions cause changes in facial muscles which distort the aging cues. A facial expression is explained by a combination of these changes in the face which are called Action Units [33]. A problem in age estimation is that expression-related muscles overlap with aging-induced facial changes. For example, smiling involves the activation of some facial muscles leading to raising the cheeks and pulling the lip corners. This influences the aging wrinkles around the mouth and near the eyes. Consequently, the aging cues changes caused by expressions show the necessity of separating the influence of expression when estimating the age.

Most of the existing age estimation methods assume that faces show little or no expressions and ignore the changes of the face appearance induced by them. Guo et al. [54] study human age estimation under facial expression changes. Their method learns the correlation between two expressions at a time (e.g. neutrality and happiness). To

predict the age across two expressions, the face is mapped from one expression (e.g. happiness) to another (e.g. neutrality). Next, the age is predicted from the "mapped" face. For the face aging representation, Biologically-Inspired Features (BIF) [53] and Marginal Fisher Analysis (MFA) are used. Zhang et al. [139] employ a weighted random subspace method to solve cross-expression age estimation. In their method, several feature sets are generated first, then subspaces are built for these sets. Next, a classifier is learnt for each subspace and predictions of all classifiers are fused to produce the final prediction. Their method does not require different expressions from the same subjects as opposed to [54]. However, both methods [54, 139] require the expressions of test images to be known before predicting the age which limits their applicability.

In this paper, we propose a different approach. Instead of learning the age across two expressions, we jointly learn the age and expression and model their relationship. The aim is to achieve expression-invariant age estimation. In our approach, one model is learnt for all expressions. To predict the age, the age and expression are inferred jointly, and hence prior-knowledge of the expression of the test face is not required. More specifically, we introduce a new graphical model which contains a latent layer between the age/expression labels and the facial features. This layer captures the relationship between the age and expression. During training, the age and expression variables are observed. This allows the latent layer to learn the configurations which map the features to the age for different expressions and thus obtaining expression-invariant age estimation. For testing, the age and expression labels are unknown and the method finds the values of age, expression and latent layer which together maximize their compatibility with the features.

The contributions of our work are: 1) we show how age-expression joint learning improves the age prediction compared to learning independently from expression. 2) As opposed to existing methods, the proposed method predicts the age across different facial expressions without prior-knowledge of the expression labels of the test faces. 3) Finally, our results outperform the best reported results on age-expression datasets (FACES and Lifespan).

3.2 Algorithm

The proposed graphical model aims to jointly learn the relationship between age and expression. To this end, an inter-connected latent layer is introduced. The latent variables encode the changes in face appearance. These variables are not explicitly defined, but learnt from the training data.

The graphical model has four sets of connections: First, connections between the face subregions and the latent variables. These connections are designed to capture the

changes of face appearance related to age and expression. Second, connections between the face subregions and the age/expression labels are formed. The aim here is to directly infer the age/expression from the features. Third, connections between the latent variable modeling the relationship between the face subregions. Finally, connections are established between the latent variables, the age, and the expression. The last type of connections is designed to relate the age with the expression which allows the joint learning between them. Next, we discuss the model formulation and explain the inference and learning techniques.

3.2.1 Model Formulation

Suppose we have N training samples (images) $\{\mathbf{s}_1 = (\mathbf{x}_1, \mathbf{y}_1), \mathbf{s}_2 = (\mathbf{x}_2, \mathbf{y}_2), ..., \mathbf{s}_N = (\mathbf{x}_N, \mathbf{y}_N)\}$ where $\mathbf{x}_n$ represents the features for sample $\mathbf{s}_n$ and $\mathbf{y}_n = \{y_{a,n}, y_{e,n}\} \in \mathcal{Y} = \mathcal{A} \times \mathcal{E}$ denotes the age and the expression labels. $\mathcal{A}$ and $\mathcal{E}$ are the age and the expression spaces, respectively. The image is uniformly divided into four (2×2) subregions. The feature vector extracted from each sub-region x_i is connected to the corresponding hidden variable h_i. Hence, the sample feature vector consists of four subregion vectors $\mathbf{x}_n = [x_1, x_2, x_3, x_4]$ and the corresponding latent layer is denoted by $\mathbf{h}_n = [h_1, h_2, h_3, h_4] \in \mathcal{H}^4$, where $\mathcal{H}$ is the space of the latent variable state.

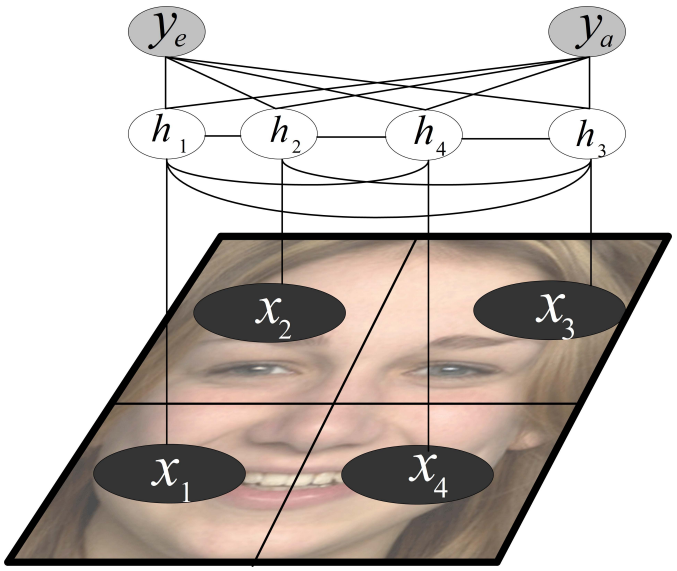


Fig. 3.1: Our graphical model to jointly learn the age and the expression. $\mathbf{x}$ represents the feature vector, $\mathbf{h}$ denotes the latent variables, y_a and y_e are the corresponding age and expression respectively. Note that, while all x_i are connected with y_a and y_e, we do not show these connections in this figure for the sake of clarity.

The aim is to learn the mapping between the features **x** and labels **y**. Our model maximizes the conditional probability of the joint assignment of **y** given observation **x**:

$$\mathbf{y}^* = argmax_{\mathbf{y}} P(\mathbf{y}|\mathbf{x};\theta). \tag{3.1}$$

Where:

$$P(\mathbf{y}|\mathbf{x};\theta) = \sum_{\mathbf{h}\in\mathcal{H}} P(\mathbf{y},\mathbf{h}|\mathbf{x};\theta) = \frac{\sum_{\mathbf{h}\in\mathcal{H}} exp(\psi(\mathbf{y},\mathbf{h},\mathbf{x};\theta))}{\sum_{\mathbf{y}'\in\mathcal{Y},\mathbf{h}\in\mathcal{H}} exp(\psi(\mathbf{y}',\mathbf{h},\mathbf{x};\theta))}.$$

Where $\psi(.)$ is the potential function which measures the compatibility between the (observed) features, the joint assignment of the latent variables, and the output labels. In the next section, the potentials are defined.

3.2.2 Potentials

The potentials measure the compatibility of the joint assignment of different variables:

$$\psi(\mathbf{y},\mathbf{h},\mathbf{x};\theta) = \sum_{i=1}^{4} \psi_1(y_a,x_i;\theta_i^1) + \sum_{i=1}^{4} \psi_2(y_e,x_i;\theta_i^2) + \sum_{i=1}^{4} \psi_3(h_i,x_i;\theta_i^3) + \psi_4(\mathbf{h},y_a,y_e;\theta^4). \tag{3.2}$$

In our model, four types of potentials are used. Hereafter, we explain each one of them. Potential ψ_1 models the compatibility of the features and the age:

$$\psi_1(y_a,x_i;\theta_i^1) = \theta_i^1 \phi_1(y_a,x_i), \tag{3.3}$$

where $\phi_1(y_a,x_i)$ represents the feature mapping function encoding the features of the joint assignment of y_a and x_i. The length of $\phi_1(y_a,x_i)$ is equal to the length of x_i multiplied by the cardinality of y_a. In case there are S different ages and the feature vector x_i has K features, the size of θ_i^1 will be $S \times K$. The mapped feature vector is given by:

$$\phi_1(y_a,x_i) = [\underbrace{0...0}_{K\times(y_a-1)dimension} \quad x_i^T ...0]. \tag{3.4}$$

The model turns into a multi-class SVM for age estimation when solely this potential is utilized with the maximum margin method. Multi-class SVM is used as a baseline in this paper. This potential models the global mapping between the input features and the output age prediction.

Potential ψ_2 models the compatibility of the features and the expression:

$$\psi_2(y_e,x_i;\theta_i^2) = \theta_i^2 \phi_2(y_e,x_i), \tag{3.5}$$

where $\phi_2(y_e, x_i)$ encodes the features of the joint assignment of y_e and x_i and is defined in the same way as in equation 3.4.

Potential ψ_3 models the compatibility of the observation and the latent states:

$$\psi_3(h_i, x_i; \theta_i^3) = \theta_i^3 \phi_3(h_i, x_i). \tag{3.6}$$

Here, $\phi_3(h_i, x_i)$ encodes the features of the joint assignment of the latent variable h_i and the features x_i. The latent variables capture the changes of face appearance. For example, a hidden state could represent whether the mouth is open (e.g. happy) or frowning (e.g. angry). Thus, the potential $\psi_3(h_i, x_i; \theta)$ learns the mapping of the observed features to the appearance changes.

The potential ψ_4 models the compatibility between the age, the expression, and the latent layer:

$$\psi_4(\mathbf{h}, y_a, y_e; \theta^4) = \theta^4 \phi_4(\mathbf{h}, y_e, y_a). \tag{3.7}$$

$\phi_4(\mathbf{h}, y_e, y_a)$ represents the feature mapping function which encodes the features of the joint assignment of **h**, y_e and y_a. The length of $\phi_4(\mathbf{h}, y_e, y_a)$ is the multiplication of the cardinalities of **h**, y_e and y_a. The element corresponding to the assignment of **h**, y_e and y_a is set to be 1 while all other elements are set to be 0.

3.2.3 Inference and Learning

Inference: Given the model parameters θ, the inference involves a combinatorial search of the joint assignment of **h**, y_e and y_a which results in the maximum conditional probability:

$$(\hat{\mathbf{y}}, \hat{\mathbf{h}}) = argmax_{\mathbf{y} \in \mathcal{Y}, \mathbf{h} \in \mathcal{H}} \psi(\mathbf{x}, \mathbf{y}, \mathbf{h}; \theta). \tag{3.8}$$

Since the proposed graphical model contains loops, it is intractable in general to perform the maximization. However, by collapsing all the latent variables **h** with the output variables y_e a new potential factor is obtained. In the same way, by collapsing all the latent variables **h** with the output variables y_a we get another new potential factor. Then the model becomes a chain structure and dynamic method is used to solve the maximization problem [80].

Learning: To learn the parameters θ, we exploit the max margin approach [121]. Since the latent variables **h** are not labeled in the training set, we need to solve the following

latent structure SVM problem:

$$\min_{\theta,\xi} \left\{ \frac{1}{2}\|\theta\|^2 + C\sum_{i=1}^{N} \xi_i \right\} \tag{3.9}$$
$$s.t.\ \forall i \in \{1,2,\ldots,N\}, \forall \mathbf{y}, \forall \mathbf{h} \in \mathcal{H}:$$
$$\xi_i \geq \Delta(\mathbf{y}_i, \mathbf{y}) + \psi(\mathbf{y}, \mathbf{h}, \mathbf{x}_i; \mathbf{w}) - \psi(\mathbf{y}_i, \mathbf{h}_i^*, \mathbf{x}_i; \theta).$$

Where $\psi(.)$ is the potential function as described in equation 3.2. $\mathbf{h}_i^*$ is the optimum state under the current parameter. The loss function $\Delta(\mathbf{y}_i, \mathbf{y})$ is defined as the following:

$$\Delta(\mathbf{y}, \hat{\mathbf{y}}) = \begin{cases} |y_a - \hat{y}_a| & if\ y_e = \hat{y}_e \\ 1 + |y_a - \hat{y}_a| & if\ y_e \neq \hat{y}_e \end{cases}. \tag{3.10}$$

This optimization problem is non-convex. Following [135], we use the CCCP concave-convex framework [137] to solve it. More details of the CCCP procedure can be found in [135, 137].

3.3 Experiments

The goal of the proposed approach is to capture the relationship between the age and expression and, hence, alleviate the influence of expression in age estimation. In this section, we conduct a number of experiments to validate our model using the age-expression datasets FACES [30] and Lifespan [78].

3.3.1 Datasets

The publicly available age estimation datasets like FG-NET [1] and MORPH [93] contain mostly neutral faces. The non-neutral faces in those datasets are not expression-labeled. Therefore, to evaluate expression-invariance age estimation, we use other datasets: FACES and Lifespan, which are recently introduced to the computer vision community [54]. FACES dataset contains face images of 171 subjects showing 6 basic expressions: neutrality, happiness, anger, fear, disgust, and sadness. Every subject shows all the expressions resulting in $1026 = 171 \times 6$ face images. The faces in the dataset are frontal with fixed illumination mounted in front and above of the faces. The ages of the subjects range from 19 to 80. The age distribution is not uniform and in total there are 37 different ages.

The Lifespan dataset is a collection of faces of subjects from different ethnicities showing different expressions. The expression subsets have the following sizes: 580, 258, 78,

64, 40, 10, 9, and 7 for neutrality, happiness, surprise, sadness, annoyed, anger, grumpy, and disgust, respectively. The ages of the subjects range from 18 to 93 years and in total there are 74 different ages. The dataset has no labeling for the subject identities. We follow the setup of [54, 139] and use the neutral and the happy subsets. Although the age distributions of both datasets cover a wide range of ages, the FACES dataset is more challenging for age prediction since its expression variation (six expressions) is larger than the one in Lifespan dataset (two expressions).

For feature extraction, eye centers are first automatically detected and the faces are registered and cropped. Then, the faces are divided into 8×8 patches and a local feature vector is extracted for each patch. Finally, the patch local descriptors are concatenated together to form the face descriptor. To extract the features from each patch, we use Local Binary Pattern (LBP) [83]. It is a simple, efficient, and rotation-invariant approach and successfully used for age prediction to capture the skin texture details [21, 133]. In our experiments, we use 8 sampling points with a radius equal to 1.

As in previous setups [54, 139], the datasets are divided into 5 folds. For the FACES dataset, the expression distributions are uniform for all the 5 folds, and none of the subjects appears in more than one fold. For the Lifespan dataset, the dataset (neutral and happy) is split randomly into 5 folds. As the subject identities are not available, a subject overlap between the training and the test samples is possible. The results are measured quantitatively by Mean Absolute Error (MAE) $\frac{1}{N}\sum_{n=1}^{N}|y_a^n - \hat{y}_a^n|$. Where y_a^n is the true age for the test sample n, $\hat{y}_a^n$ is the predicted age for the test sample n, and N is the number of the test samples.

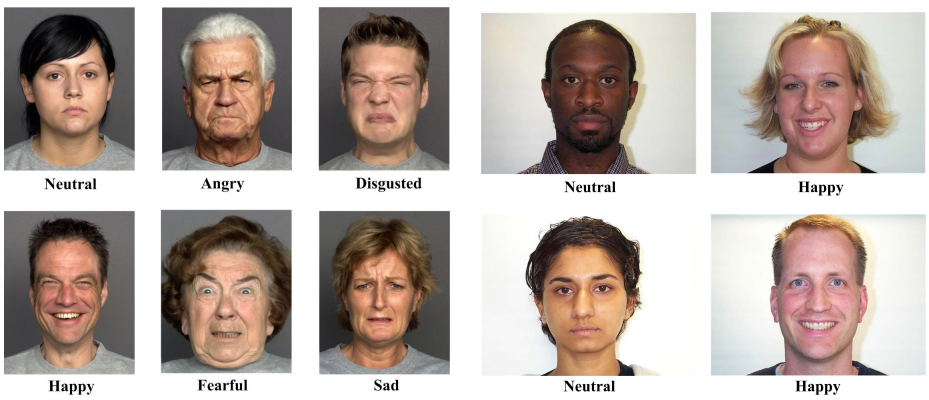


Fig. 3.2: Example faces from FACES (left) and Lifespan (right) datasets.

3.3.2 Expression-Invariant Age Estimation

In this experiment, we compare two cases. First, learning the age independently from the expression. Second, learning the age jointly with the expression. In both cases, the same 5-fold age-expression datasets are used for evaluation. For the expression-independent learning, a multi-class SVM is used as a baseline. In the expression-joint learning, we use the proposed graphical model and the number of hidden states $|\mathcal{H}|$ is set to 3. For the model learning, the expression is observed and the potential function in equation 3.2 is applied. The results for the proposed model are shown in Table 3.2. For both datasets, our graphical model significantly reduces the prediction error (14.43% for FACES and 37.75% for Lifespan). The errors reported in [54] and [139] for FACES and Lifespan datasets are shown in Table 3.2. Although both methods assume prior-knowledge of the expression of tested samples, our model outperforms their results for the two datasets.

We further compare our age estimation approach with the joint classification method by [52]. The method was proposed to recognize facial expressions while reducing the influence of human aging. In their method, the authors simply divide the dataset into four age groups ([18-29],[30-49],[50,69], and [70-94]) and consider each expression within each age group as a new class. Then, classification is performed on the newly defined classes. For facial feature extraction, they manually labelled 31 fiducial points and applied Gabor filters [24] on the locations of those points. The four age group classification accuracy using the joint learning method is reported.

To make a fair comparison, and since the authors [52] manually labeled 31 fiducial points on the face, we use our features and compare only the joint learning methods. To this end, we create a new class for each age/expression combination. Different from [52], where the datasets are divided into four age groups, we consider each age separately. The total number of new classes is $37 \times 6 = 222$ in the FACES dataset and $74 \times 2 = 148$ classes in the Lifespan dataset. The obtained errors for FACES and Lifespan are 9.94 and 8.85 years respectively, which are higher than the baseline errors and the ones obtained by our graphical model. It is worth mentioning that in [52], as the datasets are divided into four age groups, the method is tested on smaller number of "joint-classes" (24 and 8 for FACES and Lifespan respectively). In this experiment, the number of joint-classes is much higher.

Detailed results for independent and joint learning for FACES and Lifespan datasets are shown in Tables 3.2 and 3.3, where the error for each expression subset is shown separately. The error is reduced for all expression subsets, however, in different rates. The largest improvement is achieved for neutrality (with 30.13% error reduction), while the smallest improvement is obtained for the anger and the disgust expressions (4.64% and 2.43% respectively). This is explained as anger and disgust expressions induce

more profound changes in the face appearance than the other expressions which make age prediction/perception more difficult. Our model clearly outperforms the existing methods [52, 54, 139] by a wide margin which further proves the effectiveness of our approach.

The hidden states capture the changes in the face appearance. To further illustrate this point, we show the face regions corresponding to each hidden state. More specifically, the averages of the bottom and top regions are computed (Figure 3.3). For the bottom regions, the first hidden state corresponds to the face appearance where the mouth is open, the third hidden state represents a depressed lip corner, and the second hidden state corresponds to a normal face appearance. For the top regions, the second hidden state represents the face appearance where the eye is slightly closed while the first and the third states correspond to open eye appearances.

Fig. 3.3: Average face regions corresponding to different hidden states (from left to right) for the bottom and top face regions.

3.3.3 Joint-Learning for Expression Recognition

In this experiment, we consider a different, yet related, task: how age information can improve the recognition of expressions. Although aging affects how people exhibit expressions, much of automatic expression recognition methods do not use the age of the subject to recognize expressions. This is mainly due to the lack of expression datasets with a sufficiently large age range. Motivated by the introduction of recent age-expression datasets, Guo et al. [52] recently proposed a method to recognize facial expressions while reducing the influence of human aging.

We apply our model on FACES and Lifespan datasets to recognize the expression. The results are shown in Table 3.4. Our method improves the expression recognition perfor-

Tab. 3.1: Expression-independent and expression-joint learning are evaluated on FACES and Lifespan datasets. The results show clear improvement of performance when the age is learnt jointly with the expression and the age prediction error is reduced by 14.43% and 37.75% for FACES and Lifespan datasets respectively. The results of the methods [54] and [139] along with the results using the joint learning method [52] are compared with ours. Our model obtains the best performance for both datasets with a large margin. Note that [54] and [139] assume that the expressions of the tested sample is a prior-knowledge while our model has no such requirement. The last column shows the difference in error when using the joint learning in comparison with independent learning.

Dataset	[54]	[139]	[52]	Indep-Learn	Joint-Learn	Reduc-Rate %
FACES	9.12	8.33	9.94	8.66	**7.41**	14.43%
Lifespan	6.63	6.23	8.85	8.45	**5.26**	37.75%

Tab. 3.2: Age estimation error for each expression subset on the FACES dataset. The error is reduced for all expressions using the expression-joint learning. The largest error reduction is achieved for neutral faces (30.13%) while the smallest error reduction is obtained for anger and disgust (4.64% and 2.43% respectively).

Test Data	Indep-Learn	Joint-Learn	Reduc-Rate %
Neutrality	8.54	**5.97**	30.13
Anger	8.61	**8.21**	4.64
Disgust	8.37	**8.17**	2.43
Fear	9.79	**8.25**	15.71
Happiness	8.42	**6.77**	19.58
Sadness	8.17	**7.07**	13.44
Average	8.66	**7.41**	14.43

Tab. 3.3: Age estimation error for each expressions subset on the Lifespan dataset. The error is reduced for both neutrality and happiness expressions. Note that, since the numbers of happy and neutral faces are not equal, the weighted average is computed.

Test Data	Indep-Learn	Joint-Learn	Reduc-Rate %
Neutrality	8.66	**5.72**	33.94
Happiness	7.96	**4.14**	47.91
Average	8.45	**5.26**	37.75

mance for the FACES dataset by 2.38%. However, the accuracy on the Lifespan dataset is comparable to the one acquired by independent learning. This maybe explained by the observation that there are only two expressions in Lifespan compared to six ones in FACES, and hence the expression variation within Lifespan dataset is smaller than it is

Tab. 3.4: Expression recognition using age-joint and age-independent learning evaluated on FACES and Lifespan datasets. Joint-learning improves the accuracy by 2.38% on the FACES dataset while the accuracy on the Lifespan dataset is comparable. The method in [52] is further tested on our features, and the results show degrading in the performance for both datasets.

Dataset	Indep-Learn %	Joint-Learn %	[52] %
FACES	90.05	**92.19**	84.68
Lifespan	93.91	93.68	91.05

within the FACES dataset. Consequently, the margin of improvement is smaller for the Lifespan dataset and the joint learning method obtains comparable accuracy.

We compare the proposed method with the one in [52]. As the authors manually labeled 31 fiducial points on the face and extracted the features using their locations, a direct comparison of the results will not be fair. Thus, we test the method in [52] using our features. The datasets are divided into the same four age groups ([18-29],[30-49],[50,69], and [70-94]). Then, a new class is created for each expression/age group combination resulting in 24 and 8 new classes for the FACES and Lifespan dataset respectively. The obtained accuracy (see Table 3.4) is lower than the one acquired by our model.

3.4 Discussion

The results obtained using our graphical model show the strength of joint-learning to alleviate the influence of facial expression in age prediction. Some existing works [54, 139] approached age prediction with variant facial expressions. Our method is different in two aspects: First, in our model, the age is jointly learnt with all expressions instead of learning the cross-expression for two expressions at a time. This property allows our model to be extended to a broader group of tasks where the changes are not restricted to the basic (profound) expressions. For example, the changes can be described by a group of smaller units (e.g. action units [33]). These changes can describe various face (undefined) expressions. In such cases, the hidden layer will learn the relationship between the age and multiple variables (action units) instead of one variable (expression) at a time. Moreover, beside facial expressions, other attributes can be learnt collectively within the proposed graphical model such as gender and race. Second, the proposed approach does not require the expression labels of the test samples to be known while the existing methods [54, 139] assume prior-knowledge of the expressions.

3.5 Conclusions

In this paper, an expression-invariant age predictor is proposed by jointly learning the age and the expression. We introduce a graphical model with a latent layer to learn the relationship between the age and the expression. This layer is designed to capture the changes in the face which induce the aging and the expression appearance.

Conducted on two age-expression datasets (FACES and Lifespan), our experiments show the improvement in performance when the age is jointly learnt with the expression in comparison to expression-independent age estimation. The age estimation error is reduced by 14.43% and 37.75% for FACES and Lifespan datasets respectively. Furthermore, using our model, without prior-knowledge of the expressions of the test faces, the acquired results are better that the best reported ones for both datasets.

Acknowledgment

This research is partly supported by NICTA. NICTA is funded by the Australian Government as represented by the Department of Broadband, Communications, and the Digital Economy, and the Australian Research Council (ARC) through the ICT Centre of Excellence Program.

4

Combining Facial Dynamics with Appearance for Age Estimation

4.1 Introduction

Age estimation from human faces is a challenging problem with applications in forensics, security, biometrics, electronic customer relationship management, entertainment and cosmetology [4, 38, 89]. Automatic age estimation can augment many computer applications in these domains, but it can also be used as a stand-alone tool, since humans are not universally successful in estimating age. The most frequently used measure of age estimation is the mean absolute error (MAE), and a recent crowd-sourcing study performed with frequently used aging databases show that humans have a MAE of 7.2–7.4 years for estimating the age of a person over 15, depending on the database conditions [56].

The main challenge of age estimation is the heterogeneity in facial feature changes due to aging for different humans. To determine facial changes associated with age is a hard problem, because they are related not only to gender and to genetic properties, but also to a number of external factors such as health, living conditions and weather exposure. Gender can play a role in the aging process as there are differences in aging patterns and features in males and females. Furthermore, facial cosmetics, surgical operations, the presence of scars, and even the presence of facial hair can be mitigating factors for age estimation.

Age estimation is an active topic today due to the growing necessity of including this information in real-world systems. This necessity comes from the fact that age is impor-

tant to understand requirements or preferences in different aspects of the daily life of a person. Systems implementing age specific human computer interaction can cope with these aspects. Some examples are biometric systems that filter their database for the estimated age range of a subject, vending machines capable of denying some products such as alcohol or cigarettes to an underage customer, or advertisements in different automated environments (web pages, displays in stores, etc.) that can be personalized according to the age of the individual interacting with the system.

Automatic facial age estimation is affected by the traditional factors that make face analysis difficult in general. Unknown illumination conditions, non-frontal facial poses, and presence of facial expressions, are some issues that such systems need to deal with. Especially, facial expressions might negatively affect the accuracy of automated systems: When a person smiles, for instance, wrinkles are formed and these can be misleading when only the appearance cues are taken into account [73]. Similarly, sagging of the face in a sad expression can resemble the effects of aging.

The most important cues that are used in age classification are appearance-based, most notably the wrinkles formed on the face due to deformations in skin tissue. For this reason, current systems mainly focus on static appearance features of the face, as it is the easiest way to obtain satisfactory results [38]. Hence, the dynamics of facial movement are largely ignored.

In this paper, instead of only considering static appearance features, we explore a novel set of dynamic features for age estimation. As movement features can be observed from facial expressions, the aim is to use dynamic features derived from these facial expressions for estimating the age. Since the smile is one of the most frequently used facial expressions, as well as the easiest emotional facial expression to pose voluntarily [31], we first focus on smiles and analyze the discrimination power of smile dynamics for age estimation. Once we verify that smile dynamics can improve discrimination, we validate the effectiveness of the proposed approach on a different facial expression.

There are a number of changes that happen on the face with aging, including loss of muscle tone, loss of underlying fat tissue, which reduces the smoothness of the face and creates wrinkles, receding gums (and sometimes, missing teeth), increased crows feet around the eyes, sunken eyes as a consequence of fat from eyelids settling into eye sockets, texture changes like blotches, dark spots, bone mass reduction causing lower jaw to reduce in size, and cartilage growth to lengthen the nose [12, 77, 97]. All these morphological changes alter the overall appearance of an expression on the face, but especially the loss of muscle tone directly affects the dynamics, along with the appearance. It is well known that the elastic fibers on the face show fraying and fragmentation at advancing age [98]. By leveraging the deformation features of the facial surface patches, age estimation with dynamic features may improve over systems

that use solely appearance-based cues.

The main contribution of this paper is to show, on multiple expressions, that expression dynamics can be used to better estimate the age of a person. We propose a fully automatic age estimation framework, and show that it significantly outperforms the generic approach. We also introduce the high-resolution UVA-NEMO Disgust Database, which we make publicly available. We report our results with smile and disgust expressions, and make our experimental protocols available.

We extend our previous study [26] in many ways. Apart from a more in-depth treatment and extended literature, (1) we use 3D volume changes via surface patches instead of landmark movements, (2) we add frequency and facial asymmetry descriptors to the feature set, (3) we use a two-level adaptive classification scheme, (4) we evaluate four appearance features, (5) we systematically analyze gender-specific and spontaneity-specific effects of aging features, (6) we introduce an adaptive grouping procedure, (7) we introduce a new public database for disgust expression and report results on it.

The next section introduces related work in age estimation. Since there are comprehensive surveys in this area ([38, 89]), we focus on the most successful approaches, and the most recent work. Section 4.3 describes the proposed system of age estimation. In particular, we describe the detection and tracking of facial landmarks, the set of dynamic features, and the two-level classification scheme. Section 4.4 describes the experimental protocol and the UvA-NEMO Smile Database, as well as introducing the new UVA-NEMO Disgust Database. Section 4.5 reports extensive comparative results. We analyze the contribution of appearance and dynamic features in detail, selecting four different state-of-the-art appearance-based approaches to serve as baselines. We test the influence of different facial regions in age estimation, augment the method by using gender-specific analysis, and study the effects of spontaneity in facial expressions. It is followed by a discussion in Section 4.6. Section 4.7 concludes the paper.

4.2 Related Work

Several works propose to determine facial pattern changes and evolution associated with the aging process, both from psychological and biological points of view. These studies are mostly aimed at age synthesis, i.e. changing the appearance of a rendered face to show proper effects of aging. Some of these works are useful in the determination of appropriate facial features for age estimation. For instance, O'Toole *et al.* [86] use 3D models of faces to apply caricaturing processes in order to describe age variations between samples. Wu *et al.* [129] develop a system for the simulation of wrinkles and skin aging for facial animation. Suo *et al.* [113] present a model for face aging by ana-

lyzing it as a Markov process through a graph representing different age groups. Tiddeman *et al.* [119] also develop prototype models for face aging using texture information. In [84], a quantitative approach to face evolution of aging is presented.

The results of these studies show that the craniofacial development and skin texture are the most important features for age estimation. In fact, one of the first approaches for age estimation is proposed by Kwon and Lobo [67], where individual faces are classified into three age groups (baby, young and senior). This classification is performed using the theory of craniofacial development [5] and facial skin wrinkle analysis. Lanitis *et al.* [68] propose an age estimation method based on regression analysis of the *aging function*. During the training procedure, a quadratic function of facial features is fitted to each individual in the training set as his/her aging function. As for age estimation, they propose four approaches to determine the proper aging function for the unseen face image. The Weighted Person Specific (WPS) approach achieves the best performance in the experiments. This function, however, relies on profiles of the individual containing external information such as gender, health, living style, etc.

Image processing methods, including tools for subspace learning and dimensionality reduction, are also used to automatically estimate the age. In [51], faces are projected into manifolds by using subspace learning followed by a regression model to estimate the age. The aging pattern subspace (AGES) method [46] models a sequence of individually aging face images by learning a subspace representation. The age of a test face is determined by the projection in the subspace that can reconstruct the face image best. This model is later extended by the authors to model the nonlinear nature of human aging by considering learning of nonlinear subspaces, using a model called KAGES (Kernel AGing pattErn Subspace) [43]. Zhan *et al.* [138] propose an extended non-negative matrix factorization method to learn a subspace representation, which could recover age information while eliminating variations caused by identity, expression, pose, etc. Chen *et al.* propose a method that employs pairwise age ranking based on subspace learning for age prediction [20]. In their approach, age ranks from unlabeled data are incorporated by semi-supervised learning. [18] applies age-oriented local regression using distance metric learning and dimensionality reduction.

Feature extraction is one of the key issues of automatic age estimation. In [53], Guo *et al.* introduce biologically-inspired aging features (BIF) for age estimation. These features are based on Gabor filter responses for different orientations and scales. Alnajar *et al.* propose intensity- and gradient-based features to adopt a learning-based encoding method for age estimation under unconstrained imaging conditions [7]. For each pixel, neighboring pixels are sampled in a ring-based pattern to form a low-level feature vector. Then, the features are encoded using a PCA-tree-based codebook. [134] models the completed local binary patterns (CLBP) using an SVM regressor. Initially, the method fine-tunes facial alignments in terms of facial shape and pose. The similarity

transformation is based on local binary pattern distributions.

Aging patterns show significant differences in young and elderly people, and human performance in age estimation shows differences for these groups. It seems possible to break the age estimation problem into simpler subproblems by adopting different strategies for different age groups. In [70], fuzzy age labels (human annotations) are used in combination with the real age labels to train an age estimation system. Fuzzy age labels are defined as the upper and lower bounds of human estimation. Hybrid constraint supported vector regression is proposed to model both deterministic and fuzzy labels. In [56], a hierarchical age estimation is proposed. It classifies each facial component into one of four disjoint age groups using an SVM-based binary decision tree. For each age group, a separate SVM regressor is trained to fine-tune the age prediction. Then, outputs for different components using different features are fused to estimate the final age.

Age estimation and expression recognition are rarely coupled, although several systems rely on similar features and classification paradigms for both problems. In [54], an age estimation method is proposed to cope with significant expression changes, using correlation learning and discriminant mapping. However, this methodology requires both neutral and expressive facial images for the same subject, since it is based on the correlation between pairs of expressions. More recently, Zhang and Guo propose a weighted random subspace method to deal with expression changes by improving the discriminative power of the aging features [139].

Remarkably, the temporal dynamics of faces have been ignored in age estimation. Until [26], the precursor of the present work, the only study is by [55] in which Hadid proposes to use volume LBP (VLBP) features to describe spatio-temporal information in videos of talking faces and classifies the ages of the subjects into five groups (child, youth, adult, middle-age, and elderly). However, VLBP features alone are not powerful enough and the proposed system could not reach the accuracy of static image-based age estimation. Therefore, we propose to use facial dynamics and explore the potential for obtaining useful cues from facial expressions which have been unexplored so far.

4.3 Method

The aim of the proposed method is to estimate the age of subject by using a sequence of images that show the subject displaying a facial expression as input. To this end, we focus on the smile expression, since it is one of the most frequently shown facial expression. Additionally, disgust expression is considered to evaluate the reliability and generalizability of the approach.

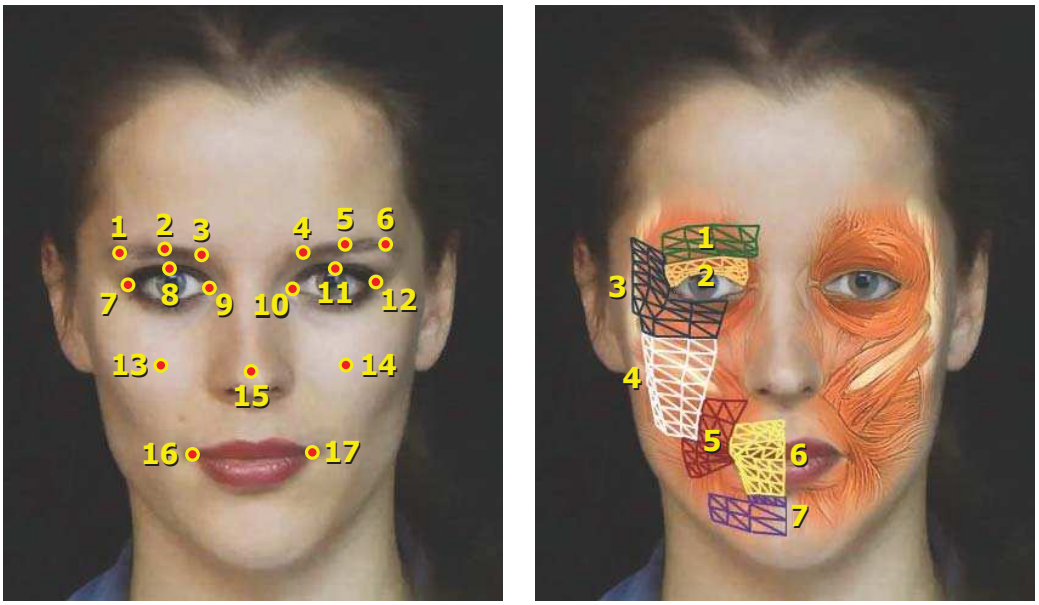
The proposed approach combines appearance features with facial expression dynamics. The method assumes that the input video starts with a moderately frontal face, and has the entire duration of a smile (or disgust) expression. These are typical assumptions of video-based expression recognition approaches.

The flow of the system is summarized as follows. Initially, a mesh model is fitted to face using 17 fiducial points, and tracked during the rest of the video. The surface deformations on different regions are computed using the tracked mesh points. Temporal phases (onset, apex, and offset) of the expression are estimated using the mean displacement signal of the lip corners. Then, dynamic features for each regional patch are extracted from each phase. Appearance features are extracted using the first frame of the onset phase, in which the face is neutral. After a feature selection procedure, the most informative dynamic features are selected and fused with appearance features to train Support Vector Machine (SVM) classifiers/regressors. In the rest of this section, we will outline the different components of our approach in detail.

4.3.1 Smile and Disgust Expressions

In this paper, we extract appearance and dynamic features from smile and disgust videos. In general, a smile can be modeled as the upward movement of the mouth corners, which corresponds to Action Unit 12 (AU12) in the facial action coding system (FACS) [33]. In terms of anatomy, the *zygomatic major* muscle contracts and raises the corners of the lips during a smile [34]. On the other hand, the disgust expression is the display of intense displeasure or condemnation that is shown by narrowing eyebrows, curling upper lip, and wrinkling nose. In terms of dynamics, smile and disgust expressions are composed of three non-overlapping phases; the onset, apex, and offset, respectively. Onset is the initial phase of a facial expression and it defines the duration from neutral to expressive state. Apex phase is the stable peak period of the expression between onset and offset. Likewise, the offset is the final phase from expressive to neutral state.

According to Ekman [31], there are many smiles, which are different in terms of their appearance and meaning. Ekman identified 18 of them (such as enjoyment, fear, miserable, embarrassment, listener response smiles) by describing the specific visual differences on the face and indicating the accompanying action units [31]. In this paper, we focus on enjoyment smiles for the detailed analysis, because they are frequently shown and can easily be induced. Subsequently, we use posed and spontaneous smiles. To test whether the approach generalizes to other expressions, we use posed disgust expressions.



(a) (b)

Fig. 4.1: (a) Used facial feature points with their indices. (b) Regional surface patches (with their indices) and their underlying facial muscle structure. For simplicity, patches are shown on a single side of the face.

4.3.2 Facial Feature Tracking and Alignment

To analyze facial dynamics, surface deformations of seven facial regions (eyebrow, eyelid, eye-side, cheek, mouth-side, mouth, chin) are tracked in the videos [see Fig. 4.1(b)]. Patches for these regions are initialized in the first frame of the videos, using automatically detected 17 landmarks (corners and center of eyebrows, eye corners, center of upper eyelids, nose tip, and lip corners) for precise tracking and analysis [see Fig. 4.1(a)]. For automatic facial landmark detection, the method proposed by Dibeklioğlu *et al.* [28] is used. This method models Gabor wavelet features of a neighborhood of the landmarks using incremental mixtures of factor analyzers and enables a shape prior to ensure the integrity of the landmark constellation. It follows a coarse-to-fine strategy in which landmarks are initially detected on a coarse level and then fine-tuned for higher resolution. To track the facial features and pose, we use a piecewise Bézier volume deformation (PBVD) tracker, originally proposed by Tao and Huang [118].

The PBVD tracker employs a model-based approach. A 3D mesh model of the face [see Fig. 4.1(b)] is constructed by warping the generic model to fit the facial features in the first frame of the image sequence. The generic face model consists of 16 surface patches. To form a continuous and smooth model, these patches are embedded in Bézier

volumes. If $x(u, v, w)$ is a facial mesh point, then the Bézier volume is defined as:

$$x(u,v,w) = \sum_{i=0}^{n}\sum_{j=0}^{m}\sum_{k=0}^{l} b_{i,j,k} B_i^n(u) B_j^m(v) B_k^l(w), \tag{4.1}$$

where points $b_{i,j,k}$ and variables $0 < \{u, v, w\} < 1$ control the shape of the volume. $B_i^n(u)$ denotes a Bernstein polynomial:

$$B_i^n(u) = \binom{n}{i} u^i (1-u)^{n-i}. \tag{4.2}$$

After fitting the face model, facial feature points (as well as head motion) are tracked in 3D according to the movement and the deformations of the mesh. To measure 2D motion, template matching is used between frames at different resolutions. The estimated 2D image motion is modeled as a projection of the 3D movement onto the image plane. Then, the 3D movement is calculated using the projective motion of several points.

The tracked 3D coordinates of the facial feature points ℓ_i [see Fig. 4.1(a)] are used to align the faces in each frame. We estimate the 3D pose of the face, and normalize the face with respect to roll, yaw, and pitch rotations. Since three non-collinear points are enough to construct a plane, we use three stable landmarks (eye centers and nose tip) to define a plane $\mathcal{P}$. Eye centers are defined as middle points between inner and outer eye corners and denoted by $c_1 = \frac{\ell_7+\ell_9}{2}$ and $c_2 = \frac{\ell_{10}+\ell_{12}}{2}$. Angles between the positive normal vector $\mathcal{N}_\mathcal{P}$ of $\mathcal{P}$ and unit vectors U on X (horizontal), Y (vertical), and Z (perpendicular) axes give the relative head pose as follows:

$$\theta = \arccos \frac{U.\mathcal{N}_\mathcal{P}}{\|U\| \, \|\mathcal{N}_\mathcal{P}\|}, \text{ where } \mathcal{N} = \overrightarrow{\ell_{15}c_2} \times \overrightarrow{\ell_{15}c_1}. \tag{4.3}$$

In Equation 4.3, $\overrightarrow{\ell_{15}c_2}$ and $\overrightarrow{\ell_{15}c_1}$ denote the vectors from point ℓ_{15} to points c_2 and c_1, respectively. $\|U\|$ and $\|\mathcal{N}_\mathcal{P}\|$ are the magnitudes of U and $\mathcal{N}_\mathcal{P}$ vectors. According to the face geometry, Equation 4.3 estimates the roll (θ_z) and yaw (θ_y) angles of the face with respect to the camera. However, the estimated pitch (θ_x) angle is subject-dependent, since it is relative to the constellation of the eye corners and the nose tip. If we assume that the face is approximately frontal in the first frame, then the actual pitch angles (θ'_x) are calculated by subtracting the initial value. Once the pose of the head is estimated, tracked points are normalized with respect to rotation, scale, and translation by:

$$l_i = \left[\ell_i - \tfrac{c_1+c_2}{2}\right] R_x(-\theta'_x) R_y(-\theta_y) R_z(-\theta_z) \tfrac{100}{\rho(c_1,c_2)}, \tag{4.4}$$

where l_i is the aligned point and R_x, R_y, and R_z denote the 3D rotation matrices for the given angles. ρ denotes the Euclidean distance between the given points. On the normalized face, the middle point between eye centers is located at the origin and the inter-ocular distance (distance between eye centers) is set to 100 pixels.

4.3.3 Dynamic Features

To analyze the dynamics of facial deformations during an expression, we extract a set of dynamic features from seven different patches on the face [see Fig. 4.1(b)]. These patches are defined based on the underlying facial muscle structure, since the direction and the length of such muscles cause the visual variations of expressions [95].

When the tracked points are normalized, onset, apex, and offset phases of the smile and disgust expressions are detected, using the approach proposed by Schmidt *et al.* [99], by calculating the amplitude of the expression as the distance of the right lip corner to the lip center during the expression. Since the faces are normalized, the lip center is calculated only once in the first frame. Differently from [99], we estimate the expression amplitude as the mean amplitude of right and left lip corners, normalized by the length of the lip. Let $\mathcal{D}_{\text{lip}}(t)$ be the value of the mean amplitude signal of the lip corners in the frame t:

$$\mathcal{D}_{\text{lip}}(t) = \frac{\rho(\frac{l^1_{16}+l^1_{17}}{2}, l^t_{16}) + \rho(\frac{l^1_{16}+l^1_{17}}{2}, l^t_{17})}{2\rho(l^1_{16}, l^1_{17})}, \tag{4.5}$$

where l^t_i denotes the 3D location of the i^{th} point in frame t. This estimate is smoothed by the 4253H-twice method [127]. Then, the longest continuous increase in $\mathcal{D}_{\text{lip}}$ is defined as the onset phase. Similarly, the offset phase is detected as the longest continuous decrease in $\mathcal{D}_{\text{lip}}$. The phase between the last frame of the onset and the first frame of the offset defines the apex.

To extract dynamic features from the given facial regions, deformation amplitude ($\mathcal{D}$) of the j^{th} patch at time t is estimated by:

$$\mathcal{D}_j(t) = \frac{\sum_{i=1}^{n_j} \lambda\,(j,i,t)}{\sum_{i=1}^{n_j} \lambda\,(j,i,1)}, \qquad j = \{1,2,\ldots,7\}, \tag{4.6}$$

where n_j shows the number of meshes in patch j. $\lambda\,(j,i,t)$ denotes the area of the i^{th} triangular mesh of patch j at time t. Let p_1, p_2, and p_3 be the corner points of the related mesh, then its surface area is calculated by:

$$\lambda = \sqrt{\gamma(\gamma - \rho\,(p_1,p_2))(\gamma - \rho\,(p_1,p_3))(\gamma - \rho\,(p_2,p_3))}, \tag{4.7}$$

where

$$\gamma = \frac{\rho\,(p_1,p_2) + \rho\,(p_1,p_3) + \rho\,(p_2,p_3)}{2}. \tag{4.8}$$

Deformation amplitudes $\mathcal{D}_j$ are hereafter referred to as amplitude signals. As shown in Eq. 4.6, amplitude signals ($\mathcal{D}_j$) are normalized by the initial patch area (area in the first frame of the onset) for the sake of analysis. In addition to the amplitudes, speed $\mathcal{V}$ and

acceleration $\mathcal{A}$ signals are computed by using the first and the second derivatives of the amplitudes, respectively:

$$\mathcal{V}(t) = \frac{d\mathcal{D}}{dt}, \tag{4.9}$$

$$\mathcal{A}(t) = \frac{d^2\mathcal{D}}{dt^2} = \frac{d\mathcal{V}}{dt}. \tag{4.10}$$

All the calculated amplitude signals are smoothed by the 4253H-twice method [127], and then split into three phases as onset, apex, and offset, which are previously defined using the amplitude signal $\mathcal{D}_{\text{lip}}$ of the lip corners.

A summary of the proposed dynamic features is given in Table 4.1. Note that the defined features are extracted separately for each phase of the expression. As a result, we obtain three feature sets for each of the surface patches. Each phase is further divided into increasing ($^+$) and decreasing ($^-$) segments, for each feature set. This allows a more detailed analysis of the feature dynamics. Most of these features were originally proposed to analyze smile expressions [26], and a similar set has been employed for automatic kinship estimation through smile dynamics [29]. The present study demonstrates that they are also powerfully descriptive for the disgust expression.

In Table 4.1, signals symbolized with superindex ($^+$) and ($^-$) denote the segments of the related signal with continuous increase and continuous decrease, respectively. For example, $\mathcal{D}^+$ pools the increasing segments in $\mathcal{D}$. η defines the length (number of frames) of a given signal, and ω is the frame rate of the video. $\mathcal{D}_L$ and $\mathcal{D}_R$ define the amplitudes for the left and right sides of the face, respectively. ψ denote the Discrete Cosine Transform (DCT) coefficients of $\mathcal{D}$ and computed by:

$$\psi(k) = \frac{1}{\varphi(k)} \sum_{t=1}^{\eta(\mathcal{D})} \mathcal{D}(t) \cos\left(\frac{\pi(2t-1)(k-1)}{2\eta(\mathcal{D})}\right), \tag{4.11}$$

where

$$\varphi(k) = \begin{cases} \sqrt{\eta(\mathcal{D})} & : k = 1 \\ \sqrt{\frac{\eta(\mathcal{D})}{2}} & : 2 \leq k \leq \eta(\mathcal{D}) \end{cases} \tag{4.12}$$

Since a low frequency signal can be reconstructed efficiently by using only a few DCT coefficients, we enable the first 10 DCT coefficients ($\psi(k)$, $k = \{1, 2, \ldots, 10\}$) of the amplitude signals in the feature set. As a result, for each face region, seven 35-dimensional feature vectors are generated by concatenating these features.

In some cases, features cannot be calculated. For example, if we extract features from the amplitude signal of the mouth patch using the onset phase, the decreasing segments can be an empty set ($\eta(\mathcal{D}^-) = 0$). For such exceptions, all the features describing the related segments are set to zero. This is done to have a generic feature vector format which has the same features for different phases of each face region.

Tab. 4.1: Definitions of the extracted features

Feature	Definition
Frequency Components:	$[\,\psi(1)\,,\psi(2)\,,\ldots,\psi(10)\,]$
Duration:	$\left[\,\frac{\eta(\mathcal{D}^+)}{\omega}\,,\frac{\eta(\mathcal{D}^-)}{\omega}\,,\frac{\eta(\mathcal{D})}{\omega}\,\right]$
Duration Ratio:	$\left[\,\frac{\eta(\mathcal{D}^+)}{\eta(\mathcal{D})}\,,\frac{\eta(\mathcal{D}^-)}{\eta(\mathcal{D})}\,\right]$
Maximum Amplitude:	$\max(\mathcal{D})$
Mean Amplitude:	$\left[\,\frac{\sum\mathcal{D}}{\eta(\mathcal{D})}\,,\frac{\sum\mathcal{D}^+}{\eta(\mathcal{D}^+)}\,,\frac{\sum\lvert\mathcal{D}^-\rvert}{\eta(\mathcal{D}^-)}\,\right]$
STD of Amplitude:	$\mathrm{std}(\mathcal{D})$
Total Amplitude:	$[\,\sum\mathcal{D}^+\,,\sum\lvert\mathcal{D}^-\rvert\,]$
Net Amplitude:	$\sum\mathcal{D}^+-\sum\lvert\mathcal{D}^-\rvert$
Amplitude Ratio:	$\left[\,\frac{\sum\mathcal{D}^+}{\sum\mathcal{D}^+\!+\!\sum\lvert\mathcal{D}^-\rvert}\,,\frac{\sum\lvert\mathcal{D}^-\rvert}{\sum\mathcal{D}^+\!+\!\sum\lvert\mathcal{D}^-\rvert}\,\right]$
Maximum Speed:	$[\,\max(\mathcal{V}^+)\,,\,\max(\lvert\mathcal{V}^-\rvert)\,]$
Mean Speed:	$\left[\,\frac{\sum\mathcal{V}^+}{\eta(\mathcal{V}^+)}\,,\frac{\sum\lvert\mathcal{V}^-\rvert}{\eta(\mathcal{V}^-)}\,\right]$
Maximum Acceleration:	$[\,\max(\mathcal{A}^+)\,,\,\max(\lvert\mathcal{A}^-\rvert)\,]$
Mean Acceleration:	$\left[\,\frac{\sum\mathcal{A}^+}{\eta(\mathcal{A}^+)}\,,\frac{\sum\lvert\mathcal{A}^-\rvert}{\eta(\mathcal{A}^-)}\,\right]$
Net Ampl., Duration Ratio:	$\frac{(\sum\mathcal{D}^+-\sum\lvert\mathcal{D}^-\rvert)\omega}{\eta(\mathcal{D})}$
Left/Right Ampl. Difference:	$\frac{\lvert\sum\mathcal{D}_L-\sum\mathcal{D}_R\rvert}{\eta(\mathcal{D})}$

4.3.4 Appearance Features

To describe the appearance of faces, we use four different state-of-the-art descriptors: namely, intensity-based encoded aging features, gradient-based encoded aging features, biologically-inspired aging features, and local binary patterns (LBP). Details of these appearance descriptors are given in this section.

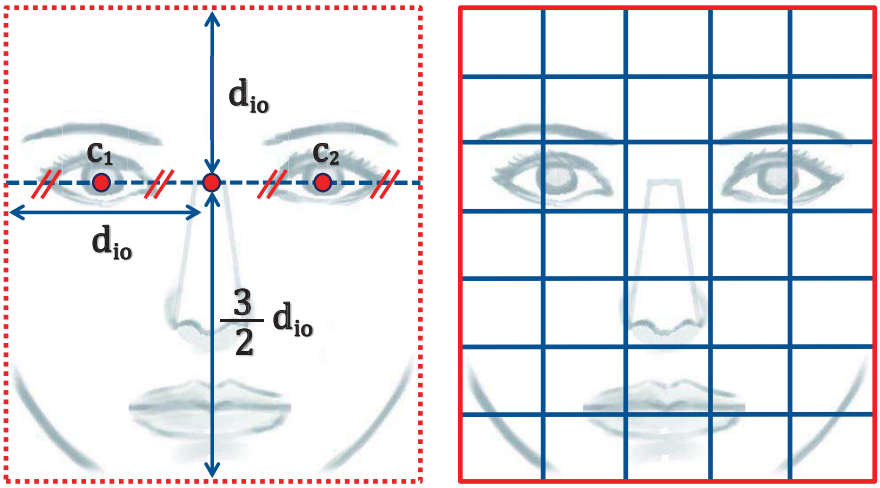
Intensity-based (IEF) and gradient-based encoded aging features (GEF) are proposed by Alnajar *et al.* [7]. These features are based on a learning-based encoding. A dis-

criminative low-level feature is computed for each pixel. Then, the features are encoded by a PCA-tree-based codebook [36]. The face is divided into patches and the codes in each patch are described by a histogram. Finally, the patch histograms are concatenated together to form the aging descriptor. Two versions of the descriptor are used based on low-level features: GEF based on gradient histogram (to capture wrinkle details) and IEF based on intensity sampling (to capture skin texture and fine wrinkle details). For IEF, the neighboring intensities are sampled around each pixel in a ring-based pattern. 25 intensity values are sampled at the circumferences of two rings with $r = 1$ (8 values) and $r = 2$ (16 values) including the central pixel value itself. To extract GEF, the gradient directions are computed in an 8×8 neighborhood of each pixel. The gradient orientations are binned to equally-spaced bins over $0^\circ - 360^\circ$, where the gradient magnitudes are accumulated. As in [7], Gaussian derivatives are chosen for calculating the gradient and the number of bins equals eight.

Biologically-inspired aging features (BIF) are introduced by Guo *et al.* [53] for age estimation. The features are extracted by applying two-layer filters. In the first layer, BIF uses Gabor filter responses for different orientations and scales. In the second layer, it assembles the responses from the first layer in a local area (with the same directions and adjacent scales “band") to a single value using max or standard deviation functions. The authors adapt the descriptor from [102] by introducing the standard deviation operation in creating the second layer and making the number of bands and orientations adaptive to the data. In our experiments, for sake of simplicity, 16 orientations and eight bands are computed to build the descriptor.

The original local binary patterns operator, which is proposed by Ojala *et al.* [82], takes the intensity value of the center pixel as threshold to convert the neighborhood pixels to a binary code. Computed binary codes describe the ordered pattern of the center pixel. This procedure is repeated for each pixel on the image and the histogram of the resultant 256 labels can then be used as a texture descriptor. In [83], Ojala *et al.* show that a large number of the local binary patterns contain at most two bitwise transitions from 0 to 1 or 1 to 0, which is called a uniform pattern. Therefore, during the computation of the histograms, the size of the feature vector can be significantly reduced by assigning different bins for each of the 58 uniform patterns and one bin for the rest. Uniform local binary patterns are used in experiments, and are hereafter referred to as LBP. Eight neighborhood pixels (on a circle with a radius of 1 pixel) are used to extract the LBP features.

Since the onset of a facial expression starts with a neutral face, the first frame of the previously detected onset phase is selected to extract the appearance features. On the selected frame, the roll rotation of the face is estimated and normalized using the eye centers c_1 and c_2. Then, the face is resized and cropped as shown in Fig. 4.2(a). The inter-ocular distance d_{io} is set to 50 pixels to normalize the scale and cropping. As



(a) (b)

Fig. 4.2: (a) Cropping of a face image, and (b) the defined 7×5 blocks to extract appearance features.

a result, each normalized face image has a resolution of 125×100 pixels. After the preprocessing step, appearance features (IEF, GEF, BIF, and LBP) are computed. IEF, GEF, and LBP descriptors are extracted from 7×5 non-overlapping (equally-sized) blocks [see Fig. 4.2(b)]. For all descriptors, the dimensionality of the appearance feature vectors is reduced by Principal Component Analysis (PCA) so as to retain 99.99% of the variance.

4.3.5 Feature Selection and Classification

Estimating the age of a person by using a generic classifier/regressor is an inherently challenging problem, since many factors influence the age for different age groups (mainly shape in early ages and appearance in later ages [89]) and the learning-based predictor should capture all these details from the training data to produce a correct age estimation. One solution for this problem is dividing the prediction of the age into two phases: The first one predicts the age group. Next, a second fine-tuned age prediction model is learned to estimate the exact age.

In the two-level age prediction, the sample is first classified into an age group (first-level prediction). Later, another predictor will place the sample in its exact age (second-level prediction). In [26], the age groups are determined in a uniform way ($8-14$, $15-17$, $18-21$, $22-28$, $29-35$, $36-54$, $55-76$). However, problems may arise when boundary ages between two adjacent groups are not distinctive (i.e. the aging features are similar). In such cases, the first-level prediction is more prone to go wrong which

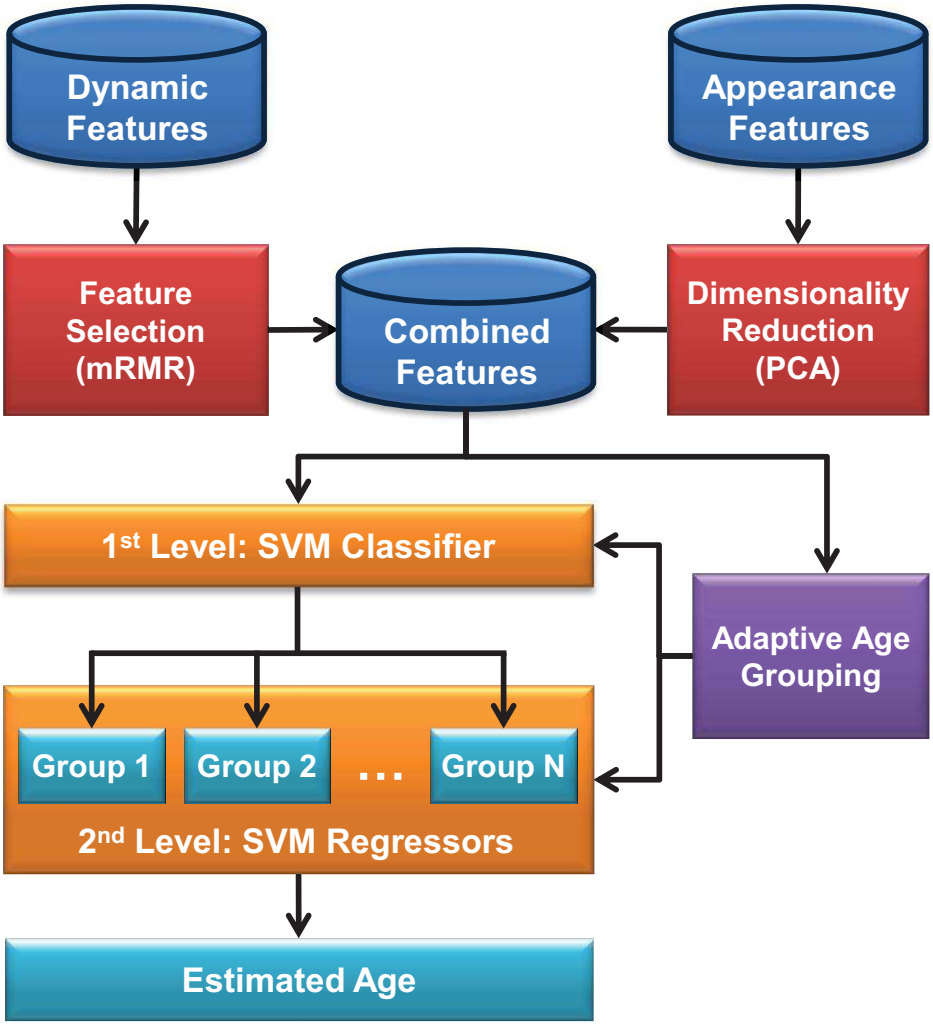


Fig. 4.3: Two-level age estimation architecture using both appearance and dynamic features.

is likely to propagate the error to the second level prediction. To overcome this issue, we propose a method that computes the ages which are dissimilar with their neighbors. The whole age range is divided into groups in such a way that the boundary between each two adjacent groups is discriminant. To this end, the average cosine similarity $\mathcal{S}$ between each age a and its $2q$ neighbors $h = \{a-q, a-q+1, \ldots, a+q\} - \{a\}$ is computed by:

$$\mathcal{S}_a = \frac{1}{2qn_a} \sum_{i=1}^{2q} \sum_{j=1}^{n_a} \sum_{k=1}^{n_{h_i}} \frac{d_a^j . d_{h_i}^k}{n_{h_i} \left\| d_a^j \right\| \left\| d_{h_i}^k \right\|}, \tag{4.13}$$

where d_a^j denotes the feature vector of age a's j^{th} sample. n_a denotes the number of samples for age a. After smoothing $\mathcal{S}$, age a is set to be a group boundary if $\forall h_i$, $\mathcal{S}_{h_i} >$

$\mathcal{S}_a$. Minimum and maximum age values in the whole range are also set as boundaries. Each boundary age is included in the same group with its most similar adjacent neighbor. The number of neighborhood levels $q \geq 2$ is selected automatically on the validation data.

In the given two level architecture (see Fig. 4.3), we use Support Vector Machine classifiers and regressors for age estimation. In the first level, one-vs-all SVM classifiers are used to classify the age of a subject into automatically defined age groups. Then, the age of the subject is fine-tuned using an SVM regressor which is specifically trained for the related age group. For an improved estimation, the regressor of each age group is trained with an age interval of -10 to $+10$ years of group boundaries. Then, the results are limited by the age range (if the estimated age is less/more than the group boundaries, it is set to the minimum/maximum age of the group). The resulting estimation of the age is given as an integer with a 1 year resolution.

As described in Section 4.3.3, we extract three 35-dimensional dynamic feature vectors for each face region. To deal with feature redundancy, we use the Min-Redundancy Max-Relevance (mRMR) algorithm to select the discriminative dynamic features [87]. mRMR is an incremental method minimizing the redundancy while selecting the most relevant information as follows:

$$\max_{f_j \in F - S_{m-1}} \left[I\left(f_j, c\right) - \frac{1}{m-1} \sum_{f_i \in S_{m-1}} I\left(f_j, f_i\right) \right], \tag{4.14}$$

where I shows the mutual information function and c indicates the target class. F and S_{m-1} denote the feature set, and the set of $m-1$ features, respectively. Then, all the selected dynamic features are concatenated with the appearance features (which are extracted from the first frame of the expression onset and reduced by PCA) to train the system (see Fig. 4.3). Minimum classification error on a separate validation set is used to determine the most discriminative dynamic features. Similarly, to optimize the SVM configuration, different kernels (linear, polynomial, and radial basis function) with different parameters (degree of polynomial kernel) are tested on the validation set and the configuration with the minimum validation error is selected. The test partition of the dataset is not used for parameter optimization.

(a) (b) (c) (d)

Fig. 4.4: Sample frames from the UvA-NEMO Smile and the UvA-NEMO Disgust Databases: (a) Showing neutral face, (b) posed enjoyment smile, (c) spontaneous enjoyment smile, and (d) disgust expression.

4.4 Experimental Settings

4.4.1 UvA-NEMO Smile Database

The UvA-NEMO Smile Database[1] has been collected to analyze the change in dynamics of smiles for different ages [27]. Data collection was carried out in the Science Center NEMO (Amsterdam) [101] as part of Science Live, an innovative research programme. NEMO visitors are the volunteers for the data collection. The database and its evaluation protocols are made available to the research community.

This database is composed of videos (in RGB color) recorded with a Panasonic HDC-HS700 3MOS camcorder, placed on a monitor at approximately 1.5 meters away from the recorded subjects. Videos are recorded with a resolution of 1920×1080 pixels at a rate of 50 frames per second under controlled illumination conditions. Additionally, a color chart is present on the background of the videos for illumination and color normalization. Sample frames from the database are shown in Fig. 4.4.

The database has 1240 smile videos (597 spontaneous, 643 posed) from 400 subjects (185 female, 215 male). The ages of subjects vary from 8 to 76 years. 43 subjects do not have spontaneous smiles and 32 subjects have no posed smile samples. Age and gender distributions of the subjects in the database are given in Fig. 4.5(a).

For posed smiles, each subject was asked to pose a smile as realistically as possible, sometimes after being shown the proper way in a sample video. Short, funny video segments are used to elicit spontaneous smiles. Approximately five minutes of recordings are made per subject, and genuine smiles are segmented.

[1] [Online] Available: http://www.uva-nemo.org

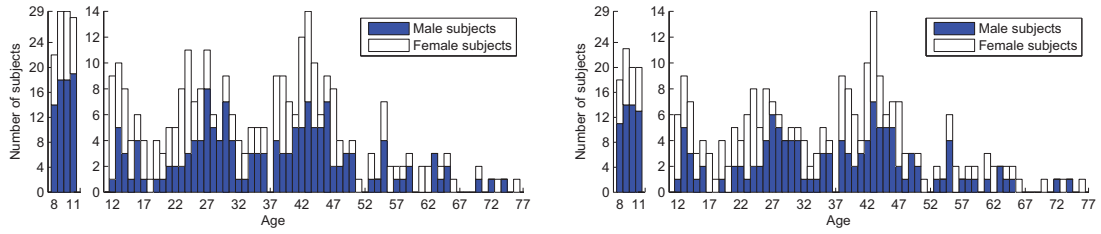


Fig. 4.5: Age and gender distributions of the subjects in (a) the UvA-NEMO Smile and (b) the UvA-NEMO Disgust Databases.

For each subject, a balanced number of spontaneous and posed smiles are selected and annotated by seeking consensus of two trained annotators. Each segment starts and ends with neutral or near-neutral expressions.

4.4.2 UvA-NEMO Disgust Database

To show the applicability of the proposed approach on other facial expressions, we introduce the UvA-NEMO Disgust Database[1] in this paper. This database is composed of posed (deliberate) disgust expressions, and recorded during the collection of the UvA-NEMO Smile Database using the same recording/illumination setup. Sample frames from the database are shown in Fig. 4.4.

Each subject was asked to pose a disgust expression as realistically as possible, sometimes after being shown a sample video. For each subject, one or two posed disgust expressions were selected and annotated by seeking consensus of two trained annotators. Each segment starts and ends with neutral or near-neutral expressions. The resulting database has 518 deliberate disgust videos from 324 subjects (152 female, 172 male). 313 of 324 subjects are also included in the 400 subjects of the UvA-NEMO Smile Database. The ages of subjects vary from 8 to 76 years. Age and gender distributions of the subjects in the database are given in Fig. 4.5(b). The database and its evaluation protocols are made available to the research community.

4.4.3 Settings

To evaluate our system and assess the reliability of facial expression dynamics and facial appearance information for age estimation problem, we first use the UvA-NEMO Smile Database of 400 subjects. We then show the effectiveness of the proposed approach for disgust expression using the UvA-NEMO Disgust Database. In our experiments,

the two-level classification/regression system is used as described in Section 4.3.5. The optimum number of selected dynamic features, adaptive age ranges, and kernels of the SVM classifiers/regressors are determined on a separate validation partition. To this end, a two level 10-fold cross-validation scheme is used. Each time a test fold is separated, a 9-fold cross-validation is used to train the system, and parameters are optimized without using the test partition. Candidate settings for these parameters are set as follows: Neighborhood level $q = \{2, 3, 4\}$ for adaptive age grouping; candidate kernels for the SVMs are linear, polynomial, and radial basis function. In polynomial and radial basis function kernels, the gamma parameter is set as $1/d$, where d is the number of features.

There is no subject overlap between folds in either database. We initialize the tracking by automatically annotated facial landmarks. For automatic facial landmark detection, we use the system proposed by Dibeklioğlu *et al.* [28]. The mean localization error for the related landmarks [corners and center of eyebrows, eye corners, center of upper eyelids, nose tip, lip corners; see Fig. 4.1(a)] is 3.84% of the inter-ocular distance to the actual location of the landmarks. Correlation coefficients between the extracted amplitude signals with manual and automatic initializations ranged between 0.93 and 1.

4.5 Experiments

In this section, we present the results of our experiments on exact age estimation. First, we evaluate the accuracy of the proposed system when only facial dynamics are used, either individually for each facial region, or taken together, on smile expressions. We compare these results with the combined use of appearance and dynamics. We then test the influence of gender and expression spontaneity on the accuracy of the system using the combined features. Finally, we report age estimation results using disgust expression dynamics.

4.5.1 Dynamics

Since the proposed dynamic features are extracted from the deformations of seven different surface patches, we analyze the individual discrimination power of these deformations and their combination for age estimation. Furthermore, to assess the reliability of the feature selection step, performance of using automatically selected (most) informative dynamic features and the use of all features without any selection are compared. The resulting *mean absolute error* (MAE) is given in Table 4.2.

As shown in Table 4.2, the feature selection increases the accuracy by approximately

Tab. 4.2: Effect of using different facial regions with and without feature selection on the UvA-NEMO Smile Database

Regions	MAE (years)	
	without Feat. Selection	with Feat. Selection
1: Eyebrow	15.34 ($\pm$10.59)	13.32 ($\pm$9.63)
2: Eyelid	15.87 ($\pm$11.38)	13.50 ($\pm$10.21)
3: Eye-sides	14.74 ($\pm$10.15)	12.93 ($\pm$9.52)
4: Cheek	13.88 ($\pm$10.44)	12.14 ($\pm$9.35)
5: Mouth-sides	14.98 ($\pm$12.36)	13.27 ($\pm$11.42)
6: Mouth	15.74 ($\pm$13.73)	14.15 ($\pm$12.11)
7: Chin	28.70 ($\pm$29.70)	24.42 ($\pm$26.29)
1–7: All	12.04 ($\pm$9.81)	10.81 ($\pm$8.85)

13% (relative) on average, while reducing the dimensionality of the feature space. Since the efficacy of the feature selection step is confirmed by these results, it is used in the remainder of our experiments. By analyzing the regional results with feature selection, it can be derived that the dynamics of cheek's surface deformations are the most reliable features, with an MAE of 12.14 (± 9.35) years. Deformation dynamics on the sides of the eyes follow closely with an MAE of 12.93 (± 9.52) years. The chin region provides an MAE of only 24.42 (± 26.29) because of its stationary surface characteristic. By combining the dynamic features of different facial regions, the MAE of the age estimation is decreased to 10.81 (± 8.85) years.

4.5.2 Dynamics versus Appearance

The aim of this work is to improve the accuracy of age estimation by combining facial appearance with expression dynamics. However, it is also important to show the discriminative power of facial expression dynamics and appearance, individually. For this purpose, the individual and combined uses of these features are evaluated.

As shown in Table 4.3, combining dynamics with appearance features significantly improves the age estimation accuracy in comparison to the individual use of dynamic and appearance features ($p < 0.0015$ for GEF, BIF and LBP; $p < 0.01$ for IEF). It is clear that the use of only facial dynamics is not sufficient for accurate age estimation. The MAE of using dynamic features is 10.81 (± 8.85) years, where the MAEs for different facial appearance descriptors range from 4.80 (± 4.77) to 5.78 (± 6.15) years. However, by combining dynamic and appearance features, the proposed system is able to achieve

Tab. 4.3: Mean Absolute Errors for dynamics, appearance, and combined features on the UvA-NEMO Smile Database

Features	MAE (years)	
	without *Dynamics*	with *Dynamics*
Appearance: None	N/A	10.81 (±8.85)
Appearance: IEF	4.80 (±4.77)	4.33 (±4.03)
Appearance: GEF	5.48 (±5.57)	4.82 (±4.89)
Appearance: BIF	5.78 (±6.15)	5.03 (±5.10)
Appearance: LBP	5.46 (±5.58)	4.77 (±4.66)

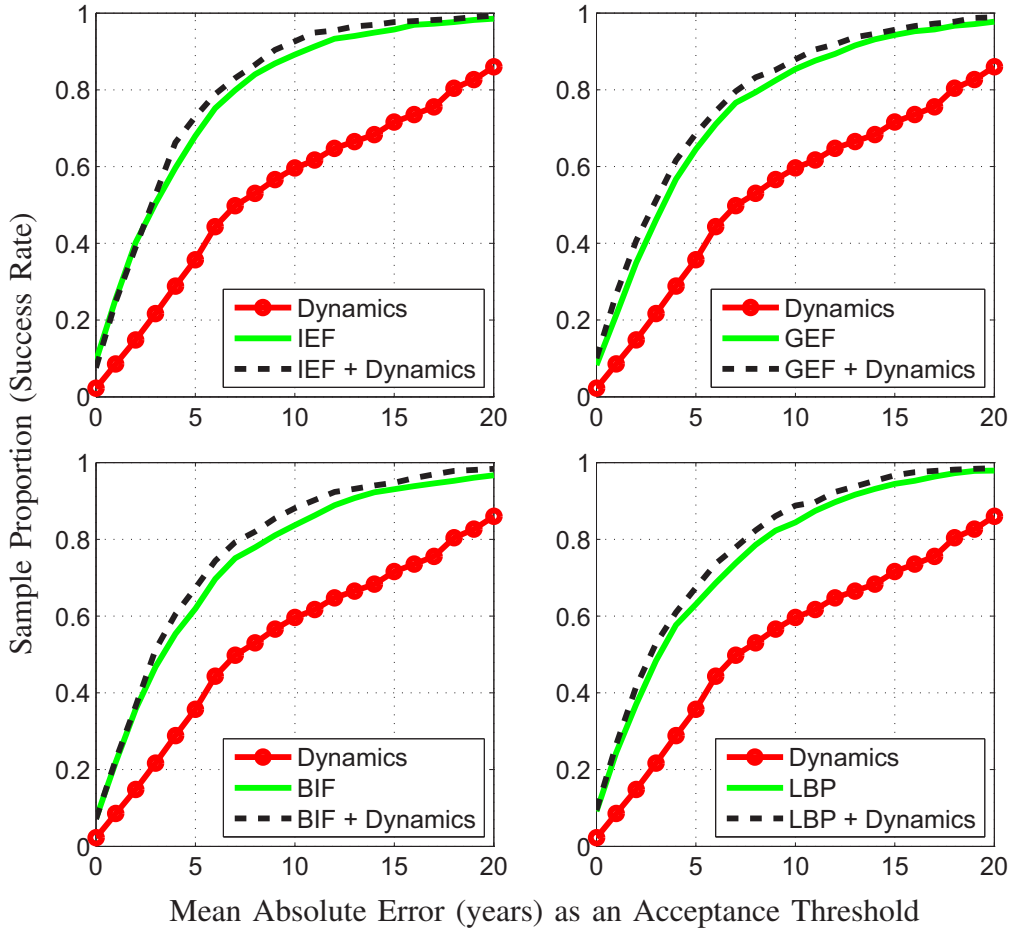


Fig. 4.6: Cumulative distribution of the mean absolute error for different features on the UvA-NEMO Smile Database.

Tab. 4.4: Performance of adaptive age grouping, grouping into bins of 10 years and the regression without grouping (none) on the UvA-NEMO Smile Database

Features	MAE (years)		
	None	10-years	Adaptive
IEF + Dynamics	5.00 ($\pm$4.25)	4.40 ($\pm$4.11)	4.33 ($\pm$4.03)
GEF + Dynamics	5.63 ($\pm$4.86)	4.97 ($\pm$5.07)	4.82 ($\pm$4.89)
BIF + Dynamics	5.12 ($\pm$4.91)	5.94 ($\pm$5.24)	5.03 ($\pm$5.10)
LBP + Dynamics	5.29 ($\pm$4.36)	4.83 ($\pm$4.60)	4.77 ($\pm$4.66)

the best results. The combination of dynamics and IEF provides the highest accuracy with an MAE of 4.33 (± 4.03). The cumulative distribution of the MAEs for individual and combined features are shown in Fig. 4.6.

The UvA-NEMO Database has uniform D65 illumination that does not specifically highlight wrinkles. For images with direct illumination, we observe that wrinkles become much more pronounced, and gradient based descriptors perform better than intensity based features.

4.5.3 Assessment of Adaptive Age Grouping

We now test the use of a two-level classification/regression strategy. We evaluate the performance of using three different classification/regression approaches, namely direct regression (no grouping), classifying age groups into bins of 10-years and into adaptive bins (based on training) before group-specific regression. The resulting MAEs of each method for different feature combinations and their cumulative distributions are shown in Table 4.4 and in Fig. 4.7, respectively.

The results show that the adaptive grouping outperforms other approaches for all features. 10-years grouping follows it, and provides a more accurate estimation in comparison to that of direct regression in most cases.

4.5.4 Effect of Gender

To assess the effect of gender on the accuracy of age estimation using the combined features, a gender-specific age estimation approach is implemented. In the gender-specific method, different classifiers/regressors are trained and tested for both males and females, separately. For this method, we assume that the gender labels of all samples are given correctly. The MAEs for both gender-specific and the general approach are given in Table 4.5.

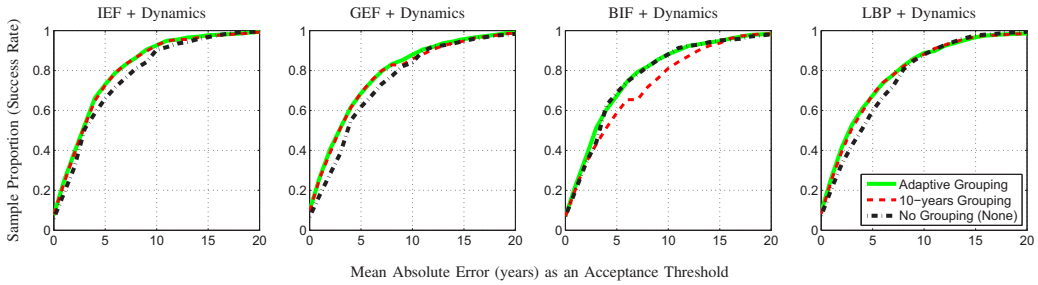


Fig. 4.7: Cumulative distribution of the mean absolute error for different grouping strategies on the UvA-NEMO Smile Database.

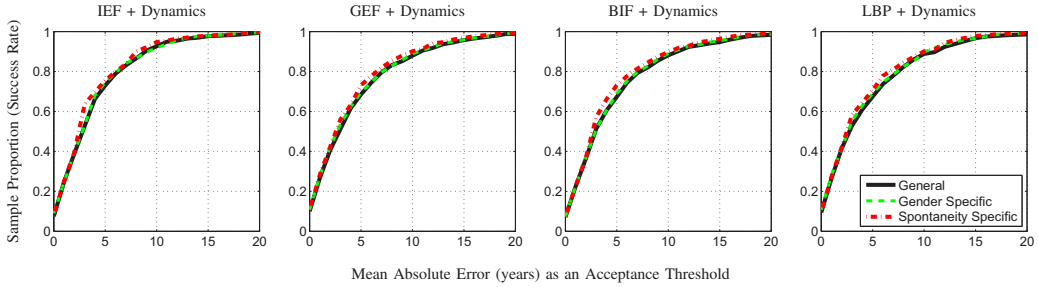


Fig. 4.8: Cumulative distribution of the mean absolute error for general, gender-specific, and spontaneity-specific methods on the UvA-NEMO Smile Database.

Tab. 4.5: Comparison of the gender-specific method with the general method for age estimation on the UvA-NEMO Smile Database

Features	MAE (years)	
	Gender-specific	General
IEF + Dynamics	$4.25\ (\pm 3.95)$	$4.33\ (\pm 4.03)$
GEF + Dynamics	$4.67\ (\pm 4.71)$	$4.82\ (\pm 4.89)$
BIF + Dynamics	$4.91\ (\pm 4.88)$	$5.03\ (\pm 5.10)$
LBP + Dynamics	$4.58\ (\pm 4.47)$	$4.77\ (\pm 4.66)$

Our results show that the gender-specific training decreases the overall MAE in comparison the MAE of general-training. The MAE of the gender-specific approach for different features range from $4.91\ (\pm 4.88)$ to $4.25\ (\pm 3.95)$ years. Although the improvement is not statistically significant, the gender-specific training provides a 3% MAE enhancement (relative) on average.

In particular, the improvement for males is more than that of females. The cumulative distribution of the MAE for general and gender-specific methods are shown in Fig. 4.8.

4.5.5 Effect of Expression Spontaneity

To assess the effect of expression spontaneity on the accuracy of using combined features, a spontaneity-specific age estimation system is implemented. For this purpose, separate classifiers/regressors are trained for spontaneous and posed smiles. Spontaneity of smiles is classified using the system of [27]. This system uses similar expression dynamics to distinguish between spontaneous and posed smiles. Correct classification of the system is 87.02% on the UvA-NEMO Smile Database.

Tab. 4.6: Comparison of the spontaneity-specific method with the general method for age estimation on the UvA-NEMO Smile Database

Features	MAE (years)	
	Spontaneity-specific	General
IEF + Dynamics	4.00 (± 3.74)	4.33 (± 4.03)
GEF + Dynamics	4.40 (± 4.44)	4.82 (± 4.89)
BIF + Dynamics	4.59 (± 4.59)	5.03 (± 5.10)
LBP + Dynamics	4.38 (± 4.23)	4.77 (± 4.66)

As shown in Table 4.6, the MAE of the spontaneity-specific approach ranges from 4.00 (± 3.74) to 4.59 (± 4.59), therefore improving the accuracy by 8% (on average) with respect to the general approach. This means a statistically significant ($p < 0.04$) improvement. Spontaneity-specific training decreases the MAE for both posed and spontaneous smiles. Since the automatically detected neutral faces are used to extract the appearance features for both approaches, accuracy improvements by performing spontaneity-specific training indicates the differences between spontaneous and posed smiles in terms of expression dynamics. The cumulative distribution of the MAE for general and spontaneity-specific methods are shown in Fig. 4.8.

4.5.6 Effect of Temporal Phases

An expression video from onset to offset contains a lot of frames. The system we have proposed gives a decision when the expression is completed, i.e. at the end of the offset phase. However, it may be necessary to give a decision while the expression is in progression. To understand how partial information would fare, we implement a version of the proposed method. Spontaneity specific approach (with automatic spontaneity detection) is used with adaptive age grouping in this experiment.

Since the order of the temporal phases during a facial expression is fixed, the online system starts classification in the onset mode, where appearance features are combined

with onset dynamics. When the apex is reached, it uses both the onset and the apex in addition to appearance. In the final stage, dynamics of all three phases are used together with appearance features. For these three modes, separate classifiers are trained.

Tab. 4.7: Effect of using dynamics of different temporal phases for age estimation on the UvA-NEMO Smile Database

Features	MAE (years)			
	w/o Dynamics	+Onset	+Onset to Apex	+Onset to Offset
IEF	$4.80\ (\pm 4.77)$	$4.36\ (\pm 4.24)$	$4.23\ (\pm 4.09)$	$4.00\ (\pm 3.74)$
GEF	$5.48\ (\pm 5.57)$	$4.92\ (\pm 5.02)$	$4.75\ (\pm 4.71)$	$4.40\ (\pm 4.44)$
BIF	$5.78\ (\pm 6.15)$	$5.08\ (\pm 5.36)$	$4.87\ (\pm 4.95)$	$4.59\ (\pm 4.59)$
LBP	$5.46\ (\pm 5.58)$	$4.83\ (\pm 4.98)$	$4.66\ (\pm 4.67)$	$4.38\ (\pm 4.23)$

The performance of the implemented system for the UvA-NEMO Smile Database is given in Table 4.7. The results show that while all phases contribute to the accuracy, the highest improvement rates are provided by combining onset dynamics with appearance features. Including onset dynamics in the feature set decreases the MAE by 11.76% (relative) on average. Including apex and offset phases in the analysis, increases the accuracy further.

4.5.7 Comparison to Other Methods

To the best of our knowledge, this is the first study using facial expression dynamics (such as speed, acceleration, amplitude, etc.) for age estimation. Except for the recent work by [55], none of the previous studies in the literature focus on using temporal information for age estimation.

In [55], Hadid proposes to use spatio-temporal information to classify the ages of the subjects into five groups (child: 0 to 9 years old; youth: 10 to 19; adult: 20 to 39; middle-age: 40 to 59 and elderly: above 60). They use volume LBP (VLBP) features with a tree of four SVM classifiers. VLBP features are extracted from different overlapping face blocks. Then the AdaBoost learning algorithm is used to determine the optimal size and locations of the local rectangular prisms, and select the most discriminative VLBP features for classification, automatically. To evaluate the system, 2000 videos of about 300 frames are randomly segmented from a set of video sequences mainly showing talking faces (collected from the Internet). Additionally, an appearance-based (static) system is implemented for comparison. This baseline method classifies each frame

Tab. 4.8: Mean Absolute Error on the UvA-NEMO Smile Database for different methods.

	Method	MAE (years) for Different Age Ranges								
		0–9	10–19	20–29	30–39	40–49	50-59	60–69	70–79	All
Dynamics	Deformation	4.85	8.72	12.22	13.06	13.53	11.55	14.13	17.82	10.81 (±8.85)
	Displacement [26]	5.42	9.67	11.98	14.53	12.77	15.42	20.57	20.35	11.54 (±11.49)
Appearance	IEF, Fusion	2.73	3.28	4.68	5.36	5.38	9.09	12.97	14.65	4.86 (±4.54)
	GEF, Fusion	2.89	3.17	5.00	6.07	5.38	10.97	16.90	17.53	5.24 (±5.38)
	BIF, Fusion	3.92	3.73	4.90	5.61	5.86	12.30	17.33	18.24	5.63 (±5.79)
	LBP, Fusion	3.42	4.02	5.13	6.63	5.60	9.06	10.43	13.41	5.37 (±5.28)
Combined, Spon. Spec.	IEF + Dynamics	2.25	2.50	3.74	4.59	4.34	8.20	11.07	13.35	4.00 (±3.74)
	GEF + Dynamics	1.40	2.29	3.99	5.17	5.37	10.17	15.00	16.06	4.40 (±4.44)
	BIF + Dynamics	2.68	3.21	4.23	5.09	4.62	9.44	13.17	14.12	4.59 (±4.59)
	LBP + Dynamics	1.53	2.68	3.95	5.31	5.31	9.33	11.83	13.94	4.38 (±4.23)
Spatio-temporal	VLBP [55]	10.69	12.95	15.99	18.54	18.43	16.58	23.80	26.59	15.70 (±12.40)
	LBP-TOP	9.71	11.01	14.19	15.88	16.75	15.29	19.70	23.71	13.83 (±10.97)
	Number of Samples	158	333	215	171	250	66	30	17	1240

in a video, individually, using LBP features with SVM classifiers. Majority voting is used to fuse the classification results of each frame. Hadid reports that the static image (appearance) based approach provides 77.4% correct classification, where the performance of the spatio-temporal approach reaches only 69.2%.

VLBP is a straightforward extension of the original LBP operator to describe dynamic textures (image sequences) [142]. VLBP enables the use of temporal space (T), models the face sequence as a volume, and the neighborhood of each pixel is defined in three dimensional space. In contrast, LBP uses only X and Y dimensions of a single image. Then, the histograms of VLBP are used as features. In [142], Zhao *et al.* have proposed to extract LBP histograms from Three Orthogonal Planes (LBP-TOP) XY, XT, and YT, individually, and concatenate them as a single feature vector.

To compare our system with related approaches using smiles, we implement three baseline methods: (1) VLBP-based spatio-temporal approach, (2) spatio-temporal approach using LBP-TOP features, and (3) appearance-based approach which classifies the first and the last frame of a smile onset (a neutral and an expressive face, respectively) using appearance features, individually, and fuses the estimations by mean operator. All methods use the same classification/regression architecture as our method. Tests are performed on the UvA-NEMO Smile Database. For a fair comparison, all of the compared methods use automatically annotated facial landmarks to initialize the tracking (for face alignment and feature extraction), and 7×5 non-overlapping blocks on the face to compute the histograms. To generate histograms, uniform patterns are used

for LBP-TOP and LBP. The neighborhood size is set to eight for LBP and LBP-TOP, and two for VLBP. Time interval for the volumetric approaches is set to three frames. Zhao *et al.* [142] show that these neighborhood and time interval parameters perform well for facial expression classification. To provide comparable smile durations for spatio-temporal descriptors, each smile phase (onset, apex, and offset) is temporally interpolated to 25 frames using bicubic interpolation. The dimensionality of IEF, GEF, BIF, LBP, VLBP, and LBP-TOP features is reduced by Principal Component Analysis (PCA) so as to retain 99.99% of the variance.

As shown in Table 4.8, the combination of dynamic and appearance features provides the most accurate results. The spontaneity-specific method that combines dynamics and IEF achieves the minimum MAE of 4.00 (± 3.74) years. Note that combining appearance with expression dynamics provides more accurate age estimation than using neutral and expressive frames in a video and averaging the results.

Spatio-temporal methods can only reach a mean accuracy of 15.70 (± 12.40) and 13.83 (± 10.97) years with VLBP and LBP-TOP features, respectively. By the sole use of proposed dynamic features, the system is significantly more accurate than when it uses the spatio-temporal features ($p < 0.001$). Finally, we also compare the deformation-based dynamic features that proposed in this paper (first row of Table 4.8) with the displacement dynamics (of eyelids, cheeks, and lip corners) introduced in our previous study [26] (second row of Table 4.8), and show that surface deformation dynamics perform better.

4.5.8 Computational Load

In this section, we report average time requirements of different feature extraction modules. All the modules except the Bézier volume tracker, are composed of non-optimized MATLAB code or C++/compiled MATLAB code. Speed tests are conducted on an Intel i7-3687U, 2.1GHz (dual core) processor with 16GBs of RAM.

Average time requirements for each module used for different features are given in Table 4.9. For a smile of 3.2 seconds (160 frames), facial expression dynamics can be extracted in 8.0436 seconds (based on landmarking of 17 points, tracking, normalization of landmarks, and dynamics extraction). Similarly, dynamics of a 0.6 seconds (30 frames) onset phase can be extracted in 2.4796 seconds. Extraction of IEF, GEF, BIF, and LBP features requires 1.7785, 4.0825, 38.7947, 0.4599 seconds per frame, respectively (based on landmarking of four points, cropping/alignment of the face image, and feature extraction).

Newer generation of trackers use local binary features for accurate and very fast face alignment, achieving speeds around 3000 fps on desktop computers [91]. Consequently,

Tab. 4.9: Average time requirements of different feature extraction modules

Module	Duration (seconds per frame)
Initial landmarking (four points: eye corners)	0.4323
Initial landmarking (17 points)	1.1204
Tracking	0.0411
Normalization of the landmarks	0.0017
Cropping/alignment of the face image	0.0234
IEF feature extraction	1.3228
GEF feature extraction	3.6268
BIF feature extraction	38.3390
LBP feature extraction	0.0042
Dynamics extraction (per signal)	0.0752

commercial systems will be able to perform face tracking and expression analysis in real time. Our results show that age estimation will also benefit from the availability of accurate and computationally cheap dynamic information.

4.5.9 Application to Disgust Expression

To evaluate the effectiveness of the proposed approach on a different facial expression, we conduct experiments on disgust expression. To this end, the UvA-NEMO Disgust Database is used. Adaptive age grouping is enabled in these experiments. Here, appearance features are extracted from the first, neutral frames of the onset of the disgust videos.

Tab. 4.10: Mean Absolute Errors for dynamics, appearance, and combined features on the UvA-NEMO Disgust Database

Features	MAE (years)	
	without *Dynamics*	with *Dynamics*
Appearance: None	N/A	$10.32(\pm 7.97)$
Appearance: IEF	$5.04\ (\pm 4.90)$	$4.21\ (\pm 4.07)$
Appearance: GEF	$5.50\ (\pm 5.67)$	$4.56\ (\pm 4.74)$
Appearance: BIF	$7.24\ (\pm 8.46)$	$6.09\ (\pm 7.08)$
Appearance: LBP	$5.29\ (\pm 5.05)$	$4.38\ (\pm 4.18)$

As shown in Table 4.10, similar to the results on smiles, combining dynamics of disgust expression with appearance features significantly improves the age estimation accuracy in comparison to the individual use of dynamic and appearance features ($p < 0.05$). MAE improvement over the accuracy of appearance features ranges from 0.83 to 1.15 years. These improvement rates are higher than those given in Table 4.3 for smile videos. One reason is that all the disgust expressions are deliberate, causing the system to act like a spontaneity-specific system and thus providing better modeling.

4.5.10 Classification of Age Ranges

Based on different application requirements, many studies report automatic age estimation results as classification of age groups [40, 90, 123]. Since facial dynamics are much more informative for large age differences, we conduct a set of experiments to show the usefulness of dynamic features in classifying age groups using smile and disgust expressions. Two different set of age groups are evaluated in our experiments: (a) 7 age groups of 10 years (8–17, 18–27, ..., 68–77) as in [26], and (b) 5 age groups (8–12, 13–19, 20–36, 37–65, 66+) as in [40]. Our system is trained for these two different sets of age ranges. Since the UvA-NEMO Smile Database includes both spontaneous and posed smiles, spontaneity specific system is used for smiles by employing the automatic spontaneity detection (for smiles) as proposed in [27]. The general classification approach is used for disgust expression, since the UvA-NEMO Disgust Database has only posed disgust expressions.

As shown in Table 4.11, combining dynamics with appearance features significantly ($p < 0.01$) improves the classification accuracy of age groups. When smile videos are used, mean accuracy improvement for 7 and 5 age groups are 9.58%, and 13.03% (absolute), respectively. When disgust videos are used, mean accuracy improvement for 7 and 5 age groups are 11.78%, and 11.83% (absolute), respectively. In comparison to the improvement on exact age estimation, these results display a more visible enhancement. This is based on higher reliability of expression dynamics for classifying subjects with large age differences. When the estimation of exact age is considered, the use of the expression dynamics improves the first level classification accuracy in a visible way, but in the second level (regression), while the exact age within the classified group is being determined, dynamics fall short of fine-level estimation, and results mainly rely on appearance features. As a result, expression dynamics are much more reliable and discriminative for classifying age groups.

Tab. 4.11: Classification accuracy of age ranges for appearance, and combined features on the UvA-NEMO Smile and the UvA-NEMO Disgust Databases

Expr.	Feature	Classification Accuracy (%)			
		7 Groups		5 Groups	
Smile	IEF	67.50	75.73	76.13	88.55
	GEF	65.48	74.84	74.11	87.58
	BIF	60.56	72.34	71.94	84.76
	LBP	65.40	74.35	70.24	83.63
Disgust	IEF	64.29	76.06	79.34	90.93
	GEF	64.09	75.68	75.68	86.68
	BIF	56.95	69.50	64.67	78.76
	LBP	60.62	71.81	78.96	89.58
		without *Dynamics*	with *Dynamics*	without *Dynamics*	with *Dynamics*

4.6 Discussion

In our experiments, we show that deformation dynamics of cheek's during smiles, perform best for individual regions. Additionally, fusion of all regions (with a feature selection step) improves the accuracy of the cheek dynamics by 10.96%. For dynamic features, using feature selection increases the accuracy approximately by 13% on average, as well as reducing feature dimensionality. This finding indicates that there is a significant amount of noise or confusing information in surface deformation dynamics.

Our results show that the individual use of the facial expression dynamics is not sufficient to obtain an accurate age estimation system. However, accuracy of using solely appearance features of a neutral face (automatically detected as the first frame of the onset phase) is significantly outperformed ($p < 0.0015$ for GEF, BIF and LBP; $p < 0.01$ for IEF) by enabling the surface deformation dynamics of smile expression. Moreover, the use of combined features outperforms all the baseline methods tested in this study. These results confirm the importance of facial expression dynamics. We also show that the deformation-based dynamics outperform the displacement dynamics (of eyelids, cheeks, and lip corners) introduced in our previous study [26].

To obtain the most informative dynamic features for age estimation, we use the frequently selected descriptors (in feature selection procedure). Significant ($p < 0.001$,

$\eta^2 > 0.15$) differences of these features between different ages are investigated using multivariate analysis of variance (MANOVA). Majority of the frequently selected features are extracted from onset and offset of smiles. Additionally, the differences of these features among different ages display lower significance level (p) in comparison to the apex features. Such findings indicate that the deformation dynamics of smile onsets and offsets are more discriminative than those of apex phase for age estimation. During the onset phase of smiles, the mean speed and the mean amplitude of deformation decrease ($\mathcal{D}^-$) on eyelids significantly change among different ages. During the apex phase, the maximum and mean amplitude of deformation on the mouth region are significantly different for different ages. When the offset features are analyzed, it is shown that the maximum and net amplitude of deformation on the mouth region significantly change. Additionally, the second frequency component ($\psi(2)$) of deformation amplitude on the mouth region significantly differs among ages. Note that $\psi(2)$ denotes the lowest frequency coefficient of the deformation amplitude, since $\psi(1)$ is always the DC-component ($\frac{\sum \mathcal{D}}{\sqrt{\eta(\mathcal{D})}}$). So it can be inferred that during smile offsets, the rough shape of the mouth deformation amplitude is an informative feature for age estimation.

Then, we have analyzed the significant differences of these features between spontaneous and posed smiles using the t-test. Our results show that the mean speed and the amplitude of deformation decrease ($\mathcal{D}^-$) on eyelids are significantly higher ($p < 0.005$) for posed smiles during the smile onsets. During the offset phase, on the mouth region, the second frequency component of deformation is lower ($p = 0.002$) for spontaneous smiles, whereas the net amplitude of deformation is ($p < 0.001$) significantly higher. Similarly, the t-test analysis is repeated for male and female differences. The results indicate that during smile apexes, the maximum and mean amplitude of the deformation of the mouth region is significantly lower for males ($p < 0.001$). During the offset phase, on the mouth region, the second frequency component and the maximum amplitude of deformation is lower ($p < 0.001$) for males, whereas the net amplitude of the deformation is ($p < 0.001$) significantly higher. These findings can explain the higher accuracy of the spontaneity- and gender-specific systems. The reader is referred to [25] for further analysis of smile dynamics.

Experimental results show that spatio-temporal approaches (using smiles) based on VLBP and LBP-TOP are not efficient for age estimation. Even the individual use of our dynamic features outperforms these methods significantly. Spatio-temporal features describe the change of facial appearance in time, but our proposed approach models the appearance on a single neutral image (which is automatically selected as the first frame of the onset phase) and adds the surface deformation dynamics of the facial expression (such as amplitude, speed, acceleration, etc) on it. As a result, the proposed system (using spontaneity-specific approach) is significantly ($p < 0.001$) more accurate than all the competitor methods used in our experiments. When we evaluate the performance

of combining proposed features with appearance for disgust expression, similar to our results on smiles, age estimation accuracy is significantly improved.

Our additional experiments on the classification of age ranges demonstrate that the facial expression dynamics are much more reliable for group classification tasks. This is due to higher discrimination power of expression dynamics for classifying subjects with large age differences. However, dynamics are not discriminative enough for discriminating similar ages. This finding can be explained by the large variation of expression dynamics within a narrow age range.

4.7 Conclusions

Our study shows that dynamic facial features obtained by analyzing a frequently occurring facial expression improves appearance-based age estimation. While appearance is more informative than facial dynamics, it is affected by many external factors, like make-up, scars, and wrinkles resulting from exposure to harsh weather conditions. Such factors do not concern facial dynamics. Consequently, dynamics are sufficiently uncorrelated with appearance to allow fusion approaches for age estimation.

In our previous work, we have assessed a range of dynamical features in an exploratory fashion, and have shown that if landmark movements are employed, eyelid dynamics are the most revealing in terms of age estimation, followed by lip corners and cheeks [26]. The present work improves these results by using surface area features (instead of landmark movements) for characterizing 3D facial dynamics. We also introduce in this paper a two-level classifier, where the age range for each classifier is adaptively selected in the first level. We test four different features for appearance to show that the improvement by dynamic features is consistent across representations, and we also introduce an appearance fusion baseline. We study gender effects systematically, to conclude that the improvement due to gender-specific models is not significant. We show that spontaneous and posed smiles have different and distinct dynamics, and spontaneity-specific age estimation significantly outperforms the general approach. Finally, we demonstrate that the method we propose is usable with other expressions, and report results on the new UvA-NEMO Disgust Database we introduce in this paper. Subsequently, this is the most extensive dynamic age evaluation study to this date in the literature.

5

Deep Aging Features

5.1 Introduction

Automatic age estimation is an important research problem in the field of computer vision. Its applications range from surveillance and security control to human-computer interaction and online marketing. In general, the human age is derived from facial aging cues where for adults it is primarily perceived via skin changes [38]. During aging, wrinkles become more apparent in some regions of the face (e.g. around the eyes). The facial skin turns thinner, darker, and rougher. The age estimation schemes consist primarily of two component; aging feature extraction and features classification/regression [38]. In the last decade, there have been many research works to design [53] or use [7, 21, 133] appearance features that capture the aging cues on the face. Some standard features are taken from other tasks such as Local Binary Patterns (LBP) [83] or Gabor filters, while others are devised or modified to represent the age(e.g. Biologically-Inspired Features (BIF) [53]). These aging cues are inferred from different regions of the face. For example, wrinkles develop mainly around the nose, the eyes and the mouth corners, while skin texture changes mainly appear on the cheeks and the forehead. Therefore, a proper age estimation system should employ suitable aging features for the related face regions. As the age from faces is inferred mainly from the wrinkles and the skin texture, some of the aging features are aimed to capture the details of the wrinkles on the face (wrinkle features) [41, 53, 112, 114] while others measure aging changes of skin texture (skin texture features) [7, 133]. These features are either applied exhaustively to cover the whole face [7, 53, 133] or, in a better way, applied to the corresponding face regions [112]. Some recent age estimation systems combine both types of features (each extracted from different face regions) [21] to obtain improved age descriptors.

However, the design/choice of the feature types and the corresponding face regions are handcrafted. For example, Choi et al. [21] extract the wrinkle aging cues using a bank of Gabor filters where the number, the orientations and the scales of the filters are estimated. The locations of the wrinkle analysis regions in [21] are inferred from studies on cosmetic surgery and biology. However, these regions are prone to mis-localization (due to errors in facial landmark detection) and the hence the filters may not match the intended regions. This suggests designing feature extraction scheme that is tolerant to misalignment. To address this, we propose in this paper to learn the most suitable aging features for each part of the face. The learnt features are adapted to slight misalignment of face regions. Hence, producing more discriminate and, yet, more robust aging descriptors.

More specifically, we decompose the face into different parts and assign a convolutional network for each one. The aim is to find the set of filters that fit each corresponding part. The pooling layer in the network serves to alleviate the effects of small registration error. Each part-based convolutional network is trained independently. Although different parts of the face show different aging cues, we assume that the adjacent ones are similar. To this end, a combination layer is topped over the part-based convolutional networks. The output of the combination layer summarizes the features from all the learnt part-specific face features into a single aging descriptor. The resulting descriptor has significantly lower number of dimensions (176) than in other methods, yet more discriminative.

The contributions of our paper are as follows: First, we propose an architecture to automatically learn the region-specific aging features. The resulting aging descriptor is discriminative and robust to slight misalignments. Second, the conducted experiments show the discriminative power, efficiency, and generalizability (cross-dataset evaluation) of the proposed method. This suggests the suitability of the proposed method for real-time applications.

5.1.1 Related Work

In the last decade, many methods have been proposed to automatically estimate the age from the human face. Typical age estimation systems consist primarily of two steps; aging feature extraction and age classification/estimation. An elaborate survey on extracting and classifying aging features can be found in [38].

Shape features are used in the early days by Kwon and Labo [67] to estimate the age from infancy to adulthood. As face size changes during this period, measurements and ratios of distances between facial points are computed to estimate the age. Other works utilize the face shape to infer the age (in combination with appearance features).

[17, 21, 51, 68, 141] use Active Appearance Model (AAM) [22] to model the shape and the appearance. The model parameters are used to estimate the age. As the face shape stops developing at early adulthood, the shape-induced features are limited in providing cues about the age for later age stages.

Appearance-based features are utilized to derive the age from the wrinkles and the skin texture. Features like LBP [83, 133], Encoding-based Sampling [7], or Discrete Cosine Transformation [130] are used to primarily capture the skin texture changes and the fine wrinkles. For more pronounced wrinkles, gradient-based filters are used like Sobel [114, 122], Gabor filters [41], or Biologically-Inspired Features (BIF) [44, 53]. Like fine wrinkles features, these filters are convolved over the entire face [53] or constrained to specific regions [122]. Other approaches combined both types of features to produce a single age descriptor [21].

Recently, the focus of research efforts in computer vision have been shifted to deep models. Driven by the recent boosts in the scale of computational power and large amounts of labeled data, convolutional networks are utilized in face-related tasks like face detection and parsing [74, 85]. A similar approach to ours is employed to face alignment and facial points detection [111, 140]. Sun et al. [111] use three-level convolutional networks to estimate the positions of five facial landmarks. The first level contains three networks to estimate the rough locations of the eye centers, the nose, and the corners of the mouth. The second and the third level aim to refine the estimation of the facial landmarks. In a similar approach, Zhang et al. [140] use successive stacked auto-encoder networks to accurately (and gradually) locate the landmarks on the face. An advantage of their method is refining the landmarks collectively as opposed to [111], where the landmarks are estimated independently. We do not aim at landmark estimation.

In this paper, we apply deep convolutional networks to learn region-based features. The networks aim to learn the filters that fit best each face regions. The outputs of the networks are further fed to *combination* layer which allows merging the information from these networks. In such a case, the resulting age descriptor contains region-specific features but, still, allows interaction and weighting of the different regions.

5.2 Deep Aging Features

Our goal is to design a feature extractor that provides discriminative age features per face region. To this end, we aim at learning specific filter banks for different areas of the face. In this way, features extracted for each face region provides an estimation of the age. Then, on top of these per-region features, a combination layer is used to learn relationships between the different face regions to provide the final output.

5.2.1 Motivation

Aging cues may vary spatially depending on the part of the face. In fact, the type of feature to be extracted depends on the location of the regions in the face. Therefore, the aim is to provide small trainable networks which are fine-tuned for different areas of the face. As lesser data is required for training, smaller nets can used to avoid overfitting. In general, aging cues are dependent on the age range. Some approaches divide the age range into segments to perform multi-class classification by assigning each age-range to a different class. Optimally, there is a class for each different age. Other approaches consider the age as a continuous value and use regression to infer the age of the test sample. The former approach may fail due to differences within the age segment. Ages near the boundaries of each segment are closer than ages in the middle of the age segment. Using regression, these differences are not represented. On the other hand, a multi-class classifier considers different relationships between aging cues in the different areas of the image. Therefore, our approach takes the benefits of both methods by performing regression via multi-class classification using a piecewise regressor.

In general, a network is trained in a supervised way and then the last layer before the soft-max operation is used to provide the feature vector. This approach is suited even if the features do not contain useful information for the task at hand. The size of the feature vector depends on the complexity of the architecture. For large networks, methods use dimensionality reduction to generate a smaller fixed-sized vector. We depart from a different approach in which the last layer is used to explicitly encode the age of the sample and then regression is computed for training. The advantage is twofold. First, the dimension of the feature vector is fixed and does not depend on the number of kernels in the last convolutional layer. Second, the feature vector provides direct age information that can be used for higher level reasoning.

5.2.2 Region-Specific Features

Our region-specific feature is a CNN consisting of two convolutional layers with linear rectification units as activation neurons and average pooling operations. More specifically, the structure of our network is expressed as (18×18) – $(14 \times 14 \times N_{f1})$ – $(5 \times 5 \times N_{f2})$ - $(1 \times 1 \times N_{f3})$ - $(1 \times 1 \times 1)$, where the last unit is a continuous value representing the age. In this way, the neural network is based on regression. Regression has been used before for age estimation [38]. However, these approaches use multi-class classification by assuming that the number of classes is the same as the number of different ages in the dataset.

Face regions of interest are first resized to 64×48 and then partitioned using a 4×4 overlapping grid as shown in Figure 5.1. Each patch of size 18×18 is locally normal-

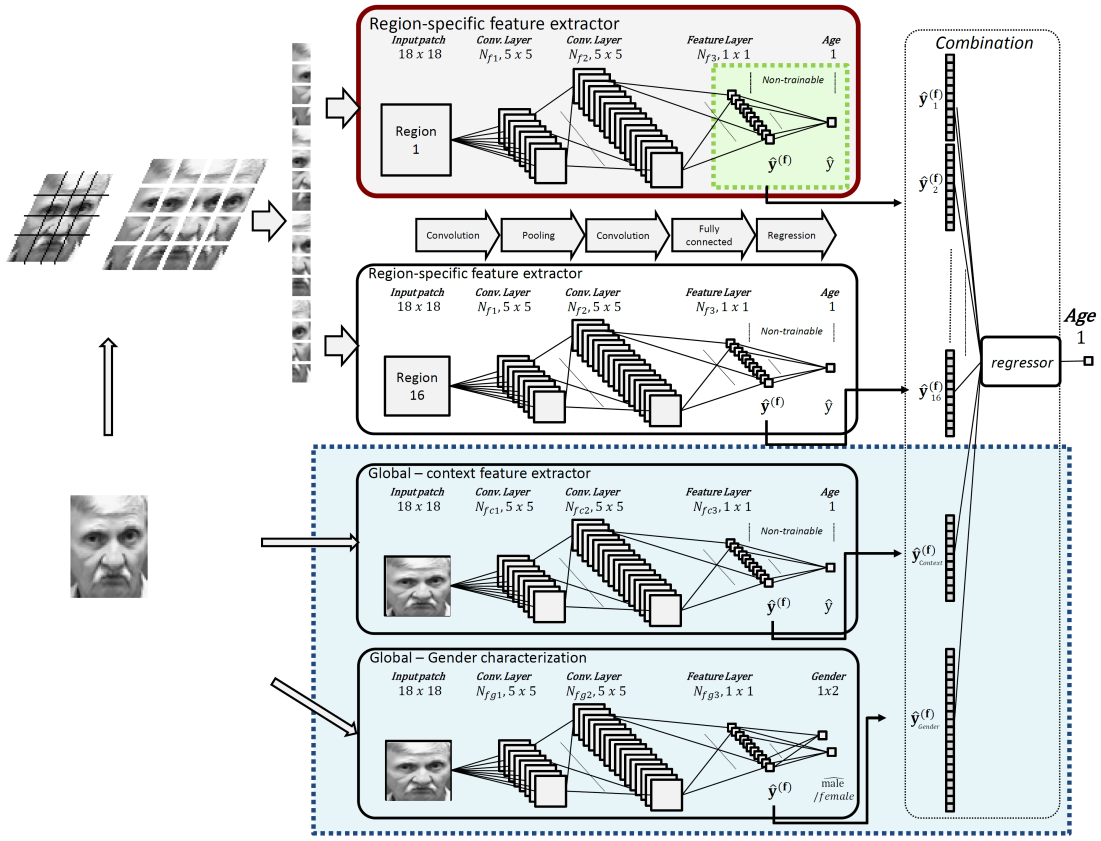


Fig. 5.1: Proposed deep aging features framework. The region-specific feature extractor is defined in the red box. Feature extractor uses piecewise regression for learning (green box – see Figure 5.2). In addition, the framework considers global appearance and gender characterization to predict the final age (blue box).

ized and used as input for a different network. In our experiments, the first convolutional layer is set to $N_{f1} = 24$ kernels of size 5×5 with rectifier linear units $(max(out, 0))$, followed by a max pooling operation of size 2×2. This pooling operation reduces the dimensionality and provides robustness to small displacements during the face alignment process. The second convolutional layer contains $N_{f2} = 72$ kernels of size 5×5 also with rectifier linear units and its output is fully connected to the feature layer consisting of $N_{f3} = 11$ kernels. The fully connected layer uses a dropout rate of 0.5. All these parameters are obtained empirically to result in a suitable trade–off between overfitting and discriminative power. Kernels are relatively small to capture fine details in facial features such as wrinkles. The number of kernels per layer is kept low to cope with datasets containing a relatively small number of training samples.

The cost function is the Euclidean distance between the age output (in years) of the net-

work and the ground-truth age of the training sample. The aim is to combine the last two layers to perform regression using piecewise linear decomposition. Usually, learning is done by a multiclass classifier where each possible age or age-range is represented by a different class. Using age-ranges has the drawback that the distances between the predicted and real ages are not represented. Other approaches apply regression to represent ages as continuous values. Our approach differs from previous methods. We consider aging features to be continues value and dependent on the age range. To this end, our feature extractor is trained using piecewise regression. The basic idea is to partition the age range into segments that potentially share similar properties. Regression is applied to each segment. This approach is based on the two last layers of the network. The last layer represents the age value while the feature layer encodes the age in segments. The feature layer is trainable and fixed to $N_{f3} = 11$ dimensions while the last layer is non-trainable and provides the desired output. The benefit of this structure is to provide suitable and compact features in the feature layer (Figure 5.1) and, at the same time, provide a continuous output for regression. The feature layer splits the age axis (real output) into age intervals defined by different thresholds $\Theta = \theta_1 < \theta_i, \ldots, < \theta_{N_{f3}}$ as shown in Figure 5.2. Given these segments, the age in the last layer is computed as:

$$\hat{y} = \sum_{k=1}^{N_{f3}} \theta_k \hat{y}_k^{(f)}, \tag{5.1}$$

where Θ is the set of weights in the last layer and $\hat{\mathbf{y}}^{\mathbf{f}} = [\hat{y}_1^f, \ldots, \hat{y}_{N_{f3}}^f]$ is the output of the feature layer. The loss function is defined as:

$$\mathcal{L}(W, X) = \sum_{i \in N} \|y_i - \hat{y}_i\|, \tag{5.2}$$

where $\hat{y}_i$ is the age estimated for the i-th sample using Eq. (5.1) and N is the number of samples in the training set. In this paper, we fixed the weights of the last layer Θ using the age ranges. More specifically, the age axis is partitioned into 11 segments (i.e., $\hat{\mathbf{y}}_{\mathbf{f}} \in \mathbb{R}^{11}$), the size of the segments is empirically fixed to $\Delta = 8$ years resulting in $\Theta = [3.5, 11.5, 19.5, 27.5, 35.5, 43.5, 51.5, 59.5, 67.5, 75.5, 83.5]$.

During training, for each training sample, we minimize the squared distance in the feature layer by $\|\mathbf{y}^{(\mathbf{f})} - \hat{\mathbf{y}}^{(\mathbf{f})}\|$. The real value (ground truth) in the feature layer $\mathbf{y}^{(\mathbf{f})}$ is computed as follows (Figure 5.2). First, the real age value is projected in the closest age segment ($i = \lceil \frac{y-\theta_1}{\Delta} \rceil$). This projection defines the entries of $\mathbf{y}^{(\mathbf{f})}$ since age segments correspond to consecutive dimensions in the feature vector. Finally, the two consecutive values of the feature vectors are computed using the distance between the projected age and the segment limits:

$$y_i^{(f)} = 1 - \frac{d_i}{\Delta}, i \in [\lceil \frac{y-\theta_1}{\Delta} \rceil, \lceil \frac{y-\theta_1}{\Delta} \rceil + 1] \tag{5.3}$$

where $d_i = \|\theta_i - y\|$ is the distance between the projected age and the limit of the corresponding age segment. A complete illustrative example is shown in Figure 5.2. Importantly, $\mathbf{y}^{(\mathbf{f})}$ directly encodes age information based on appearance features.

Holistic Age Features

Figure 5.1 outlines the proposed algorithm for age estimation. The first component of the pipeline is the region-specific feature extractor applied to different parts of the face. Features for each region are given by $\mathbf{y}^{(\mathbf{f})}$ which is the output after removing the non-trainable layer. Each feature directly encodes the estimated age for that patch. This is different from other methods where features provide support to a specific age (using multi-class classifiers) or simply represent the visual appearance. The framework includes a combination layer integrating the per-patch estimated age resulting in the final output. The input to the combination layer is a 176 dimensional vector constructed by concatenating the output of each patch (16 in our experiments). The framework also incorporates high level facial cues by fine tuning features of the entire image (blue box in Figure 5.1). The aim is to represent global (coarse) differences between faces at different ages (e.g. differences in gender or ethnicity may result in a different global appearance). The input image is cropped to the face area and resized to 64×48 and then partitioned by fixed 4×4 overlapping grids (Figure 5.1). Then features for each patch are extracted and concatenated leading to a 176 feature vector.

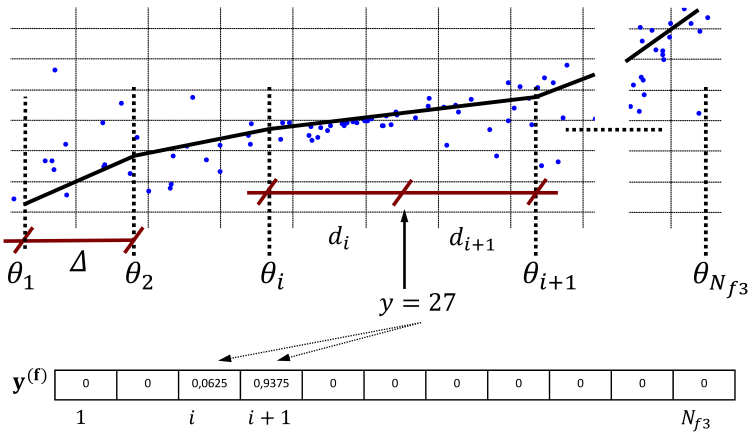


Fig. 5.2: The proposed region-specific feature encodes ages into the feature layer using piece-wise regression (corresponding to the green box in Figure 5.1.

5.3 Experiments

The goal of the proposed approach is to learn region-specific aging features using an architecture based on convolutional networks. The resulting features capture the age-telling cues for each region independently. Their learnt outputs are further fed into a combination layer which aims to integrate their information. In this section, we conduct a number of experiments to provide insights into the proposed architecture (Section 5.3.5) and evaluate the discriminative power, the efficiently, and the generalizability of the learnt features.

5.3.1 Datasets and Experimental Setting

Two datasets are used in our experiments. The first dataset is Morph-III [93] aging dataset. It is one of the largest aging dataset with accurate ground truth and around 200K images of more than 13K subjects. The images in other large-scale datasets are crawled from the internet with no ground truth [81]. The female to male ratio is around 1 to 6. The dataset contains facial images of subjects from different ethnicities (African, Caucasian, Asian, Hispanic, and others). However, the samples of African and Caucasian descendants constitute around 95% of the total dataset samples (73% and 22% respectively). The ages of the subjects range from 16 to 80 (Figure 5.4), however, around 99% of the samples are of ages less than 60 years.

The second dataset is FGNET [1] which contains 1002 facial images from 82 subjects of Caucasian descendant. The ages range from 0 to 69 years, however the age distribution is skewed to younger ages (around 70% less than 20 years). Figure 5.4 shows the age distributions of both datasets. Sample faces from both datasets are shown in Figure 5.3.

Our experiments are conducted primarily using Morph dataset due its large size. FGNET is used for validating our features in a cross-dataset experiment. For both datasets, the eye centers are detected and the face is registered and cropped. For faces where eyes are not detected, the samples are discarded. To isolate the ethnicity influence in our experiments, we conduct our experiments on Caucasian descendant samples resulting in 45878 samples. The dataset is split randomly into 41495 training and 4383 test sets. Figure 5.5 shows the distributions of the training and test sets. The performance is evaluated quantitatively by Mean Absolute Error (MAE) $\frac{1}{N}\sum_{i=1}^{N}|y^i - \hat{y}^i|$. Where y^i is the true age for the test sample i, $\hat{y}^i$ is the predicted age for the test sample i, and N is the number of the test samples.

[1] http://www.fgnet.rsunit.com

Fig. 5.3: Sample faces from Morph (up) and FGNET (down) datasets.

5.3.2 Region-Specific Feature Learning

In this experiment, we evaluate the features learnt by the proposed method. The network is configured as detailed in Sect. 5.2.2. Region-specific networks have the same configuration (18×18) – $(14 \times 14 \times 24)$ – $(5 \times 5 \times 72)$ - $(1 \times 1 \times 11)$ - $(1 \times 1 \times 1)$. The selection of the parameters aims at providing a trade off between discriminative power and training overfitting. On one hand, the network needs to be fine tuned for each specific region. On the other hand, the amount of training images is limited and therefore, larger networks would lead to overfitting. Different experiments have shown that this configuration leads to good trade off between training speed and accuracy. Gobal features are extracted using the same configuration. For the gender features, we use common supervised training on two classes (male / female) and then, features are extracted on the previous layer.

The features are tested with two learning methods; Support Vector Machine (SVM) and Random Forest. In addition to being widely used as the state-of-the-art methods, they are of different types. SVM uses a maximum margin approach while Random Forest is an ensemble-based learner. This is to illustrate the effectiveness of our features regardless of the learner used. The results for SVM are 4.17 (classification) and 4.04 (regression) while the errors produced by Random Forest are 4.13 (classification) and 3.87 (regression).

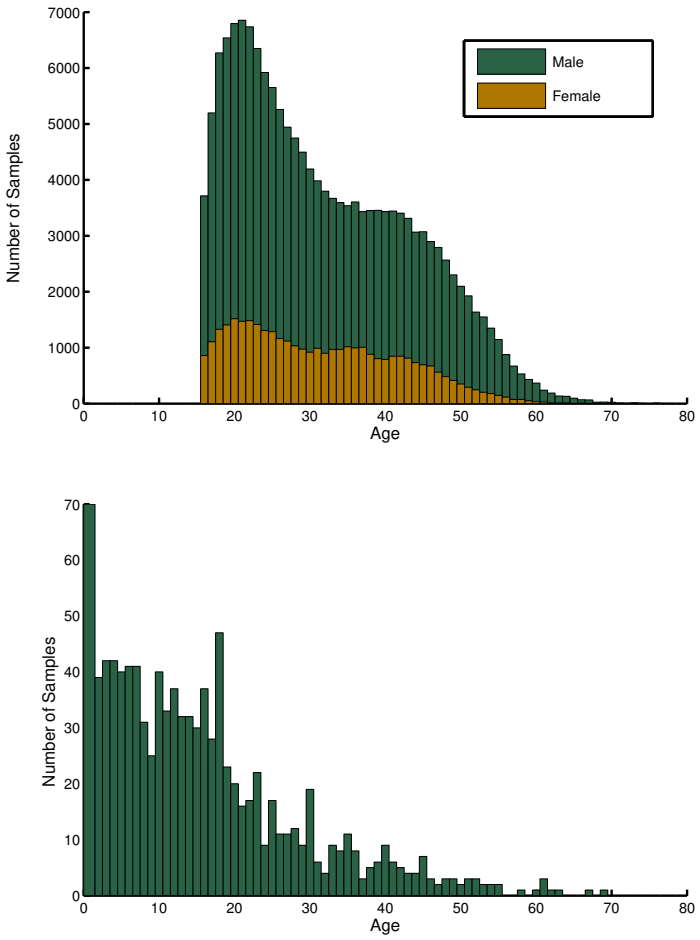


Fig. 5.4: Age distributions for MORPH (up) and FGNET (down) datasets.

5.3.3 Comparison with Other Methods

To further show the effectiveness of our architecture, we compare the proposed features with other state-of-the-art features. Namely, we extract LBP, BIF and Encoding-based Features (EBF) from the training and test sets and apply the same classification/regression setting used in Section 5.3.2. EBF learns the encoding of the intensity samples (EBF-S) or the gradient information (EBF-G). LBP and EBF-S are aimed to capture primarily the skin aging changes, while BIF and EBF-G are designed to capture the deep and apparent wrinkles on the face. The dimension of the BIF features is reduced using PCA (from 11080D) to 1000D since BIF produced large error using the unreduced dimension. As in the previous settings, SVM and Random Forest and utilized to estimate the age from the features. Tables 5.1 and 5.2 summarize the results of

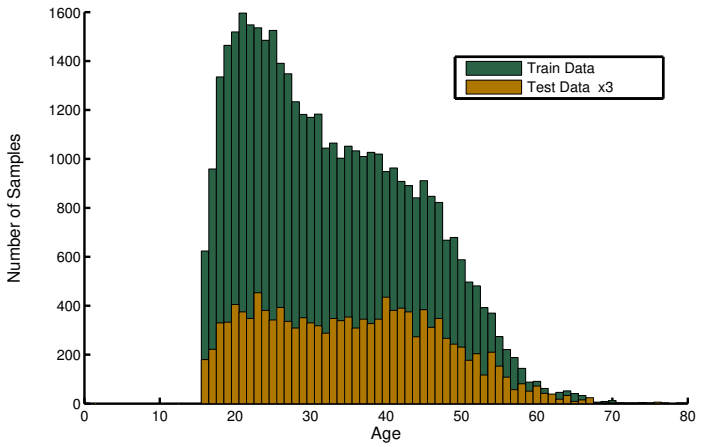


Fig. 5.5: Distributions of the training and test test. The test set distribution is scaled up 3 times for better comparison.

Tab. 5.1: Comparison between four types of features and the proposed features (Region-Specific Features (RSF)) using SVM: Local Binary Pattern (LBP), Bio-Inspired Features (BIF), Encoding-based Features EBF-S and EBF-G.

Method	Class.	Reg
LBP	5.08	4.73
EBF-S	4.23	4.14
EBF-G	4.52	4.47
BIF	4.29	4.15
RSF	**4.17**	**4.04**

Tab. 5.2: Comparison between four types of features and the proposed features (Region-Specific Features (RSF)) using Random Forest: Local Binary Pattern (LBP), Bio-Inspired Features (BIF), Encoding-based Features EBF-S and EBF-G.

Method	Class.	Reg
LBP	6.68	–
EBF-S	6.14	–
EBF-G	6.13	5.44
BIF	8.53	6.15
RSF	**4.13**	**3.87**

SVM and Random Forest respectively. The region-based features produces the lowest errors for both learning methods using classification and regression which proves the effectiveness of the proposed features regardless of the prediction method.

Tab. 5.3: Comparison between the proposed features (Region-Specific Features (RSF)) and four types of features (all reduced to 176 dimension): Local Binary Pattern (LBP), Bio-Inspired Features (BIF), Encoding-based Features EBF-S and EBF-G. SVR is used for prediction.

Method	176-D	All-D
LBP	5.19	4.73
EBF-S	4.76	4.14
EBF-G	4.87	4.47
BIF	4.70	4.15
RSF	**4.04**	**4.04**

5.3.4 Efficiency vs. Discriminative Power

The experiments in the previous section show the effectiveness of the proposed learnt features. In this section, we discuss the efficiency of the features in two aspects; feature extraction and estimation (time) cost. While the other features have relatively high dimension $>$ 1000 (1180, 5120, and 11080 for LBP, EBF, and BIF respectively), the dimension of the region-based features is 176. Besides that, the features are extracted in one pass of the architecture which is faster than the other features (pixel sampling (LBP), gradient computing (EBF), or banks of Gaussians with different orientations and scales (BIF) [2]. Efficiency is specially important on devices with limited computational power like wearable devices (e.g. Google Glass) where realtime processing is a must. To compare with other features in low-dimension scale, PCA is used to reduce the dimension of LBP, EBF-S, EBF-G, and BIF into 176. Table 5.3 shows the results using SVR (as it gives the best results for the other features). For LBP, EBF, and BIF, the error increases by at least 10% when the feature dimension is reduced to the same number as in our features. This is further suggests the compactness advantage of the proposed features over other features.

5.3.5 Contribution of the Combination Layer

Figure 5.6 shows the performance evaluation of the region-specific features applied to the different areas of the face. As shown, the larger error occurs in the upper part of the face. This area usually corresponds to hair and therefore are less discriminative for age estimation. On the other hand, higher accuracies are provided by the areas located in the second and the third rows of the face corresponding to areas around the eyes and nose. Given this accuracy distribution, an important advantage of our method is that specific areas of the image can be omitted to reduce computational cost. More specifically, we

[2]We do not report the computation time for extracting the features since different implementation might result in different computation times

omit the features from the first row and train an SVM regressor. The resulting error increased from 4.04 to 4.07 (less than 1%) while the dimensionality is reduced to 132.

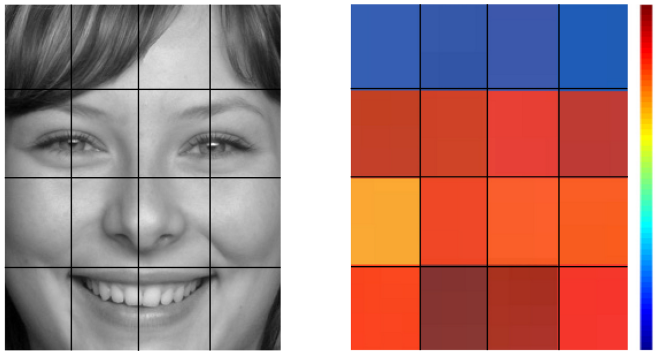


Fig. 5.6: Accuracy of region-specific features applied to different areas of the image (higher accuracy is represented by reddish colors while lower accuracy is represented by bluish colors). As shown, the accuracy decreases in the upper part.

Table 5.4 summarizes the performance of three instances of our deep aging proposal. First the combination of region-specific features. Second, region-specific features combined with global appearance features and finally, region-specific features, global features and gender features. As shown, including context improves the performance. The improvement is slightly better when gender features are considered.

Tab. 5.4: Comparison between different instances of our framework by including global features using Random Forest: the proposed features (Region-Specific Features (RSF)), features combined with global appearance information and the complete framework including gender features.

Combination	Class.	Reg
RSF	4.13	3.87
RSF + Global	4.03	3.81
RSF + Global + Gender	**3.97**	**3.78**

5.3.6 Generalizability: Cross-Dataset Evaluation

In this experiment the generalizability of our features is explored. We investigate how well our feature extraction architecture can do when applied to extract features from different dataset (cross-dataset evaluation). To this end, we estimate the age from samples in FGNET dataset using the previously explained models (each was learnt with different types of features extracted from Morph train subset). Please note that, here, the same architecture (filters) learnt from the Morph train subset is applied. In other words, the FGNET samples are used solely for testing and no retraining is performed. As shown

Tab. 5.5: Generalizability: cross-dataset evaluation on FGNET using the proposed features and four other types of features: Local Binary Pattern (LBP), Bio-Inspired Features (BIF), Encoding-based Features EBF-S and EBF-G. SVR is used for prediction.

Method	MAE
LBP	6.67
EBF-S	7.65
HA	6.42
BIF	8.13
RSF	**6.41**

in Figure 5.4, a significant number of samples in FGNET are of teenage group and younger while the age range in Morph starts from adulthood. Therefore, we evaluate only on samples with ages $>= 20$ years ($\sim$300 samples). Note that the age distributions of the Morph and the refined FGNET are still significantly different. The age ranges are only similar. Table 5.5 shows the results for different features. The best performance was achieved with our features which further suggests the robustness of the proposed features.

5.4 Discussion

We proposed a deep architecture to learn region-specific age features from the face. The focus here is primarily on the feature extraction with regards to discrimination power, efficacy and generalizability. To fairly compare with other feature-extraction methods, the experimental settings are applied to all compared features in the same manner. Two different classifiers/regressors are used; maximum-margin learner (SVM) and ensemble learner (Random Forest). Linear kernel for SVM and a default number of trees (1000) for Random Forest.

In terms of effectiveness, efficiency and generalizability, the proposed features outperform the other features. The fast feature extraction and the low-dimensionality of our age descriptor makes it especially suitable for real-time application such are wearable devices with limited processing power. By visualizing the contribution of each region-specific features, the dimensionality is further reduced to 132 while the error increases insignificantly (less than 1%).

The generalizability of the method is further evaluated by testing the features on the FGNET dataset in a cross-dataset evaluation. FGNET samples with ages 20 and higher are tested (see Section 5.3.6). The error produced is 6.41 years. While other reported errors (using methods trained and tested on FGNET) are smaller (just below 5 years), those errors are computed for the entire FGNET dataset (including the ones in [0-19]

range). The errors for age groups older than teenage on FGNET are higher than the ones in the teenage group. For example, Yan et al. [132] compute the error for each 10-year group. Although the error over all the dataset is 4.94 years, the error becomes 8.41 years when calculated over the rage [20-69]. Their error is even smaller than the errors (also calculated over the range [20-69]) produced by other methods they discuss in their papers (12.30 , 11.23, and 17 years). Please note that they report only the overall error and the per-group errors. However, it should be mentioned that their methods are trained on FG-NET where younger ages are more frequent than adult and elderly ages.

5.5 Conclusion

A novel approach is proposed to extract age features from facial images. The proposed approach extracts region-specific features as opposed to other methods, where the same filters are applied to the entire face. A convolutional network is assigned to each part of the face and a combination layer is employed to integrate the information from the convolutional networks.

The discriminative power, efficiency and generalizability of the proposed features are shown by our experiments. Compared to other state-of-the-art features, the proposed features produced smaller errors with different classifiers/regressors. Efficiency-wise, the resulting features are extracted fast and have a dimension of 176. A cross-dataset experiment to evaluate the generalizability where the model, based on our features, produced smaller errors than other models which are based on state-of-the-art features. The proposed features are suitable for real-time application especially for wearable devices with limited processing power.

Acknowledgment

The authors wish to acknowledge Jiayan Qiu for helpful discussions and for his contribution to the definition of the loss function (Section 5.2.2).

6

Age Estimation Under Changes in Image Quality: an Experimental Study

6.1 Introduction

Automatic age estimation is an important task in image processing as it has numerous applications in everyday life: i.e. security, surveillance, and online marketing. The human face reveals rich information about the age of a person and hence automatic age estimation systems are typically designed to predict the age from the human face. While aging, facial features change in response to muscle contractions and other biological changes of the skin. The facal skin becomes more leathery and rough while winkles start to appear and become more pronounced in certain face areas (e.g. around the eye and mouth corners). The shape of the face contains details indicating the age of the person. From infancy to early adulthood, the size of the head grows and certain measurements of the head size correspond to age.

A number of aging features have been designed or used to capture the aging cues. Some features [7, 83, 130, 133] are primarily applied to model the skin texture changes and the fine wrinkles on the face, while others [41, 44, 53, 114, 122] are mainly used to capture the pronounced wrinkles of the face. To measure the effectiveness of these features, they were used in age prediction pipelines.

The influence of image quality on the performance has long been a challenge in face-related image processing tasks. It is particularity important to address this challenge in age estimation since the aim here is to capture the skin texture and wrinkle changes. Moreover, the face images in real-life scenarios are taken using different capturing de-

vices and are prone to noise due to digital transmission and compression. It is this which makes it important to study the performance of aging features with varying image quality degradation.

Many research works propose or utilize features to extract the age. A thorough survey can be found in [38]. Kwon and Labo [67] first use the head size information to infer the age. They compute ratios and measurements of distances between facial points. Other methods [17, 51, 68, 141] adopt other variations of shape-based models to estimate the age. Since the shape models are limited to a certain aging period, appearance-based features are primarily utilized. They capture wrinkle changes and model the skin texture. Skin texture and fine wrinkles are represented by features like LBP [83, 133], Encoding-based Features [7] and Discrete Cosine Transformation [130]. Other gradient-based features like Gabor filters [41], Sobel [114, 122] and Biologically-Inspired Features (BIF) [44, 53] are adopted to detect pronounced wrinkles. However, since these features are designed or adjusted to capture certain aging cues, changes in image quality may introduce artifacts which influence these fine details and hence the output of these features. This suggests addressing the challenge posed by changes in image quality for age estimation.

In this paper, we investigate the effect of image quality on the performance of aging features. We mainly focus on appearance-based features. A number of age estimation datasets are used in our experiments. To investigate the robustness of the features, we simulate degrading of the image quality by applying different types of digital image noise and analyze the performance. Furthermore, we introduce a basic framework to automatically assign the best aging features based on the quality of the face image.

6.2 The Proposed Scheme

In this section, we explain the proposed scheme used to investigate the influence of image quality on aging features performance. A key point is to isolate the feature contribution from the other factors like preprocessing and learning steps. To this end, the datasets are reduced to similar sizes. The facial points are detected using the same landmarker and the faces are registered in the same manner. This leaves the image quality as a single variable to be tested.

To analyze the effect of the quality of images on the performance, we simulate artifacts in the datasets by applying different types of noise: JPEG-compression noise, quantization noise, and scaling noise. Each noise type is applied with different qualities to further analyze the performance. For JPEG-compression, the images are compressed with different compression rates. To simulate the quantization noise, the images are

quantized to different values (lower than the original 256 pixel values). Finally, the images are scaled down by different factors and then scaled up to simulate scaling noise.

Appearance aging features can be categorized into skin-texture features and pronounced-wrinkle features. However, these features differ significantly in size, number of parameters and efficiency. For certain purposes, some features are favored over others. For instance, LBP is known to be efficient and requires a small number of parameters. This may favor it over other features like BIF which requires dense convolution of multiple Gabor filters. In this work, we focus on the details captured by the features. A suitable feature extraction method for this purpose is the Encoding-based Features [7] where local features are extracted around certain positions (e.g. in a dense manner) and the features are quantized into pre-learnt patterns. This method has the advantage that the local features can vary depending on the task. If skin texture and fine wrinkles are to be captured, sampling-like local features are deployed, while gradient-based local features are used to capture more pronounced wrinkles. Consequently, this provides an adequate tool to test the performance of features for different image qualities.

6.3 Experiments

In this section, the datasets used in the experiments are first explained along with the experimental setup. We then discuss the features and the conducted experiments.

6.3.1 Datasets and Experimental Setup

Four publicly available dataset are used: FGNET [1], FACES [30], UvA-Nemo [26], and Morph [93]. The images in these datasets are collected from personal portrait (FGNET), or captured primarily to build the age-estimation dataset. FGNET is probably the most well-known publicly available dataset. It contains 1002 images of 82 subjects. The images are taken from personal portraits in different time periods. Morph dataset contains around 200K low quality images. A subset of 1K samples are selected. FACES dataset contains still images of 171 subjects showing six expressions. The imaging conditions are very good which allows the fine details of the face to be shown. In total, there are 1026 images in the dataset. Finally, UvA-Nemo dataset contains videos of 400 subjects showing happy expressions. The first frame, which shows a neutral face, is extracted from each video. The dataset is collected primarily for age and expression research purposes with fixed lighting conditions. Figure 6.1 shows example images of the four datasets.

[1] http://www.fgnet.rsunit.com

Fig. 6.1: Sample images from the datasets used in our evaluation: Morph (top left), FGNET (top right), FACES (down left), and UvA-Nemo (down right).

A facial landmarker is employed to detect the eye centers which are used to register the face. The face is then cropped to the size of 125×100. All faces are converted to gray scale. The datasets are further divided into three folds of similar sizes. The identities of the subjects are mutually exclusive between the folds and the age distributions in the folds are aimed to be as similar as possible.

In Encoding-based Features, a local feature vector is calculated around each pixel. Then, the features are encoded with a pre-learnt visual codebook. A PCA tree is used to build a 256-code visual dictionary. The type of details captured is then determined by the choice of the local features. We utilize two types of features. First, sampling features where the intensities of 25 points around the pixel are sampled. This aims to capture the skin texture and fine wrinkles. Second, gradient-based features. Here, the histogram of gradient directions are computed in a 8×8 area around the pixel. This is used to model the intensity and the direction of the apparent wrinkles. The gradients are computed using Gaussian derivatives with different sigma (width) values: 0.5, 0.75, 1.0, and 1.25. The face is divided into 7×5 patches and a histogram of the visual codes is calculated for each patch. The patch-based histograms are concatenated to form the aging descriptor.

Finally, an SVM classifier is employed to learn and predict the age. A linear kernel with $C = 1$ is applied in all experiments for fair comparison and to limit the influence of the learning process. Three-fold cross-validation is used to report the results. Mean Absolute Error (MAE) is reported for the evaluation of the performance quantitatively where $MAE = \frac{1}{N}\sum_{i=1}^{N}|y^i - \hat{y}^i|$. y^i is the true age for the test sample i, $\hat{y}^i$ is the predicted age for the test sample i, and N is the number of the test samples.

In the following experiments, we investigate the influence of image quality degradation. More specifically, we explore the influence of three different types of image noise: JPEG

Fig. 6.2: Variations of an image when JPEG-compressed with different qualities (from left to right): 100, 75, 50, 25, and 10. Where 0 is the lowest quality and 100 is the highest quality (no compression).

compression noise, quantization noise, and image scale noise. For each of these noise types, the aging features are evaluated and analyzed.

6.3.2 Compression Noise

In this experiment, we explore the performance for images contaminated by compression-related artifacts. The images are compressed using the lossy JPEG compression with different qualities (0 is the lowest and 100 is the highest quality (no compression)). Figure 6.2 shows the variations of an image from FACES dataset when compressed with qualities 75, 50, 25 and 10.

The features are evaluated on the compressed datasets and the results are reported in Table 6.1. A performance is considered better than another if the error is reduced by no less than 0.1 otherwise the two performances are deemed comparable. For 75-quality and 50-quality compressions, the best results are obtained with gradient features with sigma equals to 0.5 when applied on FACES and Morph datasets. For other datasets, the performance of sampling and gradient features are comparable. When 25-quality compression is applied, gradient features with larger sigma values (0.75-1.25) give better results. This is because larger filters are required to smooth and capture the wrinkles since fine details disappear gradually. With 10-quality compression, even larger sigma values provide better gradient features than the ones with smaller sigma values or the sampling features. The results show that the more artifacts the image contains, the wider the filter is preferred to capture the aging cues.

6.3.3 Quantization Noise

The image intensities are typically quantized into 256 values. However, fewer quantized values can be used to reduce the image size which introduces quantization noise. In this experiment, we vary the quantized values (from 256) to 64, 32, 16, 12, and 8. Figure

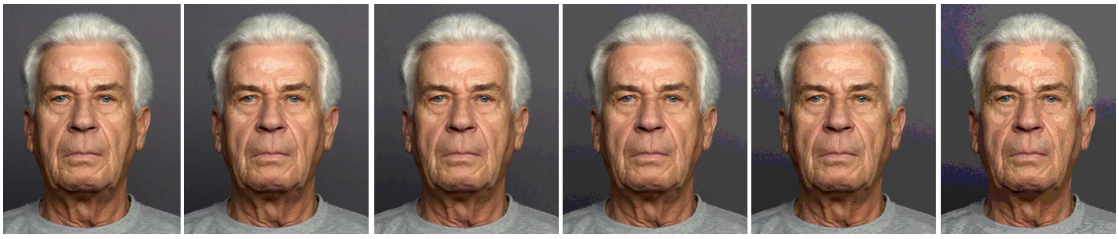


Fig. 6.3: Variations of an image when quantized to different values (from left to right): 256 (original) 64, 32, 16, 12, and 8.

Fig. 6.4: Variations of an image when scaled down with different factors (from left to right): 1 (original), 2, 4, 6, and 8.

6.3 shows the variations of an image from FACES dataset when different quantization rates are applied.

The performance of the aging features for different quantized values are shown in Table 6.4. The quantization noise gradually erase the aging details. Similar to the behavior with JPEG-compression noise, the more quantization noise the image has, the wider the filter is preferred to produce the best performance.

6.3.4 Scaling noise

We refer to the noise when scaling up an image (from a low resolution) as the scaling noise. In this experiment, we simulate scaling noise by scaling down the images by different factor and then scaling up to the original size. More specifically, the images are scaled down by 2, 4, 6, and 8 factors. Figure 6.4 shows an image from FACES dataset when scaled down with different factors.

Table 6.3 shows the performance of the aging features when scaling noise is introduced. Unlike, JPEG-compression and quantization noise, the aging features show similar behavior. i.e. all the aging features produce degrading performance and no clear preference can be suggested when scaling noise is introduced. This is because this type

Tab. 6.1: Evaluation of aging features on FGNET, FACES, UvA-Nemo, and Morph datasets when JPEG-compressed with different qualities: 75, 50, 25, and 10 (0 is the lowest quality and 100 is the highest quality (no compression)). The number after the dataset name refers to the compression quality. e.g. FACES-75 is the dataset FACES when compressed with 75 quality. SAM refers to sampling local features and GR050, GR075, GR100, GR125 refer to the gradient-based local features for sigma values of 0.5, 0.75, 1.0, and 1.25 respectively.

Dataset	SAM	GR050	GR075	GR100	GR125
FGNET-100	7.33	7.49	7.52	7.36	7.44
FACES-100	11.05	**9.15**	10.34	10.31	11.23
UvA-Nemo-100	**6.55**	6.85	6.77	7.08	7.46
Morph-100	6.70	6.64	6.92	6.74	6.86
FGNET-75	7.39	7.50	7.48	7.49	7.39
FACES-75	11.25	**9.80**	10.21	10.76	10.82
UvA-Nemo-75	6.77	7.04	6.69	7.09	7.66
Morph-75	6.82	**6.63**	7.06	**6.67**	6.88
FGNET-50	7.56	7.48	7.48	7.55	7.54
FACES-50	10.76	**10.06**	10.21	10.97	10.99
UvA-Nemo-50	7.11	7.26	7.13	7.36	7.57
Morph-50	6.89	**6.46**	7.00	6.87	6.83
FGNET-25	7.81	**7.47**	7.65	**7.48**	7.58
FACES-25	11.98	**10.42**	11.30	11.15	11.40
UvA-Nemo-25	8.03	7.76	**7.29**	**7.36**	7.44
Morph-25	7.05	6.98	6.85	6.76	**6.57**
FGNET-10	8.89	8.31	8.06	**7.86**	**7.89**
FACES-10	13.64	12.86	**12.19**	12.32	12.32
UvA-Nemo-10	9.73	9.38	9.37	9.35	**8.83**
Morph-10	7.54	7.38	7.44	7.23	**6.98**

of noise affect the aging details in a similar manner and, hence, the aging features are negatively affected similarly.

6.3.5 Automatic Feature Assignment

In the previous section, our experiments show how the performance of the aging features depends on the quality of the images. In this experiment, we aim to automatically assign the most suitable features to each test sample based on the image quality. We restrict our experiment to JPEG-compression noise since it is the most common noise among the three tested types. More specifically, we train a classifier to detect the qual-

Tab. 6.2: Evaluation of aging features on FGNET, FACES, UvA-Nemo, and Morph datasets when images are quantized with different values: 256(original), 64, 32, 16, 12, and 8. The number after the dataset name refers to the number of the quantized values. e.g. FACES-32 is the dataset FACES when images are quantized to 32 values.

Dataset	SAM	GR050	GR075	GR100	GR125
FGNET-256	7.33	7.49	7.52	7.36	7.44
FACES-256	11.05	**9.15**	10.34	10.31	11.23
UvA-Nemo-256	**6.55**	6.85	6.77	7.08	7.46
Morph-256	6.70	6.64	6.92	6.74	6.86
OKFGNET-64	7.39	7.40	7.56	7.51	7.46
FACES-64	11.06	**9.54**	10.47	10.79	11.14
UvA-Nemo-64	**6.58**	7.03	6.81	7.14	7.44
Morph-64	6.72	6.86	6.90	6.78	6.70
OKFGNET-32	7.71	7.65	7.62	7.58	7.52
FACES-32	11.18	**9.52**	10.35	11.00	10.97
UvA-Nemo-32	7.09	**6.68**	6.99	7.11	7.46
Morph-32	6.95	6.91	6.86	6.70	6.83
OKFGNET-16	8.43	**7.63**	7.80	7.64	7.69
FACES-16	10.73	**9.51**	10.48	10.54	10.74
UvA-Nemo-16	8.55	7.60	**7.17**	7.22	7.53
Morph-16	7.28	6.85	7.14	**6.77**	6.88
OKFGNET-12	9.14	7.97	7.80	7.63	**7.55**
FACES-12	10.69	**10.06**	10.71	10.58	10.67
UvA-Nemo-12	8.99	7.95	7.46	**7.41**	7.73
Morph-12	7.40	7.04	7.10	7.07	**6.99**
OKFGNET-8	10.55	8.46	8.73	8.44	**8.15**
FACES-8	11.96	**10.03**	11.57	11.08	11.51
UvA-Nemo-8	10.78	8.77	8.29	8.49	**8.23**
Morph-8	8.15	7.06	7.15	**7.00**	7.21

ity of the image and then apply the appropriate age estimator (leaned on the features most suitable for the detected quality). The results are shown in Table 6.4. The results here are reported for the dataset and its compressed variations (the original with the four compressed variations). We compare the obtained performance with the performance when using a single (best) feature type. For example, for FACES dataset, the single best feature type is GR050 and hence we compare the proposed method against these features for FACES dataset. Regarding UvA-Nemo dataset, GR100 features are compared against and so on. The results show that better or comparable performance is obtained with assigning the most suitable features in comparison to applying a single

Tab. 6.3: Evaluation of aging features on FGNET, FACES, UvA-Nemo, and Morph datasets when images are scaled down with different factors: 1(original), 2, 4, 6, and 8. The number after the dataset name refers to the the downscaling factor. e.g. FACES-4 is the dataset FACES when images are scaled down by factor 4.

Dataset	SAM	GR050	GR075	GR100	GR125
FGNET-1	7.33	7.49	7.52	7.36	7.44
FACES-1	11.05	**9.15**	10.34	10.31	11.23
UvA-Nemo-1	**6.55**	6.85	6.77	7.08	7.46
Morph-1	6.70	6.64	6.92	6.74	6.86
OKFGNET-2	7.41	**7.27**	7.58	7.40	7.49
FACES-2	11.87	**10.29**	10.51	11.31	11.29
UvA-Nemo-2	**6.95**	7.10	7.17	7.30	7.52
Morph-2	6.84	**6.54**	6.98	6.85	6.93
OKFGNET-4	7.55	**7.43**	7.59	7.70	7.71
FACES-4	13.38	**11.71**	11.95	12.52	12.32
UvA-Nemo-4	8.14	8.02	8.09	**7.82**	7.86
Morph-4	7.58	**7.25**	7.42	7.46	7.57
OKFGNET-6	7.88	7.92	8.04	8.20	7.93
FACES-6	14.41	**13.39**	13.35	13.38	13.39
UvA-Nemo-6	8.67	8.79	8.73	8.93	8.81
Morph-6	8.06	7.97	8.24	7.92	8.10
OKFGNET-8	**7.88**	8.16	8.34	8.00	8.29
FACES-8	14.78	13.58	**13.48**	14.11	13.64
UvA-Nemo-8	**9.18**	9.31	9.73	9.57	9.80
Morph-8	8.36	8.39	8.68	8.52	8.37

Tab. 6.4: The performance when automatically assigning the most suitable aging features based on the image quality. The results are compared against the single best feature type for each dataset. Here, each dataset includes also its compressed variations (the original dataset and the four compressed variations).

Dataset	Best Single Feature Type	Automatically Assigned Features
FGNET	7.57	7.57
FACES	10.46	10.43
Uva-Nemo	7.45	**7.31**
Morph	6.81	**6.77**

(best) feature type for all image qualities.

6.4 Conclusion

In this paper, the influence of image quality on aging feature performance is investigated. The motivation here is that changes in image quality affect the aging cues which primarily capture the skin texture and the wrinkles. A scheme was proposed to isolate the contribution of the features while changing the image quality. Four age estimation datasets were experimented on. The performance was studied when different types of noise artifacts were applied to the datasets.

Finally, we introduced a framework to automatically, based on the image quality, apply the most suitable feature type. The image qualities are automatically predicted. Our results show better or comparable performance when automatically applying different features, based on image quality, in comparison to a single (best) feature type.

7

Calibration-Free Gaze Estimation Using Human Gaze Patterns

7.1 Introduction

Gaze estimation is the process of determining where a person is looking at in a predefined plane. It is an important task in computer vision and has numerous applications in everyday life: i.e. human-computer interaction, assisting disabled users (e.g. eye typing) [76], and human behavior analysis [105].

In general, gaze estimation methods fall into two categories: 1) appearance-based methods [57, 72, 124] and 2) 3D-eye model-based methods [19, 49, 50, 128]. The former class extracts features from images of the eyes and map them to points on the gaze plane (i.e. gaze points). The latter tries to construct a 3D model of the eye and estimates the visual axis. Then, the intersection of the axis and the gaze plane determines the gaze point. Regardless of which gaze estimation method is used, a calibration procedure is needed. The calibration can be camera-based (estimating the camera parameters), geometric calibration (estimating the relationships between the scene components like the camera, the gaze plane, and the user), personal calibration (determining the angle between visual and optical axes), or gaze mapping correlation [58]. An extensive overview of the different approaches of gaze estimation and calibration can be found in [58].

3D-eye models require special equipment like cameras with multiple light sources and infrared. The costs and the strict requirements for their use (infrared, for example, is not reliable when used outdoors) limit their range of applicability. On the other hand, appearance-based approaches are less accurate than 3D-eye-models and less invariant

to head pose changes. Yet, low-cost cameras are common and sufficient for appearance-based approaches which makes them suitable for applications where high accuracy is not essential. Consider for example an application of people looking at advertisements for marketing research. Asking each participant to buy dedicated cameras or to do the experiment in the lab is time consuming and costly. Because low-cost cameras are integrated in almost every laptop or tablet nowadays, appearance-based methods are more suitable in such an application.

Besides the choice of the recording equipment, the adopted approach allows for a certain level of flexibility in the setup and the calibration. During calibration, users are usually asked to fixate their gaze on certain points while images of their eyes are captured. This procedure is cumbersome and sometimes impractical. In case of, for example, tracing costumers attention in malls, estimating the gaze points or regions should be done passively. Hence, some approaches propose methods to reduce the number of calibration points. However, in the case of passive gaze estimation, the calibration should be done completely automatically without an active calibration procedure imposed on the user.

Some recent studies focus on visual saliency information in images and videos to avoid applying active human calibration. Sugano et al. [109, 110] treat saliency maps extracted from videos as probability distributions for gaze points. Gaussian process regression is used to learn the mapping between the images of the eyes and the gaze points. Chen and Ji [19] use 3D models of the eye and incrementally estimate the angle between the visual and the optical axes by combining the image saliency with the 3D model. The argument for using saliency is that people look at salient regions with higher probability than other regions. However, as shown in [65], the computational saliency models do not frequently match the actual human saccades (Figure 7.1). In our previous paper [6], we propose that the gaze patterns of several viewers provide important cues for the auto-calibration of new viewers. This is based on the assumption that humans produce similar gaze patterns when they look at a stimulus. The assumption is supported by Judd et al. [65], where the authors show that fixation locations of several humans are strongly indicative, in general, of where a new viewer will look at. To the best of our knowledge, our work is the first to use human gaze patterns in order to auto-calibrate gaze estimators.

We present a novel approach to auto-calibrate gaze estimators based on the similarity of human gaze patterns. In addition, we make use of the topology of the gaze points. Consider, in a fully uncalibrated setting, a person who follows a stimulus from left to right. It would be difficult to indicate where the gaze points are on the gaze plane. However, their relative locations can still be inferred and used for auto-calibration. In a fully uncalibrated setting, when a new subject looks at a stimulus, initial gaze points are inferred. Then, a transformation is computed to map the initial gaze points to match the gaze patterns of other users. In this way, we use all the initial gaze points to match

Fig. 7.1: (Taken from [65]). Examples where saliency models do not match the human fixations. Bright spots indicate the saliency model predictions and the red dots refer to the human gaze points.

the human gaze patterns instead of using each gaze point at the time. Consequently, the transformed points represent the auto-calibrated estimated gaze points.

The rest of the paper is organized as follows. The proposed method is explained in Section 2. Next, we describe the experimental setup and evaluation in Section 3. The results are discussed in Section 4. Finally, the conclusions are given in Section 5.

7.2 Calibration-Free Gaze Estimation Using Human Gaze Patterns

We build upon the observation that gaze patterns of individuals are similar for a certain stimulus [65]. Although, there is no guarantee that people always look at the exact same regions, human gaze patterns will provide important cues about the locations of the gaze points of a new observer. The pipeline of the proposed method is as follows: when a new user is looking at a stimulus, the initial gaze points are computed first. Then, a transformation is inferred which maps the initial gaze points to gaze patterns of other individuals. In this work, we consider transformations which combine translation and scaling (per dimension). Including other transformations like rotation or shearing may

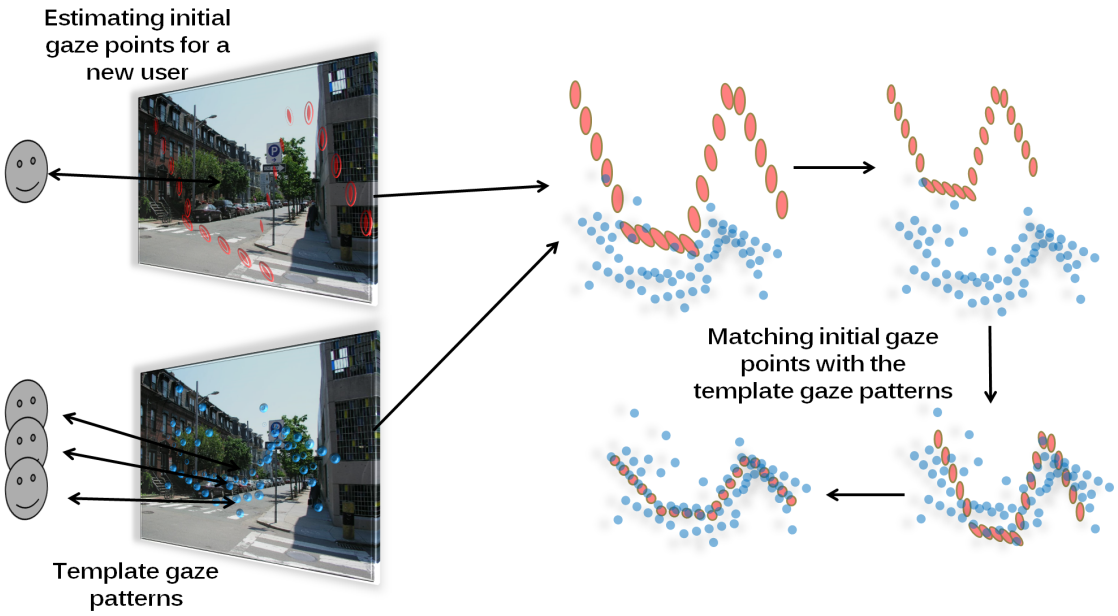


Fig. 7.2: Graphical illustration of the proposed method. Template gaze patterns refer to the gaze points of other individuals for the same gaze plane (display). When a new user looks at the stimulus, his or her initial gaze points are first estimated which preserves the relative locations between the gaze points. These points are transformed so that they match the template gaze patterns.

yield better mapping. However, they are not taken into account, since 1) translation and scaling are more common for gaze estimation, and 2) to reduce the search space. Figure 7.2 illustrates the pipeline.

7.2.1 Initial Gaze Points

The final gaze points should eventually match the human gaze patterns. However, we need to start from an *initial* estimation of the gaze points. Hereafter, we present two methods to achieve this: estimation of initial gaze points from eye templates and estimation based on 2D-manifold.

Eye templates

In this approach, the eye images of a (template) subject are captured while fixating the eyes on points on a gaze plane. The images of the eyes of a new user are captured and compared with the template eye images. The idea is to reconstruct eye images based

on the eye image templates. Note that here the eye templates are captured once for a single subject. When a new subject uses the gaze estimator, his or her eye images are compared with the already-collected eye templates. This is different from the traditional calibration-based gaze estimator where the eye templates are captured and stored for each subject and/or each different setting. The process can be performed at the raw intensity level or at the feature level. We will refer to both eye image representations as feature vectors. Consider $\{\mathbf{t}_i\}$ to be the template feature vectors, and $\{p_i\}$ denotes the corresponding gaze points. Furthermore, $\{w_i\}$ corresponds to the computed weights to reconstruct the feature vector of a new eye image $\hat{\mathbf{t}}$:

$$\hat{\mathbf{t}} = \sum_i w_i \mathbf{t}_i \qquad s.t. \qquad \sum_i w_i = 1. \tag{7.1}$$

Then the corresponding gaze point $\hat{p}$ for $\hat{\mathbf{t}}$ is calculated as follows:

$$\hat{p} = \sum_i w_i p_i. \tag{7.2}$$

To find the weights $\{w_i\}$, Tan et al. [116] suggest to first select a subset of $\{\mathbf{t}_i\}$ where the first and the second neighbors of the sample are used for training. The weight values are then computed as in [94]. Lu et al. [72] select only the direct neighbors as a training subset. Here, we select only the direct neighbors as in [72].

For a new user, potentially in a different unknown scene setup, the initial gaze points will be incorrect (without calibration). However, the relative locations between the gaze points are preserved.

2D manifold

In their work [72], Lu et al. find that the (template) eye features correspond to a 2D manifold while retaining most of the important information about the relative eye movements. The reason is that the eyes move, in the appearance-based representation, in two degrees of freedom. Figure 7.3 shows the projection of features of nine eye images on a 2D manifold and their corresponding nine gaze points on the gaze plane. It can be derived that the feature projections preserve the relative locations of the corresponding gaze points.

The 2D manifold can be obtained by projecting the template features on the first two principal components. However, the locations on the 2D manifold may be interchanged, transposed, or rotated when compared with the corresponding gaze points. For example, when the eyes move mainly vertically, the first principal component represents the

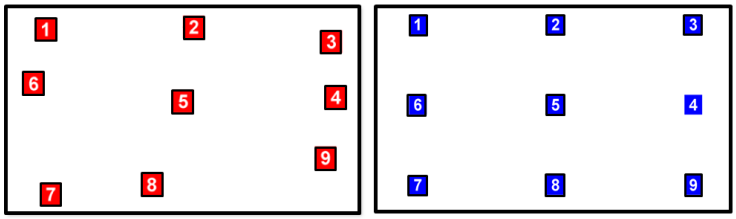


Fig. 7.3: The projection of features of 9 eye images on a 2-D manifold (red, left) and the positions of the corresponding gaze points on the gaze plane (blue, right). The 2D manifold is computed using 800 eye images corresponding to various locations on the gaze plane.

pupil changes on the Y dimension and the second principal component represents the X dimension. Hence, the projected locations need to be transposed. As this step is performed once offline, the projected locations are checked once and transformed to match the corresponding gaze points locations. As in the eye templates method, this procedure is followed once with a single (template) subject. When a new user looks at a stimulus, the eye features are projected on the offline-learned 2D manifold and the projected values are treated as initial gaze points.

The previous two methods (eye templates and 2D manifold) provide a way to find the initial gaze points. In the next section, we explain how to map these points to match the template (human) gaze patterns.

7.2.2 Gaze Points Mapping

Judd et al. [65] show that the fixation points of several humans correspond strongly with the gaze points of a new user. We aim to exploit this observation to perform calibration without the need for active user participation. To this end, we transform the initial (uncalibrated) gaze points so they match the template gaze patterns for a stimulus. By applying the aforementioned transformation, we aim to transfer the gaze points to their correct positions without explicit calibration. We present two different methods to find the transformation. Let the set $\mathbb{P} = \{\mathbf{p^1}, \mathbf{p^2}, ...\mathbf{p^M}\}$ denotes the gaze patterns of M users (hereafter, we call them *template gaze patterns*) where $\mathbf{p}^u = \{p_1^u, p_2^u, ...p_{S_u}^u\}$ consists of the S_u gaze points of user u. Let $\mathbf{p} = \{p_1, p_2, ...p_S\}$ be the initial gaze point set for a new user. The following two methods aim to transform and hence match $\mathbf{p}$ with the template gaze patterns $\mathbb{P}$.

K-closest points

This method tries to find the best mapping which minimizes the sum of distances of each point $p_j \in \mathbf{p}$ to its K closest neighbors of $\mathbb{P}$. Assume Φ is the set of all mappings. The method tries to find a mapping $\bar{\phi} \in \Phi$ which satisfies:

$$\bar{\phi} = \arg\min_{\phi} \mathcal{L}(\mathbf{p}, \mathbb{P}, \phi), \tag{7.3}$$

where:

$$\mathcal{L}(\mathbf{p}, \mathbb{P}, \phi) = \sum_{j=1}^{S} \sum_{k=1}^{K} \|\phi(p_j) - N(\phi(p_j), \mathbb{P}, k)\|. \tag{7.4}$$

$N(p_j, \mathbb{P}, k)$ is the k closest point from $\mathbb{P}$ to p_j. $\bar{\phi}$ is the computed mapping and $\bar{\mathbf{p}} = \bar{\phi}(\mathbf{p})$ represents the mapped auto-calibrated gaze points. Note that we try to match the initial gaze points $\mathbf{p}$ with all the gaze patterns in $\mathbb{P}$ simultaneously. To find $\bar{\mathbf{p}}$ and $\bar{\phi}$, we adopt a gradient-descent approach. To search for a local minimum (or maximum) using gradient descent methods, first, the gradient of the objective function is computed w.r.t to the corresponding parameters. Second, the parameters step toward the negative (positive) direction of the gradient in case of cost (reward) function. These two steps are repeated multiple times (epochs). We restrict the transformation to translation and scaling as discussed in Section 7.2. The transformation of a point $p = \begin{bmatrix} x \\ y \end{bmatrix}$ by $\phi = [s_1, s_1, h_1, h_1]$ is $\phi(p) = \begin{bmatrix} s_1 & 0 \\ 0 & s_2 \end{bmatrix} \cdot p + \begin{bmatrix} h_1 \\ h_2 \end{bmatrix} = \begin{bmatrix} s_1.x + h_1 \\ s_2.y + h_2 \end{bmatrix}$.

Here, we assume the origin to be the mean of $\mathbf{p}$. The parameter set ϕ is updated based on the derivative of the cost function $\mathcal{L}$ w.r.t ϕ:

$$\phi \leftarrow \phi - \gamma \nabla_{\phi} \mathcal{L}, \tag{7.5}$$

where γ is the learning rate and:

$$\nabla_{\phi} \mathcal{L} = \begin{bmatrix} \frac{\partial \mathcal{L}}{\partial h_1} \\ \frac{\partial \mathcal{L}}{\partial h_2} \\ \frac{\partial \mathcal{L}}{\partial s_1} \\ \frac{\partial \mathcal{L}}{\partial s_2} \end{bmatrix}. \tag{7.6}$$

The derivative w.r.t h_1 is computed as follows:

$$\frac{\partial \mathcal{L}}{\partial h_1} = \sum_{j=1}^{S} \sum_{k=1}^{K} \frac{\partial \|\phi(p_j) - N(\phi(p_j), \mathbb{P}, K)\|}{\partial h_1}. \tag{7.7}$$

Let $N(\phi(p_j), \mathbb{P}, K) = \{g_1, g_2, ...g_K\}$, then:

$$\frac{\partial \mathcal{L}}{\partial h_1} = \sum_{j=1}^{S} \sum_{k=1}^{K} \frac{\partial \sqrt[2]{(\phi(p_j)_x - g_{x,k})^2 + (\phi(p_j)_y - g_{y,k})^2}}{\partial h_1}, \tag{7.8}$$

$$\frac{\partial \mathcal{L}}{\partial h_1} = \sum_{j=1}^{S} \sum_{k=1}^{K} \frac{s_1.p_{x,j} + h_1 - g_{x,k}}{\sqrt[2]{(\phi(p_j)_x - g_{x,k})^2 + (\phi(p_j)_y - g_{y,k})^2}}. \tag{7.9}$$

And:

$$\frac{\partial \mathcal{L}}{\partial s_1} = \sum_{j=1}^{S} \sum_{k=1}^{K} \frac{s_1.p_{x,j}^2 + h_1.p_{x,j} - g_{x,k}.p_{x,j}}{\|\phi(p_j) - g_k\|}. \tag{7.10}$$

$\frac{\partial \mathcal{L}}{\partial h_2}$ and $\frac{\partial \mathcal{L}}{\partial s_2}$ can be derived in a similar manner.

Mixture model

In the K-closest points method, the matching is measured by the distances between each point of the initial gaze set and its closest neighbors in the template gaze patterns. Here, the initial gaze points are mapped to match a mixture model which is fit to the template gaze patterns. More specifically, we first model the template gaze patterns by a Gaussian mixture model. Next, the initial gaze points are transformed so that the probability density function of the transformed points is maximized. Formally, the method searches for a mapping $\bar{\phi} \in \Phi$ so that:

$$\bar{\phi} = \arg\max_{\phi} \sum_{j=1}^{S} pdf(\phi(p_j)), \tag{7.11}$$

where:

$$pdf(p) = \sum_{k=1}^{K} \omega_k \mathcal{N}(p|\mu_k, \Sigma_k), \tag{7.12}$$

and:

$$\mathcal{N}(p|\mu,\Sigma) = \frac{1}{\sqrt{(2\pi)^2|\Sigma|}}\exp(-\frac{1}{2}(p-\mu)^T\Sigma^{-1}(p-\mu)). \tag{7.13}$$

K is the number of model components, ω_k is the mixing coefficient of the k_{th} Gaussian component $\mathcal{N}(\mu_k,\Sigma_k)$ with μ_k mean and Σ_k covariance matrix. $\bar{\phi}$ is computed again by a gradient descent approach. The parameter set ϕ is updated as follows:

$$\phi \leftarrow \phi + \gamma\nabla_\phi\mathcal{F}, \tag{7.14}$$

where $\mathcal{F}$ is the reward function we aim to maximize:

$$\mathcal{F} = \sum_{j=1}^{S}\sum_{k=1}^{K}\omega_k\mathcal{N}(\phi(p_j)|\mu_k,\Sigma_k). \tag{7.15}$$

The derivative w.r.t h_1 is computed as follows:

$$\frac{\partial\mathcal{F}}{\partial h_1} = \sum_{j=1}^{S}\sum_{k=1}^{K}\omega_k\frac{\partial\omega_k\frac{1}{\sqrt{(2\pi)^2|\Sigma_k|}}\exp(-\frac{1}{2}(\phi(p_j)-\mu_k)^T\Sigma_k^{-1}(\phi(p_j)-\mu_k))}{\partial\phi(p_j)}\frac{\partial\phi(p_j)}{\partial h_1}. \tag{7.16}$$

$$\frac{\partial\mathcal{F}}{\partial h_1} = \sum_{j=1}^{S}\sum_{k=1}^{K}\omega_k\frac{1}{\sqrt{(2\pi)^2|\Sigma_k|}}\exp(-\frac{1}{2}(\phi(p_j)-\mu_k)^T\Sigma_k^{-1}(\phi(p_j)-\mu_k)) + \sum_{j=1}^{S}\sum_{k=1}^{K}\frac{\partial(-\frac{1}{2}(\phi(p_j)-\mu_k)^T\Sigma_k^{-1}(\phi(p_j)-\mu_k))}{\partial\phi(p_j)}\frac{\partial\phi(p_j)}{\partial h_1} \tag{7.17}$$

$$\frac{\partial \mathcal{F}}{\partial h_1} = \sum_{j=1}^{S}\sum_{k=1}^{K} \omega_k \mathcal{N}(\phi(p_j)|\mu_k, \Sigma_k) + \sum_{j=1}^{S}\sum_{k=1}^{K} (-(\phi(p_j) - \mu_k)^T \Sigma_k^{-1}) \cdot \begin{bmatrix} 1 \\ 0 \end{bmatrix}. \quad (7.18)$$

And:

$$\frac{\partial \mathcal{F}}{\partial s_1} = \sum_{j=1}^{S}\sum_{k=1}^{K} \omega_k \mathcal{N}(\phi(p_j)|\mu_k, \Sigma_k) + \sum_{j=1}^{S}\sum_{k=1}^{K} (-(\phi(p_j) - \mu_k)^T \Sigma_k^{-1}) \cdot \begin{bmatrix} p_{j,x} \\ 0 \end{bmatrix}. \quad (7.19)$$

$\frac{\partial \mathcal{F}}{\partial h_2}$ and $\frac{\partial \mathcal{F}}{\partial s_2}$ can be derived in a similar manner.

7.3 Experiments

In this section, we describe the experimental setup and the data used to evaluate the performance of our method. The first ten images of the eye tracking dataset of Judd et al. [65] are used as stimuli (Figure 7.4). The dataset has the advantage of containing the eye tracking data of 15 subjects for 1003 images collected from Flickr and LabelMe [96]. Hence, this data is used as template gaze patterns. The dataset contains landscape and portrait images with a 1024×768 resolution. The images contain multiple objects and they do not necessarily contain faces or objects centered in the middle of the image, which represents a realistic stimuli set.

For obtaining the ground truth for a new user, the Tobii T60XL gaze estimator [2] is used. It uses four infrared diodes mounted at the bottom of a 24 inch display with a resolution of 1920×1200 pixels. The reported accuracy of the gaze estimator is within 1°.

The design of the scene setup is to allow the subjects to look at the stimuli without hard constrains e.g. using a chin rest or sitting at a fixed distance from the stimuli. To collect the eye images, a web camera is mounted above the screen to record the face of the

Fig. 7.4: The 10 images used as stimuli in our experiments. The images show landscapes and street views where multiple objects are present in the scene.

subject. The eye image resolution is around 60×30. The coordinates and direction of the camera is unknown with regard to the gaze plane and can change for each new subject. Ten subjects were asked to sit where they wanted but within the allowed range of the Tobii system. The subject's distance from the display ranged from 55 to 75 cm. No chin rest is used in the experiments so the heads of the subjects may move during the experiment.

The subjects were asked to look at each image for three seconds followed by one second of showing a gray image. No specific task was asked and the subjects freely viewed the stimuli. The recording of each subject is stored and later analyzed to estimate the gaze points. We follow Lu et al. [72] approach to extract the images of the eyes. For each of the ten stimuli, the first corresponding web camera frame is taken as an input by the landmarker [143] to detect the eye corners. In [110], the eye corners are detected using the OMRON OKAO vision library. To detect the eye corners for the subsequent frames, we apply template matching using the eye corners of the first frame (for each stimulus) as templates. The eye images are then cropped from the corner and resized to 70×35. Histogram equalization is applied to alleviate the illumination changes. Regarding the gradient descent search in the matching methods, the number of epochs is set to 50. To prevent over downscaling the initial gaze points, we set a lower bound of scaling equal to 90% of the scale of the template gaze patterns.

7.3.1 Results on Artificially Distorted Data

Our assumption is that a collection of gaze patterns of individuals can be used to automatically infer the gaze calibration of a new user. In this section, we validate the assumption on artificially distorted data. More specifically, we use the eye tracking dataset in [65] and apply a distortion to the subject fixations. The distorted fixations are considered as a simulation of the initial (uncalibrated) gaze points. For each stimulus, we apply a random translation and scaling to the fixation set of each subject. Then, the methods in K-closest points and mixture model methods are used to transform the distorted gaze points to their correct locations. The first 30 images in the dataset are used in this experiment. For each image, we tested the subjects with 10 or more fixations. We discarded the images where the number of subjects (10 or more fixations) was less than 6 to ensure sufficient gaze patterns. Using the K-closest points, the mean accuracy across all images is 3.3°, while the accuracy is 3.5° using the mixture model fitting (the scene setup details can be found in [65]). The same procedure is applied on the ground truth gaze points obtained from our collected data. For this dataset, the K-closest points and mixture model fitting obtained accuracies of 2.6° and 2.4° respectively. The results show the validity of the proposed methods to bring the distorted (uncalibrated) gaze points closer to their correct locations for different sets of template

gaze patterns. Regarding the parameter setting, we set K in the K-closest points method to 5 and the number of Gaussian components to 7. We examined different values of K and components number and the performance difference was not significant.

7.3.2 Results on Real Data

The previous section shows how artificially distorted gaze points can be transformed to their correct locations with sufficient accuracy using the K-closest points. In this section, we use the aforementioned collected data to automatically calibrate the gaze estimator and find the gaze points from the videos acquired from the web camera. We apply the K-closest points and mixture model methods (Sections 7.2.1 and 7.2.1) to find the initial gaze points.

For the eye templates method, 25 eye templates were captured while a subject was fixating their eyes at 25 points on a 21.5 inch display. This process is followed once for a single (template) subject. Therefore, reconstructing an eye image of a new subject from the eye templates will not be optimal due to the changes in eye appearance between the template subject and the other subjects. However, we assume that it still gives a good representation of the topology of the gaze points. As in [72] we divide the eye image into a 5x3 grid and sum up the intensity of the pixel inside each grid cell. The resulting 15 values constitute the feature vector of the eye image.

Regarding the 2D manifold method, a template subject was asked to look at random points on the screen while his face was video recorded. The eye images are cropped and their feature vectors are computed as previously explained. Then, the feature vectors are projected on the first two principal components to constitute a 2D-manifold. The eye images of a new subject (while looking at a stimulus) are cropped, and then the feature vectors are extracted and projected on the same manifold to determine their relative locations. The distances between the initial gaze points are much larger than the actual corresponding gaze points. Yet, this will not affect the results as the initial gaze points will be scaled down, while finding the mapping, to match the initial gaze points with the template gaze patterns.

We select the gaze template patterns in two ways: First, we use the fixation points provided in the eye tracking dataset [65]. Second, the ground truth of our collected data (via the Tobii gaze estimator) is used. In the second case, for each subject, we consider the gaze points of the other subjects as template gaze patterns. The K-closest points and the mixture model fitting methods are applied to the initial gaze points. Table 7.1 shows the results.

The results show that the K-closest points method achieves lower error than using the mixture model method while 2D manifold outperforms eye templates for both template

Tab. 7.1: Accuracies over different methods and template gaze pattern sets. KCP denotes K-closest points method, GMM refers to Gaussian mixture model fitting. The best accuracy is yielded using 2D manifold and K-closest points.

	Template Gaze Patterns from [65]		Template Gaze Patterns from our Data	
	KCP	GMM	KCP	GMM
Eye Templates	4.6°	4.6°	4.7°	4.7°
2D Manifold	**4.2°**	4.3°	4.4°	4.5°

Tab. 7.2: Accuracies of the gaze estimation auto-calibrated using K-closest points and 2D manifold. The accuracies are shown per subject/stimulus.

	Stim. 1	**Stim. 2**	**Stim. 3**	**Stim. 4**	**Stim. 5**	**Stim. 6**	**Stim. 7**	**Stim. 8**	**Stim. 9**	**Stim. 10**	**Average**
Subject 1	4.8°	3.1°	2.1°	2.7°	6.3°	5.3°	4.9°	6.7°	6.4°	4.5°	4.7°
Subject 2	4.7°	2.1°	3.6°	2.1°	4.1°	3.8°	3.7°	5.9°	5.5°	4.8°	4.0°
Subject 3	4.4°	2.9°	1.8°	2.2°	3.6°	3.8°	3.4°	5.0°	5.3°	6.6°	3.9°
Subject 4	3.7°	2.3°	2.0°	2.8°	2.1°	2.4°	3.6°	6.2°	5.2°	6.7°	3.7°
Subject 5	5.5°	2.9°	2.8°	2.6°	3.4°	3.2°	3.6°	6.1°	4.6°	5.7°	4.0°
Subject 6	3.9°	3.0°	1.6°	3.9°	2.9°	3.5°	4.6°	5.1°	6.5°	5.3°	4.0°
Subject 7	4.2°	3.7°	3.1°	3.2°	3.5°	4.7°	5.2°	6.3°	7.7°	6.1°	4.8°
Subject 8	3.5°	3.1°	3.6°	5.0°	5.0°	5.3°	4.9°	5.4°	5.0°	4.0°	4.5°
Subject 9	3.8°	2.6°	2.7°	4.0°	4.4°	3.6°	5.5°	5.7°	5.5°	4.1°	4.2°
Subject 10	4.4°	3.3°	3.8°	4.2°	3.3°	4.7°	4.6°	6.0°	6.6°	4.8°	4.6°

gaze pattern sets. The best accuracy (4.2°) is obtained using the K-closest points and the 2D manifold. Table 7.2 details the results per subject/stimulus. Figure 7.5 shows the results for the first four images with subject 3.

Regarding the template gaze patterns, the accuracies are similar for both sets with a slight improvement when using the gaze patterns from [65] dataset. The template gaze pattern sets were collected in two different experiments on two different groups of subjects. This is interesting as it shows the general similarity of gaze patterns and hence suggests the validity of using them in auto-calibration regardless of the viewers. The gaze estimation accuracies vary for different subjects. The relatively lower accuracies for some subjects might be either due to errors in estimating the initial gaze points, i.e. because of eye appearance variations with the template subject eye templates which leads to incorrect initialization, or because of the gaze behavior of the subjects and its variation with the template gaze patterns. This point is further discussed in Section 7.3.4.

The stimuli set contains landscape and street view images, which makes the auto-calibration more challenging than images with clearly salient objects that humans usu-

ally focus on. Yet, the reported accuracy ($4.2°$) and the results in Figure 7.5 show the validity of our approach.

7.3.3 Gaze Estimation Error vs. Image Content

The primary observation behind our method is the similarities between the gaze patterns of different viewers when looking at the same stimulus [65]. These patterns may be influenced by the complexity or the scattering of the stimulus. This may, consequently, affect the gaze estimation error. In this section, we look further into the relationship between the image complexity and the gaze estimation error. More specifically, we first investigate the influence of the scene complexity to the heterogeneity between the gaze patterns of different users. Then, we examine the effect of this heterogeneity on the accuracy of the auto-calibrated gaze points estimation.

A number of approaches were proposed to model the image statistics (i.e. complexity) [47, 100, 120]. Here, we follow the approach of Geusebroek and Smeulders [47] and Scholte et al. [100] by fitting a Weibull distribution to the contrast value of the stimulus image. The Weibull distribution is defined as:

$$f(x) = C \exp(-|\frac{x-\mu}{\beta}|^{\gamma}). \tag{7.20}$$

Where C is a normalization constant and μ, β, and γ are the parameters of the Weibull distribution corresponding to the location, the scale, and the shape respectively. β and γ indicate some perceptual characteristics of the image such as regularity, coarseness, roughness, and contract [47, 115]. Geusebroek and Smeulders [47] and Scholte et al. [100] and found that images which correspond to low values of beta and gamma represent isolated objects in a plain background while their content changes gradually to contain multiple objects with higher values of beta and gamma. This suggests that image complexity can be characterized by the Weibull parameters. Scholte et al. [100] further evaluated the image complexity by compressing the images into JPEG format; the intuition here is that higher compression corresponds usually to simple scene and lower compression relates to complex scenes. They found that higher beta and gamma values correlate with higher JPEG file size (lower compression) and hence higher scene complexity. More importantly, the Weibull parameters, gamma and beta, highly correlate with neural responses in the early visual system [100].

The dissimilarity (i.e. heterogeneity) between two gaze patterns is measured by computing the KL divergence of their corresponding probability models. More specifically, each gaze pattern is modeled by a Gaussian mixture model. Next a symmetric KL diver-

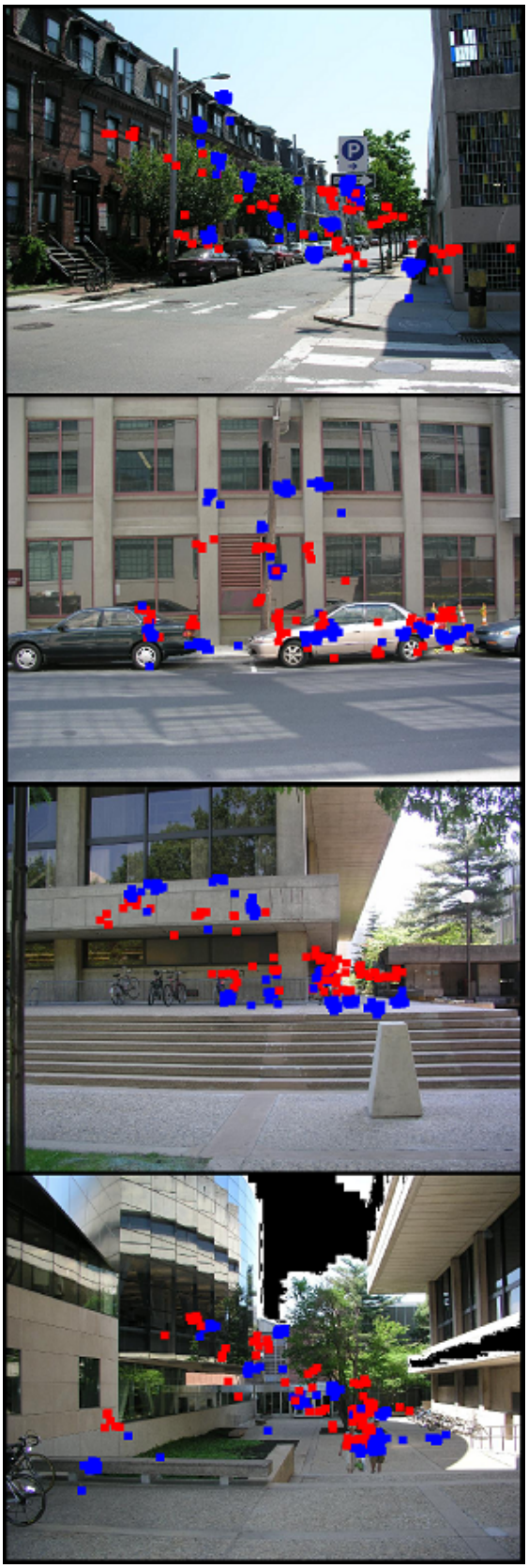


Fig. 7.5: Gaze estimation results for the first four images with subject 3. The red traces represent the estimated gaze points while the blue traces represent the ground truth obtained from the Tobii gaze estimator. The results are achieved using 2D-manifold and K-closest points.

gence is computed between the two probability distributions P and Q (corresponding to the two mixture models):

$$SD_{KL}(P\|Q) = D_{KL}(P\|Q) + D_{KL}(Q\|P), \tag{7.21}$$

where:

$$D_{KL}(P\|Q) = \int_{-\infty}^{\infty} p(x) \ln(\frac{p(x)}{q(x)})dx. \tag{7.22}$$

To compute the dissimilarity between more than two gaze patterns, the mean of the pair-wise similarities is calculated (we call it hereafter gaze patterns heterogeneity). To evaluate the relationship between the image complexity and the gaze patterns heterogeneity, we compute the correlation between the heterogeneity values and the corresponding beta and gamma parameters. The results indicate no significant correlation (r <0.1). This might be since the subjects were assigned a memory test while viewing the stimuli [65]. In this case, some regions might have been looked at to be memorized (for the test) rather than being salient or drawing attention. This reduces the impact the image complexity has on the gaze behavior.

Next, we compute the correlation between the gaze estimation error and the gaze pattern heterogeneity. Since we do not have the estimation error for all 1003 stimuli, we artificially distort the ground truth by random transformations (translation and scaling) and applying the mapping algorithm to transform the gaze points back to their correct positions (similar to the experiment in Section 7.3.1). The gaze estimation error, here, is associated with only the gaze mapping error (i.e. feature extraction error is not applied) which makes the correlation more indicative. With $\alpha = 0.01$, the correlation ($r = 0.407$) is significant which indicates an effect of the gaze pattern heterogeneity and the gaze estimation error.

The results of the previous experiments suggest an effect of the similarities between gaze patterns of viewers on the auto-calibration error. However, no correlation is found between the gaze patterns heterogeneity and the image complexity (characterized by Weibull parameters as in [47, 100]).

7.3.4 Initial Gaze Points Error vs. Auto-calibration Error

The previous experiments show how our method achieves, using visual features, an error of 4.2° without any kind of active calibration. Early steps of cropping the eye regions and extracting the visual features are likely to introduce some noise which propagates

to later steps and, consequently, contributes to the final error. In this experiment, we aim to isolate the gaze estimation error produced by the feature extraction procedure from the error induced by our auto-calibration method. To this end, we use the relative movements of eye centers provided by Tobii's infrared diodes as initial gaze points for the new users. Note that these are not the ground truth gaze points provided by the actively-calibrated Tobii system but just the changes of eye center positions measured by the infrared diodes. Since the infrared diodes produce more accurate and stable measurements than an RGB webcamera (especially when the head moves slightly during recording), we assume that such measurements alleviate the influence of initial gaze points estimation error. Please note that this is different from the experiment in Section 7.3.1, where the ground truth is distorted and realigned by the auto-calibration method. In this experiment, the measurements are more robust than the ones obtained by the RGB webcamera. However, they are still prone to some form of noise. Using the K-closest points method, the error reaches 3.1°. The results show that part of the gaze estimation error is attributed to noise in the feature extraction step (and hence the initial gaze point estimation).

7.3.5 Uncalibrated Human Gaze Patterns

In Section 7.3.2, we show how the gaze points of a new user in a uncalibrated setup can be inferred by using the information of gaze patterns of other users. The assumption in that experiment is that gaze patterns are available and calibrated, which may not be the case in some setups. In this experiment, we aim to relax this condition by using only the uncalibrated gaze patterns of other users in an incremental way. More specifically, the method starts with saliency information [59] as a template gaze pattern. For every new user, the gaze points are estimated and added to the template gaze patterns. The assumption here is that even if the gaze points are "estimated" (i.e. not accurate), they still provide some cues for other uncalibrated gaze points. After a certain number of users, since the template gaze patterns are modified, the gaze points of all users are re-estimated. Hence, the accuracy is improving gradually. In this experiments, the gaze points are re-estimated after 10 users. This process is repeated a number of times. The gaze estimation error over the iterations is plotted in Figure 7.6. Table 7.3 shows the errors after 10 iterations per user/stimulus. The results show that the error decreases gradually when adding or updating estimated gaze points. After 10 iterations, the error decreases from 4.7° to 4.3°. This suggests that our method provides comparable accuracy even when no calibrated gaze points of other users are available.

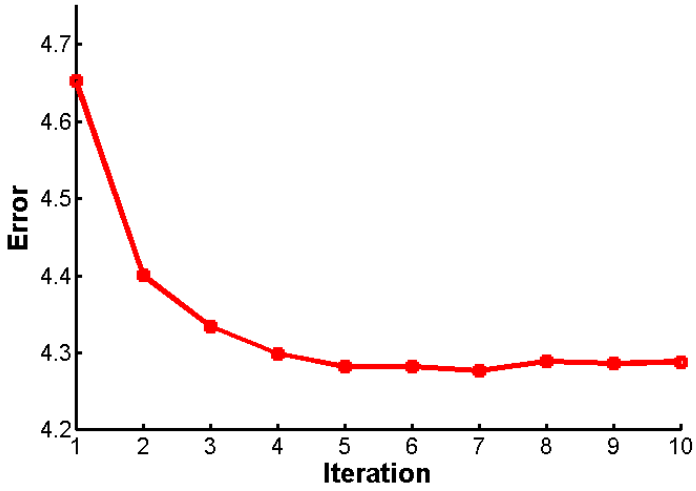


Fig. 7.6: Gaze estimation error after incrementally updating the template gaze patterns with estimated gaze points.

Tab. 7.3: Accuracies of the gaze estimation using *uncalibrated* gaze patterns after 10 iterations. The accuracies are shown per subject/stimulus.

	Stim. 1	**Stim. 2**	**Stim. 3**	**Stim. 4**	**Stim. 5**	**Stim. 6**	**Stim. 7**	**Stim. 8**	**Stim. 9**	**Stim. 10**	**Average**
Subject 1	4.5°	2.6°	2.8°	5.2°	5.2°	5.0°	3.5°	6.5°	5.0°	4.0°	4.4°
Subject 2	4.4°	2.9°	3.9°	4.2°	4.3°	3.8°	4.0°	5.3°	4.9°	4.5°	4.2°
Subject 3	3.7°	3.4°	2.0°	4.5°	3.6°	3.6°	3.4°	5.1°	4.8°	5.7°	4.0°
Subject 4	4.2°	2.9°	2.8°	3.3°	3.1°	2.1°	3.3°	5.3°	3.7°	6.5°	3.7°
Subject 5	5.0°	4.2°	3.5°	3.1°	4.5°	3.0°	3.9°	5.5°	4.5°	5.3°	4.3°
Subject 6	2.8°	3.1°	2.0°	6.0°	3.7°	3.3°	3.6°	4.9°	5.4°	5.3°	4.0°
Subject 7	4.5°	3.4°	3.5°	3.2°	4.1°	4.7°	5.2°	6.6°	6.6°	5.5°	4.7°
Subject 8	4.0°	3.3°	3.6°	5.4°	4.9°	5.3°	4.6°	4.2°	4.0°	3.2°	4.2°
Subject 9	4.4°	3.1°	3.3°	5.5°	5.4°	3.6°	4.0°	7.4°	5.4°	4.1°	4.6°
Subject 10	4.5°	3.6°	3.7°	4.8°	3.5°	4.8°	4.3°	5.9°	6.0°	4.4°	4.5°

7.3.6 Comparison to the State-of-the-art Methods

We compare our method with existing state-of-the-art auto-calibration approaches. The recent work of Chen and Ji [19] uses a single camera with multiple infrared lights to reconstruct the 3D eye model. They use the saliency to estimate the angle between the visual and optical axes. The authors reported less than 3° accuracy using five images and five subjects. Clearly, the comparison with this method is not feasible as the authors use different equipment to reconstruct an accurate 3D eye model.

Sugano et al. [110] adopt an appearance-based gaze estimator and use visual saliency for auto-calibration. The authors reported an accuracy of 3.5°. However, their experimental setup differs from ours in the following aspects: First, a chin rest is used in [110] to fixate the head during the experiment while the subjects in our experiment do not use any tool to fixate their heads. Second, the authors in [110] ask the subjects to look at a number of 30-second videos for training (5-20 videos), while in our method the subject needs to look at a single image for 3 seconds. Images contain less cues than videos in which moving objects attract the viewers attention. However, experimenting on still images is more natural and requiring motion in the scene limits the applicability of the gaze estimator. Finally, Sugano et al. analyze the performance variations with respect to different number of training videos. When training on 5 videos (each lasts 30 seconds), the average accuracy is about 5.2° (the exact accuracy is not reported as the results are plotted on a graph). While our method achieves an average accuracy of 4.2° by looking at a single image for 3 seconds.

7.4 Discussion

Our method provides sufficient accuracy to predict the areas of attention even with a flexible setup and a webcam. This is especially important for tasks where gaze estimation is required with no active participation from the user and using off-the-shelf hardware. In this work, we propose a flexible setup and use low-cost publicly available web cameras. There is a trend nowadays to use eye gaze estimation for electronic consumer relationship marketing which aims to employ information technology to understand and fulfill the needs of the consumers. These applications usually collect the data passively without active user participation. Our method is suitable for such applications. Tracing consumers attention when shopping in malls or when exploring advertisements on their laptops are examples of use.

We further make use of the "estimated" gaze points of the subsequent subjects to gradually auto-calibrate the gaze estimator by updating the template gaze patterns. This relaxes the condition of having the calibrated gaze patterns available. However, the presented method still has some limitations; the gaze behavior is, in general, expected to be different for different tasks (e.g. free viewing, searching, or memorizing). Consequently, the gaze patterns obtained from different tasks would differ and this would negatively affect the auto-calibration.

7.5 Conclusion

We presented a novel method to auto-calibrate gaze estimators in an uncalibrated setup. Based on the observation that humans produce similar gaze patterns when looking at a stimulus, we use the gaze patterns of individuals to estimate the gaze points for new viewers without active calibration.

The proposed method was tested in a flexible setup using a web camera without a chin rest. To estimate the gaze points, the viewer needs to look at an image for only 3 seconds without any explicit participation in the calibration. Evaluated on 10 subjects and 10 images showing landscapes and street views, the proposed method achieves an accuracy of $4.2°$. A number of experiments were conducted to give further insight into the method and its contribution in different cases. The gaze estimation error was reduced to $3.1°$ when infrared measurements were utilized. When using uncalibrated gaze patterns of other viewers, the estimation error was comparable to the one with calibrated gaze patterns. Finally, experiments show that the heterogeneity between the gaze patterns of the viewers has an impact on the auto-calibration error. To the best of our knowledge, this is the first work to use human gaze patterns in order to auto-calibrate gaze estimators.

8

Summary and Conclusion

8.1 Summary

Below, are the individual conclusions per chapter followed by the thesis conclusion.

8.1.1 Chapter 2

In this chapter, to alleviate the negative influence of noise in poor-quality images, a learning-based encoding method for age estimation is adopted. Soft encoding and orientation histogram of local gradients have been introduced. Experiments show that better or comparable performance are obtained using our extensions. With a discriminative codebook, our method outperforms the best performance reported on the poor-quality Gallagher dataset [103].

8.1.2 Chapter 3

In this chapter, an age-expression joint-learning approach is proposed to obtain expression-invariant age predictor. The relationship between the age and the expression is learnt by introducing a graphical model with a latent layer. This layer is designed to capture the changes in the face which induce the aging and the expression appearance.

We test our approach using two age-expression datasets (FACES and Lifespan). A reduction of the age estimation error is reported when the age is jointly learnt with the

expression in comparison to expression-independent age estimation. The age estimation error is reduced by 14.43% and 37.75% for FACES and Lifespan datasets respectively. The proposed approach has the advantage of not requiring prior-knowledge of the expressions. Experimental results show that, using our model, the acquired results are better that the best reported ones on both datasets.

8.1.3 Chapter 4

This study makes use of the facial dynamics to improve the appearance-based age estimation. To this end, landmark movements from eye lids, lip corners, and cheeks are employed. Another approach, for characterizing 3D facial dynamics, is to use the surface area features. We also introduce, in this chapter, a two-level classifier where the age range for each classifier is adaptively selected in the first level.

To show the contribution of the proposed dynamic features when accompanied with appearance features, four different appearance features are tested. The results prove the consistency of the dynamic features behavior across representations. Also, an appearance fusion baseline is introduced. We study gender effects systematically and conclude that the improvement due to gender-specific models is not significant. While this study mainly focus on dynamic features when smiling, experimental results show the positive contribution of the dynamic features with other expression (disgust) which further proves the effectiveness of the proposed features.

8.1.4 Chapter 5

The aim of this study is to automatically design region-specific, efficient and robust aging features. To this end, a convolutional network is assigned to each facial region. These are followed by a combination layer to integrate the information from the previous networks. This is different than other methods where the same filters are applied to the entire face.

We test the effectiveness, efficiency and generalizability of the proposed features. The proposed methods produce smaller estimation error than other state-of-the-art features using various classifiers and regressors. Efficiency-wise, the proposed features can be extracted fast and have a dimension of 176. Finally, a cross-dataset experiment to evaluate the generalizability where the model, based on our features, produces smaller errors than other models which are based on state-of-the-art features. Given the obtained results, the proposed features are shown to be suitable for real-time application especially for wearable devices with limited processing power.

8.1.5 Chapter 6

Motivated by the sensitivity of aging cues to image noise, in this chapter, the influence of image quality on aging feature performance is investigated. Changes in image quality affect the aging cues which primarily capture the skin texture and the wrinkles. We propose a scheme to isolate the contribution of aging features for different types and levels of noise. Three common digital noise types are investigated; JPEG-compression noise, quantization noise, and scaling noise. Four age estimation dataset are experimented on.

Finally, a framework is proposed to automatically, based on the image quality, apply the most suitable feature type. The image quality is automatically predicted. Results show better or comparable performance when automatically applying different features, based on image quality, in comparison to a single (best) feature type.

8.1.6 Chapter 7

This chapter addresses the problem of auto-calibration in gaze estimators where no active user involvement is required. Based on the observation that people have similar gaze patterns when looking at the same stimulus, a novel approach is proposed. The gaze patterns of individuals is utilized to estimate the gaze points for new users without active calibration.

The experimental setup is flexible and designed to simulate practical scenarios; 1) a web camera is mounted as a capturing device, 2) no chin rest, and 3) viewers need to look at the stimulus for only 3 seconds. The proposed method produces an error of 4.2°. To give further insight into the proposed approach, we test the method in different cases. First, to isolate the impact of eye measurements error on gaze estimation error, more accurate infrared diodes are utilized which reduces the error to 3.1°. Second, uncalibrated gaze patterns of other users, compared to calibrated gaze patterns, are utilized which produces comparable results. Finally, experiments show that the heterogeneity between the gaze patterns of the viewers has an impact on the auto-calibration error. To the best of our knowledge, this is the first work to use human gaze patterns in order to auto-calibrate gaze estimators.

8.2 Conclusions

In this thesis, we aim to find solutions to some of the problems in automatic face analysis under unconstrained conditions. Since computer scientists first began research on

automatic face analysis, many improvements and challenges have been uncovered. Solutions for automatic face analysis have matured in the last few years. Translating these solutions to real-life practical scenarios faces some challenges. This work contributes toward tackling these challenges for automatic age estimation and eye gaze estimation. It helps in applying automatic face analysis for everyday life problems.

In Chapter 2, we answer the **first research question** and show how, in a learning-based encoding scheme, soft assignment of local features alleviates the negative influence of noise on age estimation. The **second research question** concerns the negative influence of facial expressions when predicting age. In Chapter 3, we propose a graphical model to jointly predict age and expression. It contains a hidden layer to capture the relevant facial changes and, hence, alleviates the negative influence of expression when estimating age. In Chapter 4, we answer the **third research question** which concerns the use of facial dynamics to improve age prediction. To this end, we employ facial dynamic features such as landmark movements and surface area features. Results show that dynamic features, when combined with appearance features, further enhance the age estimation. The **fourth research question** raises the idea of automatically designing region-specific, efficient, and robust aging features. To this end, in Chapter 5, we suggest patch-based convolutional networks. The proposed features are fast to extract and test, yet, they produce a smaller error compared to other state-of-the-art features across various classifiers and regressors. The **fifth research question** concerns the influence of different types and levels of digital noise on the performance of aging features. An empirical study about the influence of different levels of JPEG compression, quantization, and scaling noise on aging features is discussed in Chapter 6. Finally, the **sixth research question** investigates automatic calibration of age estimation systems without active user involvement. Answering this question, we propose in Chapter 7 to use the gaze patterns of other people to auto-calibrate for a new viewer. Using an off-the-shelf web camera, no chin rest, and only 3-second viewing time, our approach produces an error of 4.2° which is sufficient to trace the attention of users.

Bibliography

[1] The fg-net aging database, http://www.fgnet.rsunit.com.

[2] Tobii technology: http://www.tobii.com/. Technical report.

[3] T. Ahonen, A. Hadid, and M. Pietikäinen. Face recognition with local binary patterns. In *Computer vision-eccv 2004*, pages 469–481. Springer, 2004.

[4] A. M. Albert, K. Ricanek Jr, and E. Patterson. A review of the literature on the aging adult skull and face: Implications for forensic science research and applications. *Forensic Science International*, 172(1):1–9, 2007.

[5] T. R. Alley. *Social and applied aspects of perceiving faces*. Lawrence Erlbaum Associates, Hillsdale, NJ, 1988.

[6] F. Alnajar, T. Gevers, R. Valenti, and S. Ghebreab. Calibration-free gaze estimation using human gaze patterns. In *Computer Vision (ICCV), 2013 IEEE International Conference on*, pages 137–144. IEEE, 2013.

[7] 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*, 2012.

[8] G. Anders. Pilot's attention allocation during approach and landing- eye-and head-tracking research in an a 330 full flight simulator. *Focusing Attention on Aviation Safety*, 2001.

[9] D. S. Berry and L. Z. McArthur. Perceiving character in faces: the impact of age-related craniofacial changes on social perception. *Psychological bulletin*, 100(1):3, 1986.

[10] D. J. Beymer. Face recognition under varying pose. In *Computer Vision and Pattern Recognition, 1994. Proceedings CVPR'94., 1994 IEEE Computer Society Conference on*, pages 756–761. IEEE, 1994.

[11] W. Bledsoe. The model method in facial recognition. Technical report, Technical Report PRI 15, Panoramic Research, Inc., Palo Alto, California., 1966.

[12] S. E. Brodie. Aging and disorders of the eye. In H. Fillit, K. Rockwood, and K. Woodhouse, editors, *Brocklehurst's Textbook of Geriatric Medicine and Gerontology*, chapter 96. Saunders Elsevier, Philadelphia, PA, 7 edition, 2010.

[13] G. Butterworth. The ontogeny and phylogeny of joint visual attention. 1991.

[14] G. Butterworth and N. Jarrett. What minds have in common is space: Spatial mechanisms serving joint visual attention in infancy. *British journal of developmental psychology*, 9(1):55–72, 1991.

[15] Z. Cao, Q. Yin, X. Tang, and J. Sun. Face recognition with learning-based descriptor. In *Computer Vision and Pattern Recognition (CVPR), 2010 IEEE Conference on*, pages 2707–2714. IEEE, 2010.

[16] J. M. Carré and C. M. McCormick. In your face: facial metrics predict aggressive behaviour in the laboratory and in varsity and professional hockey players. *Proceedings of the Royal Society of London B: Biological Sciences*, 275(1651):2651–2656, 2008.

[17] K.-Y. Chang, C.-S. Chen, and Y.-P. Hung. Ordinal hyperplanes ranker with cost sensitivities for age estimation. In *Computer Vision and Pattern Recognition (CVPR), 2011 IEEE Conference on*, pages 585–592. IEEE, 2011.

[18] W.-L. Chao, J.-Z. Liu, and J.-J. Ding. Facial age estimation based on label-sensitive learning and age-oriented regression. *Pattern Recognition*, 46(3):628–641, 2013.

[19] J. Chen and Q. Ji. Probabilistic gaze estimation without active personal calibration. In *Computer Vision and Pattern Recognition (CVPR), 2011 IEEE Conference on*, pages 609–616. IEEE, 2011.

[20] Y. Chen and C. Hsu. Subspace learning for facial age estimation via pairwise age ranking. *IEEE Trans. on Information Forensics and Security*, 8(12):2164–2176, 2013.

[21] 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):1262–1281, 2011.

[22] T. F. Cootes, G. J. Edwards, and C. J. Taylor. Active appearance models. *IEEE Transactions on pattern analysis and machine intelligence*, 23(6):681–685, 2001.

[23] N. Dalal and B. Triggs. Histograms of oriented gradients for human detection. In *Computer Vision and Pattern Recognition, 2005. CVPR 2005. IEEE Computer Society Conference on*, volume 1, pages 886–893. IEEE, 2005.

[24] J. G. Daugman. Uncertainty relation for resolution in space, spatial frequency, and orientation optimized by two-dimensional visual cortical filters. *JOSA A*, 2(7):1160–1169, 1985.

[25] H. Dibeklioglu, A. A. Salah, and T. Gevers. Recognition of genuine smiles. *Multimedia, IEEE Transactions on*, 17(3):279–294, 2015.

[26] H. Dibeklioğlu, T. Gevers, A. A. Salah, and R. Valenti. A smile can reveal your age: Enabling facial dynamics in age estimation. In *ACM International Conference on Multimedia*, pages 209–218, 2012.

[27] H. Dibeklioğlu, A. A. Salah, and T. Gevers. Are you really smiling at me? Spontaneous versus posed enjoyment smiles. In *ECCV*, pages 526–539, 2012.

[28] H. Dibeklioğlu, A. A. Salah, and T. Gevers. A statistical method for 2-d facial landmarking. *IEEE Trans. on Image Processing*, 21(2):844–858, 2012.

[29] H. Dibeklioğlu, A. A. Salah, and T. Gevers. Like father, like son: Facial expression dynamics for kinship verification. In *ICCV*, pages 1497–1504, 2013.

[30] N. C. Ebner, M. Riediger, and U. Lindenberger. Faces: database of facial expressions in young, middle-aged, and older women and men: Development and validation. *Behavior Research Methods*, 42(1):351–362, 2010.

[31] P. Ekman. *Telling lies: Cues to deceit in the marketplace, politics, and marriage*. New York: WW. Norton & Company, 1992.

[32] P. Ekman and W. V. Friesen. Constants across cultures in the face and emotion. *Journal of personality and social psychology*, 17(2):124, 1971.

[33] P. Ekman and W. V. Friesen. Facial action coding system: A technique for the

measurement of facial movement. *Consulting Psychologists Press*, 1978.

[34] P. Ekman and W. V. Friesen. Felt, false, and miserable smiles. *Journal of Nonverbal Behavior*, 6(4):238–252, 1982.

[35] D. H. Enlow and R. E. Moyers. *Handbook of facial growth*. WB Saunders Company, 1982.

[36] Y. Freund, S. Dasgupta, M. Kabra, and N. Verma. Learning the structure of manifolds using random projections. In *Advances in Neural Information Processing Systems*, pages 473–480, 2007.

[37] S. Fu, H. He, and Z.-G. Hou. Learning race from face: A survey. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 36(12):2483–2509, 2014.

[38] Y. Fu, G. Guo, and T. S. Huang. Age synthesis and estimation via faces: A survey. *IEEE Trans. on PAMI*, 32(11):1955–1976, 2010.

[39] Y. Fu and T. S. Huang. Human age estimation with regression on discriminative aging manifold. *Multimedia, IEEE Transactions on*, 10(4):578–584, 2008.

[40] A. C. Gallagher and T. Chen. Understanding images of groups of people. In *CVPR*, pages 256–263, 2009.

[41] F. Gao and H. Ai. Face age classification on consumer images with gabor feature and fuzzy lda method. In *Advances in biometrics*, pages 132–141. Springer, 2009.

[42] W. Gao and H. Ai. Face gender classification on consumer images in a multiethnic environment. In *Advances in Biometrics*, pages 169–178. Springer, 2009.

[43] X. Geng, K. Smith-Miles, and Z. Zhou. Facial age estimation by nonlinear aging pattern subspace. In *ACM International Conference on Multimedia*, pages 721–724, 2008.

[44] X. Geng, C. Yin, and Z.-H. Zhou. Facial age estimation by learning from label distributions. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 35(10):2401–2412, 2013.

[45] X. Geng, Z.-H. Zhou, and K. Smith-Miles. Automatic age estimation based on facial aging patterns. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 29(12):2234–2240, 2007.

[46] 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*, pages 307–316, 2006.

[47] J.-M. Geusebroek and A. W. Smeulders. A six-stimulus theory for stochastic texture. *International Journal of Computer Vision*, 62(1-2):7–16, 2005.

[48] T. Gritti, C. Shan, V. Jeanne, and R. Braspenning. Local features based facial expression recognition with face registration errors. In *Automatic Face & Gesture Recognition, 2008. FG'08. 8th IEEE International Conference on*, pages 1–8. IEEE, 2008.

[49] E. D. Guestrin and M. Eizenman. General theory of remote gaze estimation using the pupil center and corneal reflections. *Biomedical Engineering, IEEE Transactions on*, 53(6):1124–1133, 2006.

[50] E. D. Guestrin and M. Eizenman. Remote point-of-gaze estimation requiring

a single-point calibration for applications with infants. In *Proceedings of the 2008 symposium on Eye tracking research & applications*, pages 267–274. ACM, 2008.

[51] G. Guo, Y. Fu, C. R. Dyer, and T. S. Huang. Image-based human age estimation by manifold learning and locally adjusted robust regression. *IEEE Trans. on Image Processing*, 17(7):1178–1188, 2008.

[52] G. Guo, R. Guo, and X. Li. Facial expression recognition influenced by human aging. *Affective Computing*, 4(3):291–298, 2013.

[53] G. Guo, G. Mu, Y. Fu, and T. S. Huang. Human age estimation using bio-inspired features. In *CVPR*, pages 112–119, 2009.

[54] G. Guo and X. Wang. A study on human age estimation under facial expression changes. In *CVPR*, pages 2547–2553, 2012.

[55] A. Hadid. Analyzing facial behavioral features from videos. In A. A. Salah and B. Lepri, editors, *Int. Workshop on Human Behavior Understanding*, pages 52–61, 2011.

[56] H. Han, C. Otto, and A. K. Jain. Age estimation from face images: Human vs. machine performance. In *BTAS*, 2013.

[57] D. W. Hansen, J. P. Hansen, M. Nielsen, A. S. Johansen, and M. B. Stegmann. Eye typing using markov and active appearance models. In *Applications of Computer Vision, 2002.(WACV 2002). Proceedings. Sixth IEEE Workshop on*, pages 132–136. IEEE, 2002.

[58] D. W. Hansen and Q. Ji. In the eye of the beholder: A survey of models for eyes and gaze. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 32(3):478–500, 2010.

[59] J. Harel, C. Koch, and P. Perona. Graph-based visual saliency. In *Advances in neural information processing systems*, pages 545–552, 2006.

[60] M. P. Haselhuhn and E. M. Wong. Bad to the bone: facial structure predicts unethical behaviour. *Proceedings of the Royal Society of London B: Biological Sciences*, 279(1728):571–576, 2012.

[61] R. Highfield, R. Wiseman, and R. Jenkins. How your looks betray your personality. *New Scientist, Feb*, 2009.

[62] G. B. Huang, M. Ramesh, T. Berg, and E. Learned-Miller. Labeled faces in the wild: A database for studying face recognition in unconstrained environments. Technical report, Technical Report 07-49, University of Massachusetts, Amherst, 2007.

[63] T. E. Hutchinson, K. P. White Jr, W. N. Martin, K. C. Reichert, L. Frey, et al. Human-computer interaction using eye-gaze input. *Systems, Man and Cybernetics, IEEE Transactions on*, 19(6):1527–1534, 1989.

[64] Q. Ji and X. Yang. Real-time eye, gaze, and face pose tracking for monitoring driver vigilance. *Real-Time Imaging*, 8(5):357–377, 2002.

[65] T. Judd, K. Ehinger, F. Durand, and A. Torralba. Learning to predict where humans look. In *Computer Vision, 2009 IEEE 12th international conference on*,

pages 2106–2113. IEEE, 2009.

[66] T. Kanade. Picture processing system by computer complex and recognition of human faces. *Doctoral dissertation, Kyoto University*, 3952:83–97, 1973.

[67] Y. H. Kwon and N. da Vitoria Lobo. Age classification from facial images. *Computer Vision and Image Understanding*, 74(1):1–21, 1999.

[68] A. Lanitis, C. J. Taylor, and T. F. Cootes. Toward automatic simulation of aging effects on face images. *IEEE Trans. on PAMI*, 24(4):442–455, 2002.

[69] G. Lewis, C. Lefevre, and T. Bates. Facial width-to-height ratio predicts achievement drive in us presidents. *Personality and Individual Differences*, 52(7):855–857, 2012.

[70] J. Liu, Y. Ma, L. Duan, F. Wang, and Y. Liu. Hybrid constraint SVR for facial age estimation. *Signal Processing*, 94:576–582, 2014.

[71] D. G. Lowe. Distinctive image features from scale-invariant keypoints. *International journal of computer vision*, 60(2):91–110, 2004.

[72] F. Lu, Y. Sugano, T. Okabe, and Y. Sato. Inferring human gaze from appearance via adaptive linear regression. In *Computer Vision (ICCV), 2011 IEEE International Conference on*, pages 153–160. IEEE, 2011.

[73] M. Lucassen, T. Gevers, and H. Dibeklioğlu. The effect of smile and illumination color on age estimation from faces. *Perception*, 41, ECVP Abstract Supplement:87, 2012.

[74] P. Luo, X. Wang, and X. Tang. Hierarchical face parsing via deep learning. In *Computer Vision and Pattern Recognition (CVPR), 2012 IEEE Conference on*, pages 2480–2487. IEEE, 2012.

[75] M. J. Lyons, J. Budynek, and S. Akamatsu. Automatic classification of single facial images. *IEEE Transactions on Pattern Analysis & Machine Intelligence*, (12):1357–1362, 1999.

[76] P. Majaranta and K.-J. Räihä. Twenty years of eye typing: systems and design issues. In *Proceedings of the 2002 symposium on Eye tracking research & applications*, pages 15–22. ACM, 2002.

[77] K. L. Minaker. Common clinical sequelae of aging. In L. Goldman and A. I. Schafer, editors, *Goldman's Cecil Medicine*, chapter 24. Saunders Elsevier, Philadelphia, PA, 24 edition, 2011.

[78] M. Minear and D. C. Park. A lifespan database of adult facial stimuli. *Behavior Research Methods, Instruments, and Computers*, 36(4):630–633, 2004.

[79] B. Moghaddam and M.-H. Yang. Learning gender with support faces. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 24(5):707–711, 2002.

[80] J. M. Mooij. Libdai: A free and open source c++ library for discrete approximate inference in graphical models. *Journal of Machine Learning Research*, pages 2169–2173, 2010.

[81] B. Ni, Z. Song, and S. Yan. Web image mining towards universal age estimator. In *Proceedings of the 17th ACM international conference on Multimedia*, pages 85–94. ACM, 2009.

[82] T. Ojala, M. Pietikäinen, and D. Harwood. A comparative study of texture measures with classification based on featured distributions. *Pattern Recognition*, 29(1):51–59, 1996.

[83] T. Ojala, M. Pietikäinen, and T. Maenpaa. Multiresolution gray-scale and rotation invariant texture classification with local binary patterns. *IEEE Trans. on PAMI*, 24(7):971–987, 2002.

[84] M. Ortega, L. Brodo, M. Bicego, and M. Tistarelli. On the quantitative estimation of short-term aging in human faces. In *ICIAP*, pages 575–584, 2009.

[85] M. Osadchy, Y. L. Cun, and M. L. Miller. Synergistic face detection and pose estimation with energy-based models. *The Journal of Machine Learning Research*, 8:1197–1215, 2007.

[86] A. J. O'Toole, T. Vetter, H. Volz, and E. Salter. Three-dimensional caricatures of human heads: Distinctiveness and the perception of age. *Perception*, 26(6):719–732, 1997.

[87] H. Peng, F. Long, and C. Ding. Feature selection based on mutual information criteria of max-dependency, max-relevance, and min-redundancy. *IEEE Trans. on PAMI*, 27(8):1226–1238, 2005.

[88] T. Pfister, X. Li, G. Zhao, and M. Pietikäinen. Recognising spontaneous facial micro-expressions. In *Computer Vision (ICCV), 2011 IEEE International Conference on*, pages 1449–1456. IEEE, 2011.

[89] N. Ramanathan, R. Chellappa, and S. Biswas. Computational methods for modeling facial aging: A survey. *Journal of Visual Languages & Computing*, 20(3):131–144, 2009.

[90] K. Ramesha, K. B. Raja, K. R. Venugopal, and L. M. Patnaik. Feature extraction based face recognition, gender and age classification. *International Journal on Computer Science and Engineering*, 1(1S):14–23, 2010.

[91] S. Ren, X. Cao, Y. Wei, and J. Sun. Face alignment at 3000 fps via regressing local binary features. In *CVPR*, 2014.

[92] M. G. Rhodes. Age estimation of faces: A review. *Applied Cognitive Psychology*, 23(1):1–12, 2009.

[93] K. Ricanek and T. Tesafaye. Morph: A longitudinal image database of normal adult age-progression. In *Automatic Face and Gesture Recognition, 2006. FGR 2006. 7th International Conference on*, pages 341–345. IEEE, 2006.

[94] S. T. Roweis and L. K. Saul. Nonlinear dimensionality reduction by locally linear embedding. *Science*, 290(5500):2323–2326, 2000.

[95] L. R. Rubin. The anatomy of a smile: its importance in the treatment of facial paralysis. *Plastic and reconstructive surgery*, 53(4):384–387, 1974.

[96] B. C. Russell, A. Torralba, K. P. Murphy, and W. T. Freeman. Labelme: a database and web-based tool for image annotation. *International journal of computer vision*, 77(1-3):157–173, 2008.

[97] S. J. Salasche. Anatomy. In T. E. Rohrer, J. L. Cook, T. H. Nguyen, and J. R. Mellette Jr., editors, *Flaps and Grafts in Dermatologic Surgery*, chapter 1. Saunders

Elsevier, Philadelphia, PA, 1 edition, 2007.

[98] R. Sanders. Torsional elasticity of human skin in vivo. *Pflügers Archiv European Journal of Physiology*, 342(3):255–260, 1973.

[99] K. L. Schmidt, J. F. Cohn, and Y. Tian. Signal characteristics of spontaneous facial expressions: Automatic movement in solitary and social smiles. *Biological Psychology*, 65(1):49–66, 2003.

[100] H. S. Scholte, S. Ghebreab, L. Waldorp, A. W. Smeulders, and V. A. Lamme. Brain responses strongly correlate with weibull image statistics when processing natural images. *Journal of Vision*, 9(4):29, 2009.

[101] Science Center NEMO. http://www.e-nemo.nl/.

[102] T. Serre, L. Wolf, S. Bileschi, M. Riesenhuber, and T. Poggio. Robust object recognition with cortex-like mechanisms. *IEEE Trans. on PAMI*, 29(3):411–426, 2007.

[103] C. Shan. Learning local features for age estimation on real-life faces. In *Proceedings of the 1st ACM international workshop on Multimodal pervasive video analysis*, pages 23–28. ACM, 2010.

[104] C. Shan, S. Gong, and P. W. McOwan. Facial expression recognition based on local binary patterns: A comprehensive study. *Image and Vision Computing*, 27(6):803–816, 2009.

[105] K. Smith, S. O. Ba, J.-M. Odobez, and D. Gatica-Perez. Tracking the visual focus of attention for a varying number of wandering people. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 30(7):1212–1229, 2008.

[106] I. D. Stephen, V. Coetzee, M. Law Smith, D. I. Perrett, et al. Skin blood perfusion and oxygenation colour affect perceived human health. *PloS one*, 4(4):e5083–e5083, 2009.

[107] I. D. Stephen, M. J. L. Smith, M. R. Stirrat, and D. I. Perrett. Facial skin coloration affects perceived health of human faces. *International journal of primatology*, 30(6):845–857, 2009.

[108] M. Stirrat and D. I. Perrett. Face structure predicts cooperation men with wider faces are more generous to their in-group when out-group competition is salient. *Psychological science*, page 0956797611435133, 2012.

[109] Y. Sugano, Y. Matsushita, and Y. Sato. Calibration-free gaze sensing using saliency maps. In *Computer Vision and Pattern Recognition (CVPR), 2010 IEEE Conference on*, pages 2667–2674. IEEE, 2010.

[110] Y. Sugano, Y. Matsushita, and Y. Sato. Appearance-based gaze estimation using visual saliency. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 35(2):329–341, 2013.

[111] Y. Sun, X. Wang, and X. Tang. Deep convolutional network cascade for facial point detection. In *Computer Vision and Pattern Recognition (CVPR), 2013 IEEE Conference on*, pages 3476–3483. IEEE, 2013.

[112] J. Suo, T. Wu, S. Zhu, S. Shan, X. Chen, and W. Gao. Design sparse features for age estimation using hierarchical face model. In *Automatic Face & Gesture*

Recognition, 2008. FG'08. 8th IEEE International Conference on, pages 1–6. IEEE, 2008.

[113] J. Suo, S. C. Zhu, S. Shan, and X. Chen. A compositional and dynamic model for face aging. *IEEE Trans. on PAMI*, 32(3):385–401, 2010.

[114] H. Takimoto, Y. Mitsukura, M. Fukumi, and N. Akamatsu. Robust gender and age estimation under varying facial pose. *Electronics and Communications in Japan*, 91(7):32–40, 2008.

[115] H. Tamura, S. Mori, and T. Yamawaki. Textural features corresponding to visual perception. *Systems, Man and Cybernetics, IEEE Transactions on*, 8(6):460–473, 1978.

[116] K.-H. Tan, D. J. Kriegman, and N. Ahuja. Appearance-based eye gaze estimation. In *Applications of Computer Vision, 2002.(WACV 2002). Proceedings. Sixth IEEE Workshop on*, pages 191–195. IEEE, 2002.

[117] X. Tan and B. Triggs. Enhanced local texture feature sets for face recognition under difficult lighting conditions. *Image Processing, IEEE Transactions on*, 19(6):1635–1650, 2010.

[118] H. Tao and T. Huang. Explanation-based facial motion tracking using a piecewise Bézier volume deformation model. In *CVPR*, volume 1, pages 611–617, 1999.

[119] B. Tiddeman, M. Burt, and D. I. Perrett. Prototyping and transforming facial textures for perception research. *IEEE Computer Graphics and Applications*, 21(5):42–50, 2001.

[120] A. Torralba and A. Oliva. Statistics of natural image categories. *Network: computation in neural systems*, 14(3):391–412, 2003.

[121] I. Tsochantaridis, T. Joachims, T. Hofmann, and Y. Altun. Large margin methods for structured and interdependent output variables. *Journal of Machine Learning Research*, pages 1453–1484, 2005.

[122] J.-D. Txia and C.-L. Huang. Age estimation using aam and local facial features. In *Intelligent Information Hiding and Multimedia Signal Processing, 2009. IIH-MSP'09. Fifth International Conference on*, pages 885–888. IEEE, 2009.

[123] K. Ueki, T. Hayashida, and T. Kobayashi. Subspace-based age-group classification using facial images under various lighting conditions. In *AFGR*, 2006.

[124] R. Valenti and T. Gevers. Accurate eye center location through invariant isocentric patterns. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 34(9):1785–1798, 2012.

[125] M. F. Valstar, M. Pantic, Z. Ambadar, and J. F. Cohn. Spontaneous vs. posed facial behavior: automatic analysis of brow actions. In *Proceedings of the 8th international conference on Multimodal interfaces*, pages 162–170. ACM, 2006.

[126] J. C. Van Gemert, C. J. Veenman, A. W. Smeulders, and J.-M. Geusebroek. Visual word ambiguity. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 32(7):1271–1283, 2010.

[127] P. F. Velleman. Definition and comparison of robust nonlinear data smoothing algorithms. *Journal of the American Statistical Association*, pages 609–615, 1980.

[128] A. Villanueva, R. Cabeza, and S. Porta. Eye tracking: Pupil orientation geometrical modeling. *Image and Vision Computing*, 24(7):663–679, 2006.

[129] Y. Wu, N. Thalmann, and D. Thalmann. A dynamic wrinkle model in facial animation and skin aging. *Journal of Visualization and Computer Animation*, 6:195–205, 1995.

[130] S. Yan, M. Liu, and T. S. Huang. Extracting age information from local spatially flexible patches. In *Acoustics, Speech and Signal Processing, 2008. ICASSP 2008. IEEE International Conference on*, pages 737–740. IEEE, 2008.

[131] S. Yan, H. Wang, Y. Fu, J. Yan, X. Tang, and T. S. Huang. Synchronized submanifold embedding for person-independent pose estimation and beyond. *Image Processing, IEEE Transactions on*, 18(1):202–210, 2009.

[132] S. Yan, X. Zhou, M. Liu, M. Hasegawa-Johnson, and T. S. Huang. Regression from patch-kernel. In *Computer Vision and Pattern Recognition, 2008. CVPR 2008. IEEE Conference on*, pages 1–8. IEEE, 2008.

[133] Z. Yang and H. Ai. Demographic classification with local binary patterns. In *Advances in Biometrics*, pages 464–473. Springer, 2007.

[134] J. Ylioinas, A. Hadid, X. Hong, and M. Pietikäinen. Age estimation using local binary pattern kernel density estimate. In *ICIAP*, pages 141–150. 2013.

[135] C.-N. J. Yu and T. Joachims. Learning structural svms with latent variables. *International Conference on Machine Learning*, pages 1169–1176, 2009.

[136] A. L. Yuille, P. W. Hallinan, and D. S. Cohen. Feature extraction from faces using deformable templates. *International journal of computer vision*, 8(2):99–111, 1992.

[137] A. L. Yuille and A. Rangarajan. The concave-convex procedure (cccp). *Neural Computation*, pages 915–936, 2003.

[138] C. Zhan, W. Li, and P. Ogunbona. Age estimation based on extended non-negative matrix factorization. In *IEEE Int. Workshop on Multimedia Signal Processing*, 2011.

[139] C. Zhang and G. Guo. Age estimation with expression changes using multiple aging subspaces. In *BTAS*, 2013.

[140] J. Zhang, S. Shan, M. Kan, and X. Chen. Coarse-to-fine auto-encoder networks (cfan) for real-time face alignment. In *Computer Vision–ECCV 2014*, pages 1–16. Springer, 2014.

[141] Y. Zhang and D.-Y. Yeung. Multi-task warped gaussian process for personalized age estimation. In *Computer Vision and Pattern Recognition (CVPR), 2010 IEEE Conference on*, pages 2622–2629. IEEE, 2010.

[142] G. Zhao and M. Pietikäinen. Dynamic texture recognition using local binary patterns with an application to facial expressions. *IEEE Trans. on PAMI*, 29(6):915–928, 2007.

[143] X. Zhu and D. Ramanan. Face detection, pose estimation, and landmark localization in the wild. In *Computer Vision and Pattern Recognition (CVPR), 2012 IEEE Conference on*, pages 2879–2886. IEEE, 2012.

Samenvatting

Dit proefschrift richt zich op twee taken in gezichts analyse: het automatisch schatten van leeftijd en het automatisch schatten van kijkrichting. Anders dan in eerdere aanpakken is het vooral gericht op het aanbieden van (praktische) oplossing voor deze taken in varierende omstandigheden en opstellingen. Beneden zijn de conclusies gegeven per hoofdstuk.

Hoofdstuk 2:

In dit hoofdstuk, om de negatieve invloed van ruis in plaatjes van lage kwaliteit te verlichten, is een codering voor het voorspellen van leeftijd op basis van leren aangenomen. Continue codering en orientatie histogrammen van lokale afgeleiden worden geintroduceerd. Experimenten laten zien dat verbeterde of vergelijkbare resultaten worden bereikt door middel van onze uitbreidingen. Met een discriminatieve representatie verbetert onze methode de beste resultaten die zijn gerapporteerd op de lage kwaliteit Gallagher dataset [56].

Hoofdstuk 3:

In dit hoofdstuk wordt een aanpak geintroduceerd dat tegelijkertijd leeftijd en gelaatsuitdrukking leert, om een leeftijdsvoorspeller te krijgen dat onafhankelijk is van gelaatsuitdrukking. De relatie tussen leeftijd en gelaatsuitdrukking wordt geleerd door een grafisch model te introduceren met een verborgen laag. Deze laag is ontworpen om de verschillen in het gezicht vast te leggen die kenmerkend zijn voor leeftijd en gelaatsuitdrukking.

We testen onze aanpak op twee datasets met leeftijd en gelaatsuitdrukking annotaties (FACES en Lifespan). Een verbetering in het schatten van de leeftijd wordt gerapporteerd wanneer de leeftijd tegelijkertijd wordt geleerd met de gelaatsuitdrukking, in

vergelijking met leeftijdsschatting dat afhankelijk is van de gelaatsuitdrukking. De fout in de leeftijdsschatting wordt gereduceerd met 14.43voor de FACES en Lifespan dataset respectievelijk. De voorgedragen aanpak heeft als voordeel dat het niet kennis van gelaatsuitdrukkingen van te voren nodig heeft. Experimenten laten zien dat onze resultaten beter zijn dan de best gerapporteerde resultaten op beide datasets.

Hoofdstuk 4:

Deze studie maakt gebruik van dynamische elementen in het gezicht om het schatten van de leeftijd op basis van het uiterlijk te verbeteren. Veelzeggende bewegingen van oogleden, mondhoeken, en wangen worden gebruikt. In dit hoofdstuk introduceren we ook een leeralgoritme met twee lagen, waarbij het bereik van de leeftijd voor elk leeralgoritme adaptief wordt geselecteerd in de eerste laag.

Om de bijdrage van de geintroduceerde dynamische functies te laten zien als toevoeging op de uiterlijk functies, worden vier verschillende uiterlijk functies getest. De resultaten bewijzen de consistentie van het gedrag van de dynamische functies over verschillende representaties. Ook wordt een fusie op basis verschillende uiterlijke kenmerken geintroduceerd. Wij bestuderen de effecten van het geslacht systematisch en concluderen dat de verbetering die komt door modellen die specifiek zijn aan geslacht niet significant zijn. Alhoewel deze studie voornamelijk gericht is op de dynamische functies wanneer er wordt gelachen, laten experimenten zien dat de dynamische functies een positieve toevoeging hebben op andere uitdrukkingen (zoals afkeer), wat de effectiviteit van de voorgestelde functies verder bewijst.

Hoofdstuk 5:

Het doel van deze studie is om automatisch functies te ontwerpen die specifiek voor een regio, efficient, en robuust zijn. Daarom wordt een convolutioneel netwerk toegekend aan elk gebied van het gezicht. Deze worden gevolgd door een combinatielaag om all informatie van de netwerken te integreren. Dit verschilt van andere methoden waarbij dezelfde filters worden toegepast op het gehele gezicht.

Wij evalueren de effectiviteit, efficientie, en mogelijkheid om te generaliseren op de voorgestelde functies. De voorgestelde methoden hebben een kleinere fout in de schatting dan andere functies van de bovenste plank die gebruik maken van meerdere leeralgoritmen. De voorgestelde functies kunnen snel berekend worden en hebben een dimensionaliteit van 176. Een experiment tussen twee datasets, om de generalisatie de evalueren, produceert kleinere fouten voor onze functies dan andere modellen die gebaseerd zijn op functies van de bovenste plank. Op basis van de behaalde resultaten laten de voorgestelde functies zien dat ze geschikt zijn voor real-time applicaties, zeker

voor draagbare apparaten met gelimiteerde processoren.

Hoofdstuk 6:

Gemotiveerd door de gevoeligheid van leeftijdsaanduidingen door ruis in plaatjes, wordt in dit hoofdstuk de invloed van de kwaliteit van plaatjes op het schatten van leeftijd onderzocht. Aanpassingen in de kwaliteit van plaatjes hebben een invloed leeftijdsaanduidingen, die voornamelijk zijn gebaseerd op huid texturen en rimpels. Wij stellen een plan voor om de toevoeging van verschillende types ruis op het schatten van de leeftijd te isoleren. Drie gebruikelijke types van digitale ruis worden onderzocht.; JPEG-compressie ruis, quantizatie ruis, en schaal ruis. Vier datasets voor het schatten van leeftijd worden gebruikt in de experimenten.

Uiteindelijk wordt ook een kader voorgesteld om automatisch, op basis van de kwaliteit van een plaatje, de meeste geschikte functietype toe te passen. De kwaliteit van een plaatje wordt automatisch voorspeld. Resultaten laten zien dat automatisch verschillende functies gebruiken op basis van de kwaliteit van plaatjes leidt tot betere of vergelijkbare resultaten dan de functie die individueel het beste presteert.

Hoofdstuk 7:

Dit hoofdstuk behandelt het probleem van het automatisch kalibreren van schatters van staarrichting waarbij geen actieve inmenging van gebruikers nodig is. Op basis van de observatie dat mensen vergelijkbare staarpatronen hebben wanneer ze kijken naar dezelfde prikkelingen, wordt een nieuwe aanpak voorgesteld. De patronen in staarrichting van individuen wordt gebruiken om staarpunten te schatten voor nieuwe gebruikers, zonder actieve calibratie.

De experimentele set-up is flexibel en ontworpen om praktische scenario's te simuleren: 1)een webcamera is gebruikt, 2) geen steun voor de kin is nodig, en 3) kijkers hoeven maar voor drie seconden naar de prikkelingen te kijken. De voorgestelde methode geeft een fout van 4.2 graden. Om verdere inzichten in de voorgestelde aanpak te krijgen, testen we de methoden in verschillende gevallen. Ten eerste, om het effect van fouten in de afmetingen van het oog op de fout in de schatting van staarrichting te isoleren, worden preciezere diodes gebruikt, wat de fout terugbrengt tot 3.1 graden. Ten tweede, niet-gekalibreerde staarpatronen van andere gebruikers, in vergelijking met gekalibreerde staarpatronen, worden gebruikt, wat vergelijkbare resultaten geeft. Experimenten laten zien dat de heterogeniteit tussen de staarpatronen van de kijkers een invloed heeft op de fout van automatische kalibratie. Voor zover bij ons bekend, is dit het eerste werk dat staarpatronen gebruikt om schatters van staarrichting automatisch te kalibreren.

Acknowledgments

Seven and a half years ago, I arrived to Amsterdam to pursue a master's in AI. I would never have expected such a great journey ahead. There were many ups and downs but I have met amazing people who have influenced me and helped me to become who I am today. They provided their support, advice, wisdom, and love, each in their own way. This acknowledgment is a way to express my gratitude to those people. While I am pleased and proud of this thesis as a scientific work, I am more proud and more pleased with the social and personal experience of the past eight years. I oscillated between different priorities until I finally reached a 'local' maximum.

The first word of gratitude goes to Theo Gevers. My relationship with Theo goes back to my very first days in Amsterdam. In addition to being my PhD promotor, Theo was my teacher in my master studies, master thesis supervisor, and boss at ThirdSight. When I became homeless, he hosted me for two months while I was taking care of his two 'wild' cats (or whatever species they belong to!). I always felt welcome in his house and with his family. Whenever I had a problem, I could always go to Theo and ask for advice or an opinion, and I always found a listening ear and a smart, to-the-point answer.
Theo, I have great appreciation for your knowledge and help. I really enjoyed learning from you how to research and write papers, how to keep a positive attitude and think solution-oriented, and how to set priorities right, at work and in life. I learned a lot from you, directly and indirectly, and enjoyed your company (excluding when playing football :)). Theo...you the man!

I would like to thank Arnold Smeulders, my co-promotor, for the help and discussions. Although brief and at the end, it was insightful and interesting.

Sezer Karaoglu, thanks for being such a great friend, colleague and neighbor. We started together four and a half years ago. I could always rely on you when I had a problem. Your help and support are highly appreciated. I had really great times with you, espe-

cially when we were in a foreign country with no idea where we were or where we were going to! You are a great man.

Roberto Valenti, thanks for your collaboration and discussion during the first part of my PhD, and for being such a cool boss at Sightcorp and a good friend. I am very pleased with our collaboration which resulted in the ICCV paper and the patent. I am more grateful for knowing you as a caring and supportive friend. Hamdi Dibeklioglu, thanks for your help and collaboration especially with the TIP paper. Zhongyu Lou, we had a bumpy start in the beginning of our collaboration, but we finally managed to get something nice out of it. Thanks for your work and collaboration.

I met friendly people at ISLA lab. My officemates were very nice in putting up with me during the last four years. A big thank you to Svetlana Kordumova. You have a positive attitude which vibrates all over the office. You were always kind, supportive, and willing to lend an ear when asked. Thanks for your nice company and friendship. I also would like to thank Jan van Gemert for the scientific discussions and advices, Dennis Koelma for your technical (and transportation) help, Silvia Pintea, Pascal Mettes, and Spencer Cappallo for proofreading my last-minutes drafts, Jasper van Turnhout for designing the cover, Cees Snoek for the support, Ran, Honza, Morris, Yang, Amir, Masoud, Zhenyang, Efstratios, Marcel, Jorn, Agni, Kandan, Dung, Gosia, Stevan, Ivo, Vlado, Thomas and Mihir for giving me such a nice time during the last few years. The Sightcorp crew, Aziz, sofia, Jasper and Diego, thanks for your collaboration and friendship.

I want to thank Khalil Sima'an for the interesting lunches and the advices he gave me during my Master and PhD studies. A special word of gratitude goes to Janet. Thanks for your care and support especially during the first part of my PhD. They definitely were a driving force for me to keep up with my research. For all my friends who were supportive and understanding, a big thank you.

The last word of gratitude goes to my family: my brother Firas and sisters Huda and Nada for your love and always believing in me, my uncle Mahmoud and aunt Umayma for your love and warm support since I was a little boy. Finally, my parents Ibrahim and Mariam, who I dedicate this work to. Your love has been always unconditional and unlimited. Despite the very difficult period you had back home in the last few years, your care, love, and support were increasing by the day...an enormous thank you.