NON-RIGID REGISTRATION-BASED DATA-DRIVEN 3D FACIAL ACTION UNIT DETECTION

by

Arman Savran

B.S., in Electronics and Communication Engineering, İstanbul Technical University, 2002

M.S., in Electrical and Electronics Engineering, Boğaziçi University, 2004

Submitted to the Institute for Graduate Studies in
Science and Engineering in partial fulfillment of
the requirements for the degree of
Doctor of Philosophy

Graduate Program in
Boğaziçi University
2011

To my parents, Ayşen and Asil Savran ...

ACKNOWLEDGEMENTS

I owe my deepest gratitude to my thesis supervisor Prof. Bülent Sankur for his boundless support, guidance, patience and for allowing me to work in my own way. He is an inspirational person with his high ethical values and with his rigor in research as well as with his vast knowledge in various fields. It is a great honor for me to be student of him.

I am very grateful to Prof. Lale Akarun for providing the 3D acquisition device used in this dissertation, for allowing me to work in her laboratory and for her support.

I would like to thank Assoc. Prof. Burak Acar and Prof. Lale Akarun for being member of the thesis follow-up committee and for their feedbacks. I would like to thank Prof. Ethem Alpaydın and Assoc. Prof. Yücel Yemez for being part of my defense jury.

I would like to express my gratitude to those who made the experimentation in this thesis possible by collaborating and helping in preparation of the Bosphorus 3D face database. Foremost I offer my sincerest gratitude to Prof. Bülent Sankur and Prof. Lale Akarun for organizing the Bosphorus database campaign starting at eNTERFACE'07 workshop in Summer 2007. I also would like to thank all organizers of this workshop. I am grateful to Niyazi Ölmez for inviting many professional actors/actresses to participate in our database and for his guidance in preparing the acquisition setup. I am grateful to Taha Bilge for preparing ground-truth FACS codes meticulously, for conducting a questionnaire to analyze emotional expression content in the database and for helping in gaining deeper insight on psychological aspects of facial expressions. I am grateful to my colleagues for their efforts during and after data acquisition: Neşe Alüz, Oya Çeliktutan, Hamdi Dibeklioğlu, Erdem Akagündüz, Aydın Akyol, Nesli Bozkurt, Kerem Çalışkan, Cem Demirkır, Semih Esenlik, Bilgin Esme, Luca Teijeiro Mosquera, Tevfik Metin Sezgin, Jana Trojanova and İlkay Ulusoy. I would like to thank all the donors for their participation with great patience.

I would like to thank two institutions for allowing me to use their computer clusters on which most of the calculations reported in this dissertation were performed: TÜBİTAK ULAKBİM High Performance and Grid Computing Center, and TAM Institute in Boğaziçi University. I would like to thank the technical staff in both institutions for giving immediate support whenever I had problems.

I would like to thank my friends that I have met during my Ph.D. study for making my department a friendly and cheerful environment. Many thanks to Hatice Çınar Akakın, Çiğdem Aksu, Ebru Arısoy, Sergül Aydöre, Doğaç Başaran, Doğan Can, Oya Çeliktutan, Çağlayan Dicle, Erinç Dikici, Helin Dutağacı, Bilgin Esme, Fulya Kunter, Sıddıka Parlak, Deniz Seviş, Temuçin Som, Ekin Şahin, İpek Şen, Gönenç Tarakçıoğlu, Sinan Yıldırım, and many others. Also thanks to Assoc. Prof. Murat Saraçlar for his frequent cheerful visits to the laboratory.

I would like to thank Nilay Aktürk, Aidin Dario, Ebru Doğan, Erdem Eren, Selçuk Hazar, Özer Özcan, Hasan Sicim, and Osman Yüksel for their friendship. Having countless funny moments with them helped a lot in getting through the difficult last months of this dissertation. I am particularly very grateful to Hasan Sicim for his friendship and many helps during the last year, especially for giving me a room in his house.

Finally my mother Ayşen, my father Asil, and my sister Aslin. Without their endless love and support this thesis would not be possible. I dedicate my dissertation to them.

During my Ph.D. study I was supported by a number of institutions and projects: by the Scientific and Technical Research Council of Turkey (TÜBİTAK) BİDEB Doctorate Fellowship, by Boğaziçi University Research Fund (BAP) Project 09HA202D, by the Turkish State Planning Organization (DPT) under the TAM Project number 2007K120610, by the TÜBİTAK Project 107E001 grants, European FP6 NoEs SIMILAR and Biosecure projects, and BÜVAK.

ABSTRACT

NON-RIGID REGISTRATION-BASED DATA-DRIVEN 3D FACIAL ACTION UNIT DETECTION

Automated analysis of facial expressions has been an active area of study due to its potential applications not only for intelligent human-computer interfaces but also for human facial behavior research. To advance automatic expression analysis, this thesis proposes and empirically proves two hypotheses: (i) 3D face data is a better data modality than conventional 2D camera images, not only for being much less disturbed by illumination and head pose effects but also for capturing true facial surface information. (ii) It is possible to perform detailed face registration without resorting to any face modeling. This means that data-driven methods in automatic expression analysis can compensate for the confounding effects like pose and physiognomy differences, and can process facial features more effectively, without suffering the drawbacks of model-driven analysis. Our study is based upon Facial Action Coding System (FACS) as this paradigm is widely accepted to be capable of describing practically all types of human facial expressions and enables their systematic evaluations. Coding with FACS is done with Action Units (AUs) that are closely related with muscular activations. To validate the first hypothesis we develop person-independent detectors and intensity estimators of AUs, which use 2D maps of 3D facial surfaces. This approach enables us to compare 2D luminance modality with the 3D surface geometry data modality under the same set of algorithms. In addition, our detectors and estimators are free from biases of model-driven techniques to guarantee a fair assessment of the two modalities. For the second hypothesis, we first investigate non-rigid registration on 2D facial surface curvature maps. Our non-rigid registration algorithm is capable of handling large deformations and yet it is computationally efficient. To realize our second hypothesis we explore and develop AU detectors using this algorithm. Our work is the first example of detailed registration in data-driven expression analysis and surpasses the performance of state-of-the-art AU detectors.

ÖZET

ÜÇ BOYUTLU YÜZLERDE ESNEK KAYITLAMAYA DAYALI VERİ-GÜDÜMLÜ EYLEM BİRİMİ SAPTAMA

Yüz ifadelerinin otomatik analizi, akıllı insan-bilgisayar arayüzleri ve insan yüz davranışı araştırmaları gibi potansiyel uygulamalarından dolayı aktif bir calışma alanıdır. Otomatik ifade analizini daha ileri götürmek için bu tez şu iki hipotezi önermiş ve deneysel olarak kanıtlamıştır: (i) 3B yüz verisi geleneksel 2B kamera imgelerinden, sadece poz ve aydınlanma etkilerine karşı dayanıklılığıyla değil, aynı zamanda gerçek yüz yüzeyi bilgisini barındırdığı için de daha iyi bir veri kipidir. (ii) Detaylı yüz kayıtlamasını herhangi bir yüz modellemesi yardımı olmadan, yani veri-güdümlü olarak gerçekleştirmek, model-güdümlü analizin dezavantajları olmaksızın poz ve fizyonomi farklılıklarının bozucu etkileri telafi edildiğinden ve yüz öznitelikleri daha etkin olarak işlendiğinden, ifade analizinde gelinen en son noktayı ilerletmek için mümkündür. Bu çalışma, neredeyse bütün insan yüz ifadelerini betimleyebildiği varsayılan bir paradigma olduğundan ve sistematik karşılaştırmaları olanaklı kıldığından, Yüz Eylemi Kodlama Sistemi (FACS) temel alınarak yapılmıştır. FACS kodlamaları kas aktivasyonları ile yakından ilişkili eylem birimleri (AU) ile yapılır. İlk hipotez için, 3B yüz yüzeylerinin 2B'lu haritalarını kullanan kişiden-bağımsız AU saptayıcıları ve yoğunluk kestiricileri geliştirilmiştir. Bu yaklaşım, 2B'lu ışıklılık ve 3B'lu yüzey geometri veri kiplerini aynı algoritmalar altında karşılaştırmayı olanaklı kılar. Ayrıca, bu saptayıcı ve kestiriciler model-güdümlü yöntemlerdeki yanlılık olmadan iki kipin adil olarak değerlendirilmesini garanti ederler. İkinci hipotez için, önce, 2B yüz yüzeyi kıvrımlılık haritaları üzerinde esnek kayıtlama araştırılmıştır. Geliştirilen kayıtlama algoritması büyük deformasyonların üstesinden gelebildiği gibi işlemsel olarak da verimli çalışmaktadır. İkinci hipotezi gerçekleştirmek için bu algoritmayı kullanan AU saptayıcıları araştırılmış ve geliştirilmiştir. Önerilen yöntem, veri-güdümlü ifade analizinde detaylı kayıtlamanın yapıldığı ilk örnektir ve en güncel AU saptama yönteminden üstün gelmiştir.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

LIST OF SYMBOLS/ABBREVIATIONS

$\mathbf{d}(\mathbf{p})$	Constant 2D displacement vector at 2D point $\mathbf{p}$
D_A	Domain of the 2D-mapped reference surface Ω_A
D_B	Domain of the 2D-mapped target surface Ω_B
D_i	2D domain of input sample i
D_k	2D domain of reference sample k
$\mathbf{E}$	Green-Lagrange strain tensor
E_D	Deformation Energy
E_M	Matching Energy
E_T	Total Energy
$\mathbf{I_A}$	2D multi-modal reference image
$\mathbf{I_B}$	2D multi-modal target image
I_i	2D input image i
I_k	2D reference image k
$\mathbf{p}$	2D point
$\boldsymbol{\nabla}\mathbf{d}$	Displacement gradient tensor
$\boldsymbol{\nabla}\boldsymbol{\phi}$	Deformation gradient tensor
$\boldsymbol{\varepsilon}$	Cauchy's strain tensor
Ω_A	Reference Surface in 3D space
Ω_B	Target Surface in 3D space
$\boldsymbol{\phi}$	2D deformation
$\boldsymbol{\phi_{ki}}$	Estimated 2D deformation from the domain of k to the domain of i
ρ	Rigidity coefficient
3DMM	3D Morphable Model
AAM	Active Appearance Model
ASM	Active Shape Model
AuC	Area under the Curve

AU	Action Unit
CLM	Constrained Local Model
EMFACS	Emotional Facial Action Coding System
FACS	Facial Action Coding System
FACSAID	Facial Action Coding System Affect Interpretation Database
HMM	Hidden Markov Model
ICA	Independent Component Analysis
ICP	Iterative Closest Point
LSCM	Least Squares Conformal Mapping
NMF	Non-negative Matrix Factorization
NMFSC	Non-negative Matrix Factorization with Sparseness Constraints
PCA	Principle Component Analysis
PDM	Point Distribution Model
RBF	Radial Basis Function
ROC	Receiver Operating Characteristic
SVM	Support Vector Machine
TPS	Thin-plate Spline

1. INTRODUCTION

The human face is an effective carrier of various types of communicative and emotional information. From facial expressions one can infer about mental and/or physiological states, or can read social communicative messages instantaneously. Some examples of information displayed by the face are basic emotions like joy, anger and fear [1] or a host of other mental states like boredom and concentration [2]. Instances of reflection of physiological conditions are pain [3] and tiredness. Finally, facial expressions and head gestures are instrumental in regulating verbal communication [4]; or they produce non-verbal social interaction signals, such as winking [5]. The variety of facial signal sources indicate the potential of applications of automated expression recognition. Expression detectors can be useful in the design of intelligent human-computer interaction systems, in developing affective interfaces, or to enhance man-machine communication as in driver fatigue monitoring.

Majority of previous work on automated expression recognition has focused on the classification of the six basic emotions, which have been assumed to be universal [1]. There exist a number of detection methods that can successfully identify the basic emotions under controlled conditions, and good surveys of these methods can be found in [6, 7]. The common approach in these methods is to directly identify the expressions from the face image data. One difficulty with this approach in real life conditions is that, emotion classes and communicative intents can be displayed as facial expressions in very diverse ways. Furthermore, there exist a much richer set of identifiable facial expressions of interest beyond the basic six. As the variety and types of expressions increase, training direct face interpretation systems become increasingly cumbersome. In this respect, the paradigm of coded facial actions has the potential to be more flexible in expression interpretation. In this paradigm, facial expressions are decomposed into local deformations. Any expression can then be recognized based on the dictionary and the extracted codes of local deformations. Among the existing facial coding methodologies, a widely accepted one in the literature is the Facial Action Coding System (FACS) which defines 44 action units (AUs) to code the expressions

[8]. Therefore, automatic AU coding systems are very useful because there are several issues with manual scoring: too many rules and details defined in FACS necessitates certification, human coders can become fatigued quickly, manual coding takes a very long time, and coding is sometimes be subjective.

Automatic expression recognition methods can be grouped into two categories as model-driven and data-driven methods. In this context, model-driven analysis implies that a pre-designed model of human faces is fitted to the actual input faces before performing any analysis task. Modeling can be done by preparing geometric models, by modeling appearance variations of faces as well as their shapes, or simply by means of facial landmarks. After preparing these models in the training phase, they are used to analyze input face images by either landmark detection or model fitting algorithms. Use of face models brings several advantages: ambiguities for instance due to pose can be removed after model fitting, recognizers that are parsimonious in the number of features can be developed, and requirement on number of training samples can be reduced. On the other hand, in data-driven methods, no prior information about human faces is incorporated. These methods extract all the information from labeled training images, without the burden of model construction. They directly analyze appearance of faces in the input images without any bias due to modeling, and do not require prior fitting of the face models, which is an error prone process.

In this thesis we focus on two hypotheses:

(i) 3D facial surface measurements should achieve better performance in facial expression recognition as compared to conventional 2D camera images. This is because 2D luminance data depends non-linearly on pose, illumination and facial albedo as well as facial deformations. Due to its dependency upon several factors extrinsic to expression, luminance modality cannot represent facial deformations as faithfully as 3D measurements.

(ii) The factors of pose and of differences in physiognomy among subjects cause misalignment of crucial facial features and handicap automatic recognition of expressions. A model-driven approach has an inherent advantage of coping with

these confounding factors since correspondences between the face images are established through a face model. On the other hand, in data-driven analysis, after a coarse face registration step which is usually done based on detected face and eyes, we have to deal with all the extrinsic variations by means of statistical learning using the available data. The upside of data-driven analyses is that we do not have to struggle with building face models, error prone model fitting, and limitations or bias of the assumed model. Therefore, if we can adequately utilize detailed registration in data-driven analysis, we can improve the state-of-the-art in expression analysis, without suffering any drawbacks of model-driven analysis.

Previous 3D studies have focused on prototypical expression recognition, the so-called six universal expressions. However, in our research we follow the FACS paradigm since it has the potential to handle wide range of expressions and it is the de facto standard in facial behavior research. It enables systematic evaluation of the expression recognition methods since AUs are local and well defined, but still there are many research challenges such as discriminating between certain AUs that have subtle differences, handling of AUs which appear differently due to the interactions with certain others, and variations due to the differences in their intensity levels and asymmetries. In order to explore a wider range of facial expressions, we have developed the first 3D database consisting of a rich set of expressions with FACS annotations. Furthermore, as different from previous FACS databases, all AUs are annotated with five different intensity scales allowing for assessment of detection of low intensity expressions as well as AU intensity estimation.

We limit our scope to static 3D data. Though 3D video data is in principle more informative and could result in better recognition by learning temporal dynamics learning and using temporal features, we believe that there are significant research problems in 3D and that improvements could still be made with static data. A more compelling reason was that affordable 3D video acquisition devices in this time period were not available.

We develop person-independent automatic 3D AU detectors and intensity esti-

mators based on 2D maps of 3D facial surfaces in contrast to previous 3D studies. Thus, we efficiently process 3D data being independent of surface mesh resolution and topology. Another important point is that, 3D-to-2D mapping enables comparison of 3D geometry and 2D luminance modalities under the same set of algorithms. Moreover, we concentrate exclusively on data-driven analysis. Data-driven approach allows unbiased comparisons of 3D and 2D modalities, since model design always introduces bias that can favor one modality over other. Note that previous 3D algorithms required detection of facial landmark points, which could total anywhere between 64 and 82 points. Hence either these methods were not fully automated or they were prone to consequence of landmark detection failures. By means of data-driven analysis, we bypass the problematic landmarking requirement. Furthermore, we develop intensity estimators based on regression. Regression-based estimation has not been addressed in 2D and 3D literature so far. Our extensive experimentation prove that 3D is overall better than 2D AU detection and intensity estimation, and that fusion of 2D and 3D also yields improved results.

In order to study on our hypothesis on detailed registration in data-driven analysis, we first develop a novel non-rigid registration method. There are two crucial factors for a non-rigid registration algorithm if it has to be utilized for automatic expression analysis: First, it should be able to handle large deformations that can arise from expressions. Second, many applications as in human-computer interaction necessitate time efficient processing. It is known that computational demand of non-rigid registration algorithms, especially the ones suitable for large deformation handling, can be quite high. To satisfy these conditions we developed 2D deformable triangular mesh-based registration which utilizes hyper-elastic deformations so that large deformations can be dealt with. Although the inclusion of non-linear elasticity means heavier computation, yet the method is time efficient due to 2D triangular discretization since deformation gradient tensor over a triangular element is constant. In addition, triangular discretization is adapted to the surface geometry so that we avoid unnecessary computations.

Our novel data-driven approach benefits from detailed non-rigid registration and

its performance surpasses those of state-of-the-art data-driven techniques. Since we do not use face models, the resulting method is very practical, avoids commitment to any model, and does not necessitate intermediate model fitting steps. The two key points of our method is that, first, it provides compensation for factors like physiognomical differences (which are in effect non-rigid) as well as pose differences, and second, by using expression specific references transient features of expressions like furrows can be handled effectively.

The thesis is organized as follows: In Chapter 2 we review the Facial Action Coding System and survey such topics as, automatic AU recognition, use of 3D modality for expression recognition, AU intensity estimation methods, non-rigid registration in expression analysis, and non-rigid registration of 3D surfaces. Chapter 3 is devoted to our experimentation databases and methodology. In Chapter 4, we describe how AUs are analyzed on 2D maps of 3D facial surfaces and compare 3D versus 2D modality under several state-of-the-art methods. In Chapter 5, we develop a non-rigid 3D surface registration algorithm that is suitable for automatic expression analysis. This algorithm is then used in Chapter 6 for registration-based data-driven expression analysis. In that chapter we evaluate various aspects of the proposed approach and compare it with state-of-the-art data-driven AU detection. Finally, in Chapter 7 we summarize original contributions of our research and present several future directions.

2. LITERATURE SURVEY

Research on automatic facial expression recognition has been very active for about two decades. These studies have been conducted in two veins: direct recognition of expressions and recognition of expressions via facial action units. In the later case, the goal is to detect atomic components of expressions which can be instrumental to recognize expressions as well as being useful in facial behavior research. However, most of the existing studies follow the direct recognition paradigm, and majority of them concentrates on the six basic emotions, since they are mostly available in several databases and due to the claim [1] that they are universal. Good surveys of these two directions can be found in [6, 7, 9, 10].

2.1. Facial Action Coding System

FACS has been developed by behavioral scientists [8, 11] to enable objective measurements of facial activity. It has been employed as a solution in a diverse set of problems. Some application examples where FACS has been particularly useful are as follows: Testing theories of emotion and affective processing in the brain [12]; detecting emotions from reflexes such as startling [13]; discriminating between lying and telling the truth [14]; the way cues about smoking invoke ambivalent reactions in refrained smokers [15]; determining if variables other than nurturing attitudes exert an effect on the affective relationship between mothers and newborns [16]; discriminating a faked expression of pain from a genuine one even in children [17]; assessing how depression affects reactions to hilarious stimuli [18]; and prediction of psychopathological conditions [19].

FACS incorporates 44 AUs, which are very closely related to muscle activations, but the coding rules are based on observations made from images. FACS makes a distinction between upper and lower face AUs and necessitates evaluating them in isolation. In cases where certain action units interact and create a non-additive combination, coders use modified criteria for the rating of AUs. Finally, FACS also provides

for the intensity rating of AUs on a five point scale. In summary, FACS can be interpreted as the alphabet of facial expressions under some simplifying assumptions. So far, over 7000 different AU combinations have been observed, and substantial amount of data relating AUs to expressions has been collected. For instance, Emotional FACS (EMFACS) describes basic emotions in terms of AUs, and FACS Affect Interpretation Database (FACSAID) includes various affective states. A sample coding for fear expression is shown in Figure 2.1.

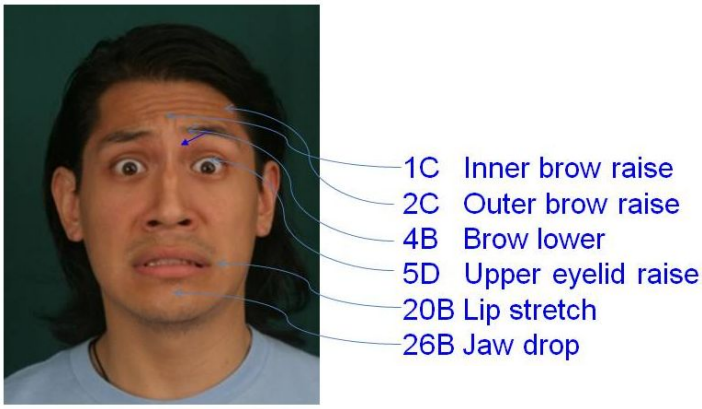


Figure 2.1. A sample FACS coded face expressing fear emotion [20]. The FACS code for this particular expression is 1C+2C+4B+5D+20B+26B.

Due to the countless rules defined in FACS and subtleties of AUs, encoding is a difficult and time consuming process, and requires certification. Also, due to the human factor in coding, analysis may be subjective by nature and is prone to human mistakes. The factors such as modified criteria for rating AUs when certain combinations are observed and inclusion of traces in the intensity rating of AUs render manual AU coding a very demanding technique. Although Sayette et al. [21] argued that the psychometric features of FACS, such as inter-coder reliability, was in the range of good to excellent, they also pointed out to the fact that the five point scale used in intensity led to lower reliability measures in the rating of certain action units. Intensity scoring in FACS is described in Section 2.1.1. Therefore, an automated coding system of facial actions not only will ease the tedious human effort, but will be free of human errors.

2.1.1. FACS Intensity Scoring

FACS has developed certain conventions and rules for scoring intensities of Action Units. Scoring is done on a five-point ordinal scale, A-B-C-D-E, if evidence of an AU is present. The interpretation of these levels are as follows: the A level refers to a trace of the action; B, slight evidence; C, marked or pronounced; D, severe or extreme; and E, maximum evidence. Scoring criteria depend upon the scale of evidences, and the evidence can be assessed in terms of the degree of appearance change or in terms of the number of appearance changes. The relationship between the scale of evidence and the scoring levels is a bit different for some AUs. Scoring criteria are listed in the FACS manual [11] for each AU, though sometimes modified criteria are used depending on the AU combinations.

By definition, each level denoted by a letter refers to a range of appearance changes, that is, they do not correspond to a single strength of AU. Notice that the intensity scale is not divided into uniform intervals; the C and D levels cover a larger range of appearance changes. The relationship between the scale of evidence and intensity scores is depicted in Figure 2.2.

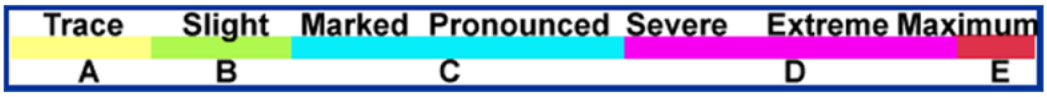


Figure 2.2. Relation between the scale of evidence and intensity scores [11].

FACS manual states that scoring of lower intensities, the A and B levels, requires particularly careful examination, and A level actions can be scored reliably only by very experienced coders. While scoring of lower intensities may not be easy, distinguishing the E level AUs can be difficult as well since the intense muscular contractions of the E level combine with the person's individual physical characteristics causing variability on the appearance changes across different people. Examples of low level (B) and high level (E) AU 5 - Upper Lid Raiser images are shown in Figure 2.3.

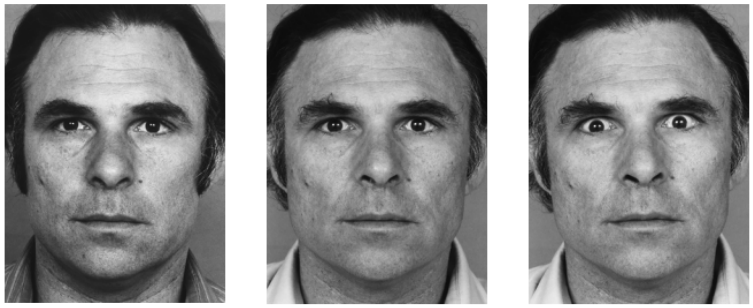


Figure 2.3. AU 5 - Upper Lid Raiser. Neutral (left), B level (middle) and E level (right) samples are shown [11].

2.2. Automated Detection of Facial Actions Units from 2D Images

FACS defines AUs in terms various facial features. Understanding of these features is important to develop analysis algorithms. There are two distinct types of facial features that convey facial expression information on 2D images:

- *Permanent facial features* are always available. Since these features may be deformed due to facial expressions, they carry information about expressions as well as identity. Expressions give rise to deformations mainly on lips, eyebrows and eyelids. We can also mention facial hair, tissue texture, and permanent furrows as helpful facial features especially in video analysis since they do not disappear in an expression analysis session.
- *Transient facial features* occur only due to the expressions. These are wrinkles, furrows, bulges, and opening/closing of eyes and mouth (Figure 2.4).

Compared to recognition of prototype expressions, automatic detection of AUs is a more challenging problem since there are only subtle differences between some of the AUs, and also the appearance of an AU can change a lot when it occurs in combination with other AUs. In the sequel we discuss previous work on AU detection in terms of the following categorizations: holistic versus local analysis, use of static images versus video data, and model-driven versus data-driven analysis.

2.2.1. Holistic versus Local Analysis

In holistic methods, face images are processed as a whole and extracted features do not represent local image regions. These methods usually perform well at identifying some common expression classes directly from images, like the basic six emotions. However, they are not sensitive to subtle changes in small areas and therefore generally cannot handle detection of locally defined AUs. Moreover, holistic analysis requires accurate registration. Since they are susceptible to background clutter, they also necessitate proper segmentation of faces from background. Donato et al. [22] compared several holistic analysis techniques with local analysis ones, and showed the disadvantages of holistic methods, as in the case Principle Component Analysis (PCA) versus Gabors. Lucey et al. [23] also compared features extracted from the whole face with features from the eye regions to detect eyebrow AUs. Their study was based on Active Appearance Models (AAMs) [24] and obtained only slightly better results with regional features.

2.2.2. Static Images versus Video Images

Video contains in principle more information about expressions than static images, and therefore many authors have tried to benefit from it. However, successful applications of completely static methods on selected video frames have also been demonstrated [25]. Video data can be used in three different ways or levels. First, it can be employed to simplify the feature extraction. Since facial motion between successive frames are not typically expected to be large, instead of feature detection, tracking techniques can be applied efficiently and then static methods can be used. As examples we can mention feature point tracking [26] or AAM tracking [27]. Second, extracted features can include temporal information, for instance by using frame velocity and acceleration of the static features [28]. As another example, Peng et al. [29] proposed dynamic haar-like features to improve AU recognition. An alternative technique is to use optic flow estimation. Donato et al. [22] showed that image-based static analysis outperforms dense optic flow-based AU detection by about 10%. However, a recent method [30], which is based on motion orientation histogram descriptors, attains high

recognition rates.

A third approach is to exploit temporal dynamics during the inference stage rather than keeping dynamic information only in features. Tong et al. [31] have employed the data-driven method of Bartlett et al. [25] together with Dynamic Bayesian Networks and thus exploited the dynamic and semantic relationships of AUs. A recent work of the same group [32] also show the benefit in learning these relationships. However, though it improves recognition, it also has the risk of increasing dependency on the context of the training data and thus they may not be generalized well to different databases exhibiting novel contexts. As another example, which combines temporal modeling with temporal features extraction, employs Hidden Markov Models (HMMs) on motion field features [30].

2.2.3. Model-driven versus Data-driven Analysis

Model-driven analysis implies that a pre-designed model of human faces is fitted to the actual input faces before performing any analysis task. These models vary from the simplest landmarking to complex approaches like AAMs (active appearance models). In the expression recognition stage, shape or appearance variations are analyzed based on the outcome of the fitted model.

Among model-based techniques, Tian et al. [33] fit geometric models of facial features in the form of ellipses and curves, and apply wrinkle detectors, as shown in Figure 2.4. Pantic and Rothcrantz [34] extract geometric features from facial landmark points, and they use profile as well as frontal views to benefit from depth information. Instead of using landmarks and local geometric models, 3D models of faces can also be fitted [35, 36].

However, these methods require detection of a high number of landmarks, which is itself a challenging problem. In view of the missing or erroneously located landmarks, some authors use Point Distribution Models (PDMs) to regularize the landmark configuration. PDMs learn facial shape variations in order to constrain them to plausible

Figure 2.4. Facial feature extraction by Tian et al. [33]. Permanent and transient facial features like nasolabial furrows are extracted by geometric model fitting. Wrinkles are detected.

face shapes. Since the introduction of Active Shape Models (ASMs) [37], PDMs have been used in various applications. The methods are composed of two stages: an exhaustive local search for each landmark, followed by optimization of the global non-rigid face shape parameters that results in the minimization of local responses for all the landmarks. Recently, these methods have been named as Constrained Local Models (CLMs) [38, 39]. In contrast to CLMs, AAMs (active appearance models) model both holistic appearance variations and the shape. For instance, Lucey et al. [23] apply both 2D and 3D AAM tracking to classify four upper face AUs. However, though AAMs perform very successfully in a person-dependent manner current AAM based expression analysis systems do not generalize well to novel subjects.

In contrast, data-driven techniques depend completely upon statistical learning, i.e., images are analyzed directly without the intermediary of any facial landmark detection and/or model fitting. This is realized by first extracting a set of facial appearance features and then classifying them. However, compared to model-based methods, literature on AU detection via data-driven methods is limited. Among the few existing works we can cite Donato et al. [22], who compared various data-driven techniques in the early period of automatic AU detection research. More recently, a data-driven method proposed by Bartlett et al. [25] achieved state-of-the-art results better than the previous person independent detectors. Their solution is to perform local analysis by Gabor wavelets from which features are chosen automatically with the AdaBoost algorithm. Figure 2.5 depicts the algorithm. First, face detection is performed, and then eyes are detected which in turn are used to register and crop the face. For AU classification, AdaBoost or Support Vector Machine (SVM) classification is applied on

Table 2.1. Advantages of model-driven and data-driven analysis approaches.

Model-driven Analysis	**Data-driven Analysis**
- Benefits from detailed registration	- Avoids error prone model fitting steps
- Works with much fewer features	- No burden of manual face model preparation
- Smaller amount of samples for training	- No bias due to an assumed face model

the selected Gabor features. The same research group developed this approach further with the inclusion of new feature types, such as box filters, edge orientation histograms, and local binary patterns, to construct a real world smile detector [40]. This detector proved to be successful over a very large database (of over 25,000 faces), which includes faces with high variability in pose, illumination, facial hair, age, etc.

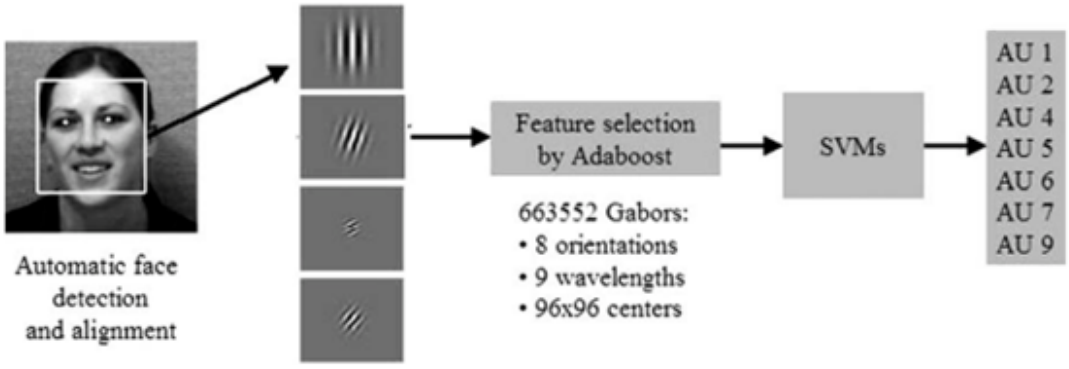


Figure 2.5. Data-driven AU detection by Gabor waveletes [25].

The difference between model-driven and data-driven analysis is the incorporation of prior information on human faces in model-driven techniques. Advantages of the two approaches, i.e., the pros and cons of including prior knowledge on faces, are compared in Table 2.1 and discussed in this sequel.

The first advantage is registration: Model-driven techniques enable detailed registration and they help remove initial ambiguities. A very common source of ambiguity is the head pose difference. In fact, for 2D images, large out-of-plane rotations have severe consequences since the appearance also change considerably. Head pose can be more effectively handled with model-driven methods, and 3D models can cope with large out-of-plane rotations. Note that though our concern in this work is facial de-

formations, nevertheless the estimated head pose as a byproduct of registration may also be useful whenever poses carry information about expressions. Differences due to subject-identity, e.g., physiognomy, form another sources of variation not related to expressions. By means of non-rigid registration, for instance based on facial landmarks, we can further suppress this source of variability due to subject-specific facial feature configuration. However, intermediate registration steps, which are performed via landmark detection or other means of model fitting, run the risk of being caught in a local minimum during fitting or has detection failures. Therefore, data-driven techniques have the advantage of bypassing the weakest-link-in-the-chain issue of these intermediate steps.

A second advantage of model-driven methods is that, recognition of expressions using only a small amount of features is possible, which is much more practical. In other words, we do not have to deal with all the image data even in training in contrast to data-driven methods. For instance, after reliable model fitting, coordinates of many landmarks become automatically available. We can then simply pick up those landmarks or texture around them known to be instrumental in expressions, e.g., to apply Gabor patches or to detect wrinkles. Another example is the use of estimated model parameters, as usually done with AAMs.

However, the cost of working over a few features is that, we need appropriate face models to extract these features. Preparing statistical face models for improved landmark detection is known to be a tedious process since precise labeling of many landmarks over many faces are required to account for non-rigid facial deformations. Alternative face models, e.g., based on geometric parametrization, are also possible, but requires extensive anatomical knowledge and tedious trial and error phase, and actually they are not commonly used models.

The third advantage of model-driven methods is the smaller amount of required training data, since they are not data-greedy as data-driven methods. Data-driven methods have to deal with variations, such as due to pose and identity, by means of learning them all, hence they require more samples to capture the spectrum of possible

variations. Moreover, model-driven methods permit design of rule-based classifiers [34, 41] to avoid complete expression training. However, it does not always mean that data-driven analysis requires more training samples. Some statistical face models, for example bilinear models [42], may also demand high amount of training data to be constructed.

Although we can utilize facial knowledge at different levels in order to simplify expression recognition, this advantage may also turn into a handicap since a model also means a prior commitment. For instance, one can miss some important image features by not going through a bottom-up analysis of the data. Another example is the modeling of shape variations as done with ASM and AAMs. The non-rigid deformations that can be estimated by these types of models are limited by piece-wise linear deformations over coarse triangulation of facial landmarks and their training database. A more severe example is the bias introduced by AAMs due to the modeling of appearance variations in addition to shape. It is known that AAM techniques that are employed in current expression analysis systems works only in a person-specific way, that is, they are biased towards specific identities. In summary, data-driven methods are free from bias or limitation due to an assumed face model.

2.3. 3D Modality for Expression Recognition

The advantages of 3D measurements have already been demonstrated in the context of face recognition, as 3D is immune to illumination and to some extent to pose variations [43]. Illumination and pose are very important issues also in automatic expression recognition. Therefore, there has been some work to explore the use of 3D expressions. These studies have been performed on the six universal emotions [1] (happiness, surprise, fear, anger, sadness, disgust). For instance, using 64 manually marked points, Wang et al. [44] have divided 3D faces into regions and extracted regional histograms of surface curvatures. Other 3D methods proposed for emotion identification also have the handicap of depending on an excessive number of feature points [45, 46, 42, 47].

There has been recently also some work on expression recognition using simultaneous 2D and 3D video [48, 49]. Both of these two independent works perform model-based analysis on 2D luminance images, by either ASMs or AAMs, to track more than 80 facial points from which 2D and 3D data features are then extracted. However, these studies are mostly on prototypical emotional expressions and consider AU detection in a very limited way. For instance, only 11 singly occurring posed AUs are recognized in [49] with a rule-based classification over a small database.

Previous work has not shown the potential of AU detection adequately with 3D observations. Our conjecture is that 3D face data can alleviate difficulties inherent in 2D modality since 3D enables true facial surface measurements, and hence, subtle differences of AUs may be better discriminated and we may go beyond the limits of 2D.

2.4. Facial Action Intensity Estimation

In contrast to AU detection, there is much less work in the literature on AU intensity estimation. The measurement of intensities can be useful in behavior research and for improved FACS coding. For instance, if the expression is surprise, and if AU 5 - Upper Lid Raiser is available, then, it should only be at B level. Second, estimating strength of the AUs we yield more information about mental state and emotional involvement of a subject. Furthermore, AU intensity outputs can be a basis for studying AU dynamics.

One of the early works on AU intensity estimation has been done by Pantic and Rothkrantz [50], who developed an expert system using geometric features extracted from dual-view (frontal and profile) images. However, their study is person-dependent and requires detection of high number of landmarks. Most of the other works on expression intensity have investigated relationships between classification decision scores and intensities. For instance, Bartlett et al. [25] investigated correlations between intensity levels and SVM classifier margins of their Gabor filter-based detectors. They reported moderate to high correlations for several AUs. One criticism of using classifier scores is

that they do not incorporate intensity knowledge. Yang et al. [51] have used the output scores of RankBoost based expression classifiers to better deal with intensity variations. They train RankBoost classifiers with onset to apex ranked image sequences, in order to rank the image pairs according to their emotion intensities. They obtained better image pair ordering performance than the linear SVM-margin approach in et al. [25]. However, though related, correctly identifying ranking of image pairs in a sequence whose intensity increases monotonically is quite a different problem than estimating intensities directly from single images. Besides, some additional techniques should be figured out to convert ordering of image pairs into intensities, also in a way that we have consistent intensity measurements between sequences. Recently Mahoor et al. [52] studied measurement of AU 6 and AU 12 intensities over six subjects via person-specific AAMs. They approach to intensity estimation as a classification problem and apply six level SVM classifiers by one-against-one technique. For feature extraction they perform AU specific dimension reduction by applying regularized locality preserving indexing on appearance data, and use delta features (i.e., by neutral face feature subtraction).

2.5. Non-rigid Registration in Expression Analysis

Registration can be said to be the most crucial preprocessing step in face analysis. For instance, by rigid registration we get rid of the pose variations. Similarly, by non-rigid registration we may compensate other sources of variations. Expressions show themselves on local face regions, as permanent and transient features (see Section 2.2). However, since locations of face parts change from person to person, analysis of these features will be affected from the person-specific variations. By means of non-rigid registration, we can better align the face parts and thus can suppress some identity related variations. Moreover, non-rigid registration can be used to estimate expression deformations, which can especially be useful to capture temporal information in video. In the following subsection, existing techniques that benefit from non-rigid face registration are covered.

2.5.1. Non-rigid Registration in Model-driven Expression Analysis

In a model-driven scheme, after facial landmark detection or face model fitting, non-rigid registration can be applied based on the correspondences of detected landmarks or fitted models. Therefore, model-driven methods are very suitable to develop expression analyzers based on non-rigid registration. Below we describe several model-driven techniques that apply non-rigid registration.

2.5.1.1. Landmark-guided Registration. Koelstra and Pantic [30] utilize non-rigid registration to estimate AU deformations in video sequences with the guidance of facial landmarks. Their method involves several steps. The first step is performed to suppress intra-sequence variations, i.e., rigid head motion throughout the sequence. This is achieved by affine registration of facial part of each frame to the facial part of the first frame by the squared sum of differences of the grey level values. For this purpose, 20 facial landmarks are detected and tracked in the video sequence. The second step aims to suppress inter-sequence variations, i.e., facial shape differences. For this purpose nine of the landmarks that are assumed to be expression invariant are used for rigid alignment to a predefined reference set of facial points. After these two alignment steps, B-spline based free-form deformation technique is applied to register successive video images. The registration process is illustrated in Figure 2.6.

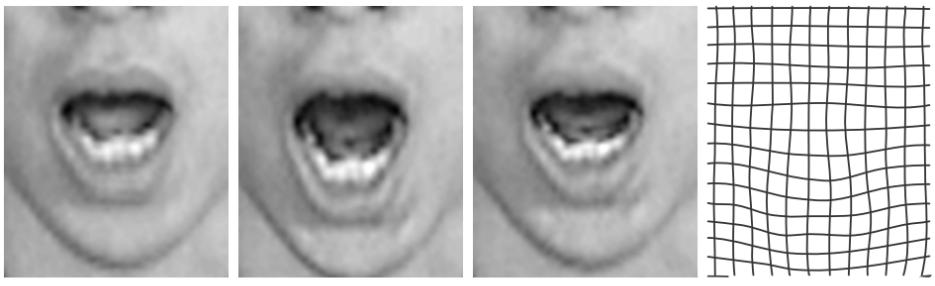


Figure 2.6. An illustration of non-rigid registration process between successive video frames [53]. (a) and (b): face parts of successive frames, (d): estimated deformation field, (c): initial frame deformed by the estimation.

2.5.1.2. Model Fitting-based Registration. Face model fitting essentially achieves the registration task also, since we can obtain correspondences between two images through the established correspondences between their fitted models. For instance, Lucey et al. [27] employ AAMs not only for locating the landmarks but also for face shape normalization in their AU recognition study. AAM is a generative linear model of face shape and appearance variations [24]. Registering an input face image via AAMs requires the estimation of geometric similarity, shape and appearance parameters so that the generated model image becomes most similar to the actual input image. Figure 2.7 illustrates an example fitting and normalization done by Lucey et al. [27]. Their AAM mesh and the input image are shown in the first column. In the middle column we see the normalization of pose by removing estimated geometric similarity transformation, i.e., rotation, translation and scale. Finally, last column shows the shape normalization realized by transforming the estimated shape to average base shape.

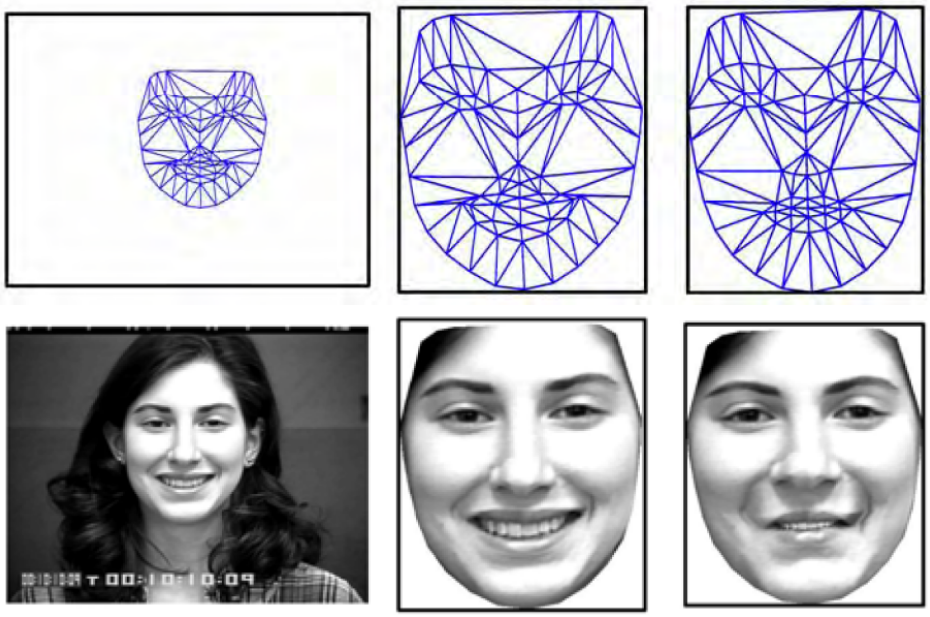


Figure 2.7. Face shape normalization of the appearance via AAM fitting [27]. First column: Input image and the AAM mesh fitted to this image. Second column: pose normalization. Third column: shape normalization by transforming to average shape.

3D Morphable Models (3DMMs) [54] can be thought as 3D counterparts of AAMs. 3DMMs work on range data and resemble to AAMs in that both texture and geometry data is modeled. However, while AAMs capture 2D landmark variations, 3DMMs deal with the 3D coordinate variations in the whole face and require dense correspondence

estimation. The accurate estimation of correspondences under expressions for model training is a major difficulty, and inaccuracies limit to usefulness of the generated models. Moreover, model fitting with 3DMMs is computationally much more demanding than AAMs. An example application of 3DMMs for expression recognition is presented in [55] where four emotional expressions are classified using the shape parameters of 3DMMs.

There are some problems with model fitting approaches. For instance, as discussed in Section 2.2.3, currently an important limitation of AAMs is the dependence on the subjects since they run in a holistic manner to match the subject's texture. Another problem is that the fitted model and extracted shape represent both identity and expression related variations. The common solution to mitigate the variations due to identity is to subtract the neutral face shape of the subject. However, one may not always have neutral faces of the subjects. A different approach to this issue is to learn the decomposition of identity and expression related shape variations. One way for this type of decomposition is to construct models that are able to generate any face, given its identity and expression parameters. For instance bilinear models [56] offer a way to combine linear identity and expression models. Recently, bilinear models have been employed in 3D emotional expression recognition [42]. Similar to 3DMMs, a model for 3D facial surface coordinates is constructed via a bilinear model instead of a linear PCA model. Thus two sets of coefficients are obtained, one for identity, the other for expression. The model is bilinearly fitted to the test face, and the resulting coefficients are used for expression classification. However, these more complex models necessitate higher amount of data to learn bilinear relationships between identities and expressions. Furthermore, they may not be adequate for detection of AUs which are more subtle and local than the prototypical expressions.

2.5.2. Non-rigid Registration in Data-driven Expression Analysis

In data-driven expression recognition, typically, face data are analyzed after an initial coarse alignment, for instance according to the detected eyes. Though pose differences may be somewhat reduced by initial rigid registration, one cannot get exact

correspondences between face points since differences due to identity and expressions result in non-rigid deformations between faces. As discussed in Section 2.2.3 absence of detailed registration is a disadvantage of data-driven methods. Therefore, if we can apply non-rigid registration without incorporating any face shape information, we can have the benefits of a data-driven solution and better quality registration without drawbacks of model-driven analysis. Recall that model-driven techniques have the disadvantage of model fitting issues, tedious face model preparation process and bias due to the assumed models (see Section 2.2.3). From the registration point of view we can mention linearity, low degree of freedoms and dependence on the training set as common bias issues. By using non-rigid registration techniques that do not depend on face models we can estimate more detailed deformations. However, registration without incorporating any face shape information is not a trivial problem for faces with expressions, and as far as we know, non-rigid registration has not been addressed in the framework of data-driven expression analysis.

2.6. Non-rigid Surface Registration

Non-rigid surface registration is an inverse problem which is ill-posed since there are numerous mappings from one surface to another. For registration one should decide first on the similarity measures to determine how well surfaces are matching and on the transformation model that constrain the mappings in order to alleviate the ill-posed nature of the problem. We overview the proposed registration techniques under landmark-based, 3D mesh-based and 2D mapping-based categories. We also describe and discuss the non-rigid 2D image registration problem in detail since our surface registration technique employs 2D image registration in the course of surface correspondence estimation for 3D expression analysis.

- *Landmark-based.* Landmark-based methods are quite common due to their simplicity. Dense correspondences over the whole surface are estimated based on given landmark correspondences. For instance, use of Thin-plate Spline (TPS) [57] or Radial Basis Function (RBF) [58] interpolation are common techniques. Another approach proposed by Praun et al. [59] is based on searching for a com-

mon parameterization of two surfaces. These methods are especially suitable when it is difficult to match local surface regions automatically, for instance, imagine our goal is to register 3D face of a dog onto a face of human. In landmark-based methods the level of detail depends on the number of given landmarks, and since their number will be limited these methods are only able to perform coarse non-rigid registration. However, it is also possible to incorporate surface similarities into to the registration process in order to achieve finer registrations. For instance, in the work of Allen et al. [60], a high resolution template mesh is optimally fitted to detailed human body range data with the given correspondences of sparse 3D markers. It is obvious that an important disadvantage of landmark-based registration is the decision on the landmarks to be used and also to be able find them automatically, which is not a trivial problem. Hence, these methods are usually employed when the task does not require fully automatic registration, or when coarse registration-based on few detectable landmarks is acceptable.

- *3D Mesh-based.* In 3D mesh-based methods correspondences are directly estimated over the surfaces based on the 3D surface mesh similarities. Shelton [61] has proposed a mesh-based method which is employed for morphable surface model construction. Their method perform registration by minimizing three energy terms defined on the surfaces so that surfaces become similar, deformations are smooth and resulting surfaces are not discontinuous. To obtain point-to-point correspondences between acquired faces, Yin et al. [62] fit a template face model. They search for similar and close vertices of the target mesh by energy minimization procedure. This algorithm is able to track facial expressions in a sequence of increasing valence values. A different approach using generalized multidimensional scaling algorithm is proposed by Bronstein et al. [63] for the dense correspondence estimation problem. They state that if objects are approximately isometric, one of them can be registered to another by finding corresponding surface points that have similar geodesic distances. The algorithm picks a small number of points from the reference, for instance 100 points to track 3D facial expressions, and thus run efficiently. However, there are some disadvantages of working on the 3D space. In general, they are computationally very intensive, sensitive to mesh resolution, their multiresolution implementation is difficult, and

they require proper initial alignment.

- *2D Mapping-based.* Some authors have found that registration performed in 2D resolves some of these handicaps of 3D processing. Blanz et al. [54] employed optical flow over the cylindrical projection of texture and range images for correspondence estimation of 3D neutral faces to construct morphable models of faces. As another example to the use of 2D mapping, Wang et al. [64] applied harmonic parameterization, which is a way of conformal mapping, to map 3D scanned faces onto image planes. They apply this mapping since it is stable and insensitive to resolution changes. Notice that their goal was not registration of highly deformed surfaces but registration of a subject expressions at high video rate by feature point tracking. Litke et al. [65] also worked in the parameterized domain. However, their surface parameterization technique is realized by non-linear optimization to minimize several types of mapping distortions. Their registration algorithm is based on variational principles and aims to register faces with different identity or expression. This algorithm minimizes image matching, regularization and feature demarcation energies to estimate deformations in the faces. As feature demarcations, they use eye and lip contours, and facial symmetry lines like the line dividing a face into left and right parts. These demarcations are created by manual segmentation to achieve accurate correspondence estimations. A drawback of 2D mapping-based methods can be the information loss due to the mapping. Depending on the surface and the mapping technique we can loose some important surface data. Various 2D mapping techniques based on surface parameterization with different characteristics are surveyed in [66, 67].

In our work, we use also 2D maps of 3D surfaces, due to the disadvantages of working on 3D surfaces, i.e., computational requirements and dependence on mesh topology and resolution. Once surface attributes are mapped onto 2D, we can employ any non-rigid registration technique on the resulting images. Non-rigid image registration is a broad topic that can be addressed from different points of view, such as similarity measure to match images, transformation model to change a reference image to match the target, or as an optimization process to estimate the transformation parameters. Various techniques of nonlinear registration are surveyed in [68, 69].

Our intent is to find a mapping between 2D surface images that establishes the image correspondences without using any landmarks, thus our work has similarities with the work of Blanz et al. [54]. However, in our case, we obtain the mapping according to the image constraint, not according to the optical flow constraint. For a pair of 2D images, I_A and I_B, the image constraint can be written as

$$I_A(u, v) = I_B(u + du, v + dv), \tag{2.1}$$

where spatial coordinates (u, v) in I_A are displaced by (du, dv) in I_B. However, as in surface registration, this is an inverse problem which is ill-posed due to several reasons. First of all, we seek a 2D vector for each coordinate, hence when working with scalar images it is an underdetermined problem. This is aggravated by the absence of image structures. Imagine an image of a gray square in a white background which is translated in another image. Though the actual transformation is the translation, we can also obtain the same image by rotations of 90° and translation, and in areas of constant intensity values numerous mappings are possible. In such case small noises can have disproportionate effect. For instance, two dots in some constant intensity region of images, but appear in different places in both images can cause wrongly a large estimated motion field. For these reasons, regularization is inevitably used in registration, and with more structure we have more chance of correct estimation of correspondences. By regularization we usually impose the resulting mapping functions are smooth, invertible and differentiable, i.e., diffeomorphisms, so that every point in one image has a corresponding point in the other and we are able to apply inverse of the mapping.

A plethora of registration algorithms, mostly for medical images, has been proposed [68, 69]. These algorithms are designed to handle specific issues and aim at very accurate registration without run time concerns, thus their computational burden becomes usually quite high for multimedia applications. Image registration is also a fundamental topic in video coding and many computer vision tasks. In many video applications, like coding, the accuracy is not as critical as in medical image registration. Moreover, in video at high frame rate, estimation of small displacements between

two consecutive video frames is sufficient. Therefore, more time efficient optical flow algorithms can be afforded in these applications. The optical flow constraint, which is the first order Taylor series approximation to the image constraint (Equation (2.1)), is used in these applications. This approximation does not allow large displacements, i.e., above one pixel, due to the omission of higher order terms.

3. EXPERIMENTATION MATERIAL AND METHODOLOGY

In this chapter we describe our experimentation material and evaluation methodology. To explore the use of 3D modality for the AU detection problem we prepared our own 3D database, since existing 3D expression databases [70, 71] are only limited to the six emotional expressions and FACS annotations do not exist. In the following section we introduce our Bosphorus 3D face database in detail and explain the two datasets from our database that we use in the experiments. In Section 3.2 we give some details on the Cohn-Kanade DFAT database that we also employed in our study. Finally, we describe our evaluation methodology in Section 3.3.

3.1. Bosphorus Database

Bosphorus database [1] is an extensive database of 3D faces which is designed to be used for 3D facial expressions and 3D face recognition under adversarial conditions. Some sample 3D faces are shown in Figure 3.1. This database is unique in three aspects:

(i) The facial expressions are composed of posed expressions including AU poses as well as the six basic emotions, and many actors/actresses are incorporated to obtain more realistic expression data. Certified FACS annotations, with complete intensity codes, are available for all the expressions.
(ii) A rich set of systematic head pose variations are available.
(iii) Different types of face occlusions are included.

3.1.1. Data Acquisition

3D faces and the companion 2D face images with a normal light camera were acquired with a structured light system [72] under good illumination conditions and without any background clutter. The acquisition setup is shown in Figure 3.2. The

[1] This database is available at http://bosphorus.ee.boun.edu.tr/

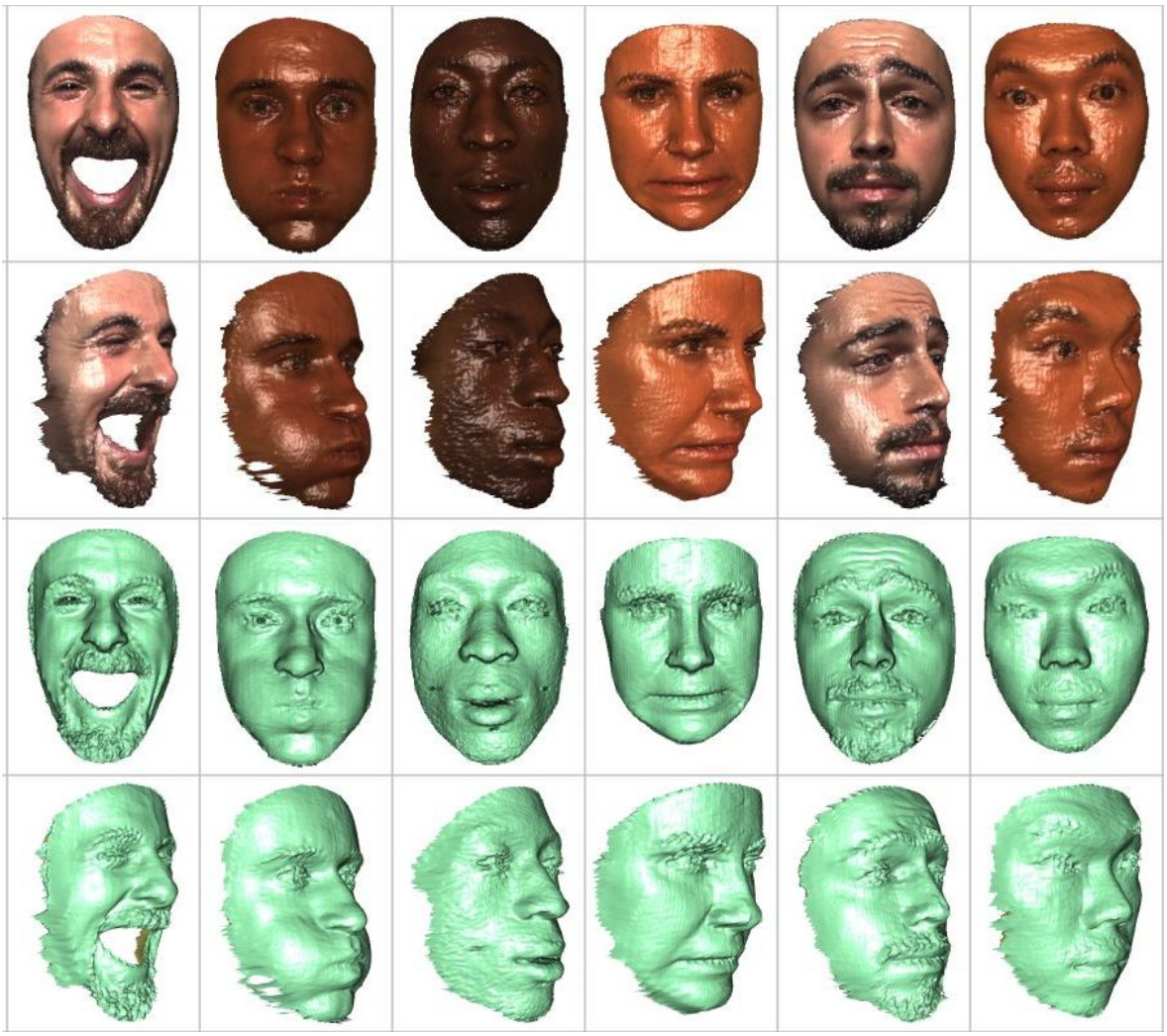


Figure 3.1. Sample 3D faces from Bosphorus database. The samples are shown with and without texture mapping by artificial lighting.

sensor resolution in x, y & z (depth) dimensions are 0.3mm, 0.3mm and 0.4mm respectively, and colour texture images are high resolution (1600x1200 pixels). The number of points on 3D faces varies roughly between 30K and 50K depending on the size of the face and due to the 1/6 down-sampling of the depth maps. The decimation is automatically made by the acquisition system on each dimension.

This database also includes 2D and 3D coordinates of 24 facial landmarks that are shown in Figure 3.3. Facial points were manually marked on 2D color images, provided that they are visible in the given scan. 3D coordinates were then calculated using the 3D-2D correspondences.

Figure 3.2. Acquisition setup composed of a 3D digitizer, a mirror, an LCD display, a 1000W halogen lamp, a seat (1.5m away from the digitizer), and yellow straps to indicate rotation angles (on the floor, just below the seat).

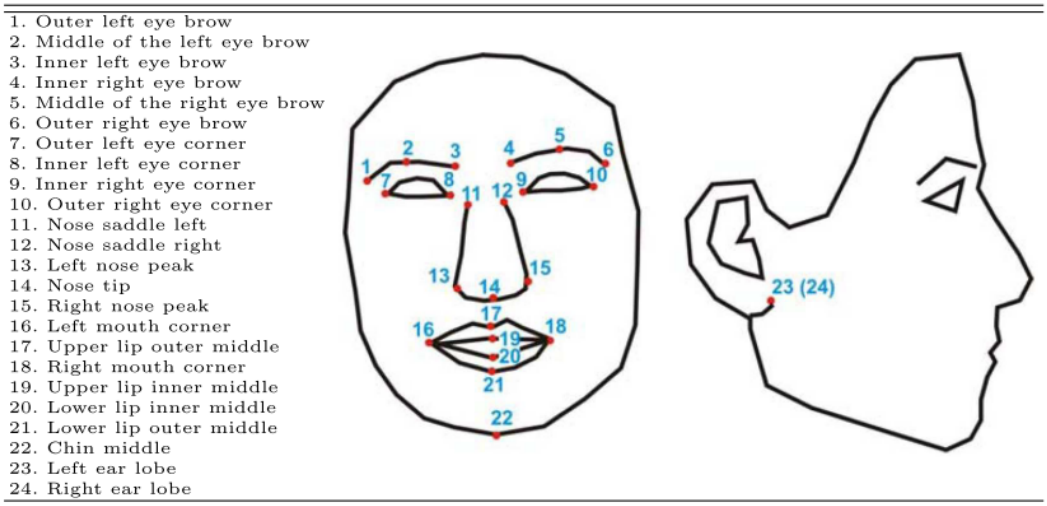


Figure 3.3. 24 labeled facial landmarks that are available in the Bosphorus database.

3.1.2. Content

Our database contains 105 subjects enacting a large repertoire of expressions, and displaying systematic head poses (13 fixed rotations including yaw, pitch and cross

rotations) and occlusions (beard, moustache, glasses, hand, hair, etc.), and the number of total face scans is 4666. The facial expressions were instructed by the experimenter and the ground-truth FACS annotations were obtained by one certified FACS coder. Some of the characteristics of the database are as follows:

- The majority of the subjects are aged between 25 and 35, mostly Caucasian.
- The cohort consists of 60 men and 45 women in total.
- 29 professional actors and actresses are employed for acting the expressions while the rest were recruited from students and staff.
- 35 men had beard/moustache (intense: 19, moderate: 16)
- 71 subjects were recorded with 54 different face scans (neutral, AU, universal emotion, systematic head poses and occlusions) while a minority of 34 subjects had 31 scans having fewer number of expressions.

3.1.3. AU Detection Datasets

We employ two different FACS dataset from the Bosphorus database for the experiments. As shown in Table 3.1 Bosphorus-DS1 AU set is a subset of Bosphorus-DS2 AU set.

- *Bosphorus-DS1.* This dataset is composed of 22 AUs and 1771 samples. In this dataset, each sample is coded with only one high intensity AU label, codes of co-occurring other AUs do not exist. However, the codes of DS1 dataset are based on instructed AUs, i.e., they are not certified annotations. Also, the AUs are instructed so that non-additive combinations with other AUs are not allowed, that is, every samples of an AU have the same deformation type. Thus DS1 becomes an easier dataset from AU recognition point of view since it only involves detection of monotype AUs. This dataset is used only in Chapter 6 to evaluate registration-based techniques using expression specific references.
- *Bosphorus-DS2.* This dataset is composed of 25 AUs and 2902 samples. This dataset has the ground-truth annotations of a certified FACS coder. The 25 AUs have been selected since other AUs do not have sufficient samples. Inclu-

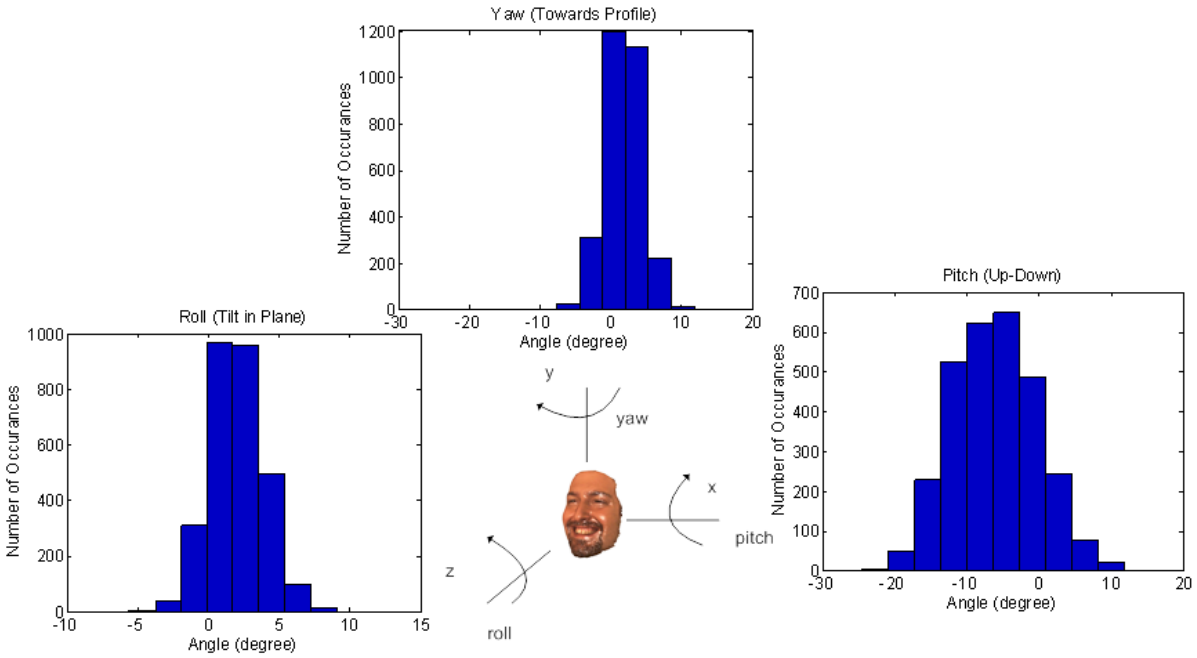


Figure 3.4. Head roll, yaw and pitch histograms in the Bosphorus-DS2 dataset.

sion of different types of variability, like various AU combinations, lower intensities, asymmetries, makes this dataset more challenging than DS1. The AUs of this dataset split as seven lower AUs and 18 upper facial AUs. This dataset is employed both in Chapter 4 and Chapter 6. In Figure 3.4 we also show the pitch, yaw and roll statistics of DS2 dataset estimated using 22 landmark points. Notice that, even with intended frontal acquisition, there exists non-negligible out-of-plane rotations especially in vertical direction.

3.2. The Cohn-Kanade DFAT Database

The Cohn-Kanade DFAT-504 [73] database is the best known AU coded facial expression database, and this was employed to compare 2D performances. The Cohn-Kanade DFAT-504 database contains digitized video clips of several (up to 23 per subject) instructed facial expressions captured from 97 university students. Each clip starts with a neutral face and ends with the apex frame of an expression. An example apex frame is shown in Figure 3.5. In contrast to Bosphorus database, this database does not contain facial hair, except a slight moustache for one of the subjects. Another important difference is that, intensity codes are available only for some of the samples

Table 3.1. Facial action units found in Bosphorus and Cohn Kanade datasets.

AU		Name	Bosphorus-DS1	Bosphorus-DS2	Cohn-Kanade
	0	Neutral	105	103	486
Upper Face AUs	**1**	Inner Brow Raise	46	232	143
	2	Outer Brow Raise	105	163	96
	4	Brow Lowerer	105	180	155
	5	Upper Lid Raise	-	154	78
	6	Cheek Raise	-	102	111
	7	Lids Tight	71	504	108
	43	Eye Closure	105	139	-
Lower Face AUs	**9**	Nose Wrinkle	99	112	50
	10	Upper Lip Raiser	70	96	12
	11	Nasolabial Furrow Deepener	-	19	33
	12	Lip Corner Puller	105	316	113
	14	Dimpler	69	76	-
	15	Lip Corner Depressor	54	68	74
	16	Lower Lip Depress	70	146	20
	17	Chin Raiser	71	134	157
	18	Lip Pucker	71	193	-
	20	Lip Stretch	63	56	69
	22	Lip Funneler	70	52	-
	23	Lip Tightener	71	63	43
	24	Lip Presser	70	155	43
	25	Lips Part	70	740	294
	26	Jaw Drop	71	220	38
	27	Mouth Stretch	105	182	76
	28	Lip Suck	105	43	-
	34	Puff	105	74	-
Total AU Samples/AUs			1771/22	4219/25	1713/19
Total Face Samples			1876	2902	972
Total Subjects			105	105	97

and AUs whereas we have complete intensity codes in the Bosphorus database.

We have used 972 neutral and apex expression frames extracted from the videos as test material for 19 AUs that had sufficient number of samples for experimentation. Note that the Cohn-Kanade AU set is a subset of the Bosphorus AU set in that every AU in the former set has a corresponding AU in the latter as shown in Table 3.1.

Figure 3.5. A sample face from Cohn-Kanade DFAT database.

3.3. Evaluation Methodology

3.3.1. Subject Independent Cross Validation

In our experimentations each AU is treated separately; as an example, we develop 25 detectors for Bosphorus-DS2 dataset, one for each AU. Any face image involving the target AU, alone or in combination with other co-existing AUs, is treated as a positive sample of that AU class, while all other images that do not involve the target AU are considered as negative samples. All the experiments are performed using 10-fold subject cross-validation, i.e., in each fold training subjects are never used in testing. However, the dataset partitioning problem is not straightforward since AUs are not distributed evenly among the subjects. We solve this problem by creating different subject partitions for each AU so that each fold becomes balanced with respect to its positive samples. We prefer balancing with respect to the positive samples because their count is much less than the negatives. The algorithm, which groups N subjects

into K folds of the cross validation, is given in Figure 3.6. After the subject folds are obtained, cross validation sample folds are created by placing the positive and negative samples of the subjects into the sample folds. The algorithm performs exhaustive search to obtain the most balanced partitioning by means of random histogram equalizations. Here, each input histogram bin corresponds to a subject storing the number of the positive samples, and each output histogram bin corresponds to the cross validation fold bin storing the positive sample count of one or more subjects.

ALGORITHM: Find the most balanced K subject sets of N subjects ($K \leq N$) according to the number of positive samples

- Create positive sample histogram of the subjects excluding the subjects without any positive samples

For each iteration

- Shuffle subject bins
- Create K bin histogram from N subject bins by histogram equalization
- Evaluate variance of the bin sizes
- Keep the current K bin histogram if its variance is smaller

- Add the unused subjects that have no positive samples to the bins
- Put positive and negative samples of each subject into respective sample bins

Figure 3.6. Sample set partitioning algorithm for subject independent cross validation

3.3.2. Intensity Separation

Every AU in the Bosphorus dataset has been annotated with its intensity level. This enables us to examine the detection performance as a function of the intensity. The five intensity levels in FACS range from slightest level A to strongest level E. In our experiments we had to exclude the AU samples with intensity level A, since even the expert annotators are not very sure of their scores at this level. The low intensity B level AUs are tested separately from the group of C, D, E levels to see the effect of weak intensity. Thus, we developed the AU detectors by training them with C, D, and

E intensity level AUs and testing them first with AUs having C, D or E intensity levels (strong set), and then with AUs at the B intensity level alone (weak set).

3.3.3. Performance Measurement

To compare the performance of different methods and compare the pros and cons of the 3D and 2D modalities, we express the detector performance in terms of Receiver Operating Characteristic (ROC) curves. ROC curves show hit rate (ratio of true positives to all positives) versus false alarm rate (ratio of false positives to all negatives) for varying thresholds. To have a single figure of merit that summarizes an ROC curve, Area under the Curve (AuC) measure is used; recall that AuC is equivalent to the theoretical maximum achievable correct rate in a binary classification problem. In this way, we avoid measures like correct recognition, hit and false alarm rates which can sometimes be quite misleading since they depend on the operation threshold. Disproportionate positive and negative populations, which is typical in AU detection experiments because of the very high number of negatives, is another cause of misinformation in case the evaluation is based on correct recognition rate alone. ROC analysis is performed over the combined cross validation sets, and thus one AuC measure is obtained for each AU. The overall performance score is found by weighted averaging of individual AU performances (AuC) where the weights correspond to the number of positive samples in each AU set.

To measure the performance of the intensity estimators we evaluate correlation coefficient between AU intensity estimates and the discrete ground-truth AU intensity levels. Again, the overall intensity estimation performances are calculated by weighted average according to the number of positive AU samples.

3.3.4. Statistical Significance

To show the significance of the findings, we estimate 95% confidence intervals ($2 \times$ standard error) over the AuC values of the test sets for each AU. We also perform paired t-test under 5% significance level in order to test whether the average AuC for

individual AUs is significantly changing with modality. The significance of the overall results is estimated by weighted sum of the distributions that are described by sample AuC means and standard errors, where weights are the number of positive samples in each AU set. By this way we obtain the 95% confidence intervals of the overall performance scores.

4. ACTION UNIT DETECTION ON 2D MAPS OF 3D FACIAL SURFACES

Previous work on 3D expression recognition process data in 3D space [44, 45, 46, 42, 47]. Although some of the authors have applied 2D techniques, either ASMs [49] or AAMs [48], their purpose was to track facial points on 2D luminance images in order to find 3D landmark coordinates. In other words, they are not pure 3D methods and do not exploit the whole facial surface information. We carry out the detection of AUs observed in 3D faces on 2D maps of these surfaces. There are three main reasons of implementing the expression recognition in 2D: First, 2D mapping enables time efficient expression recognition. Second, after resampling on image grid, analysis is not affected by differences in surface mesh resolution and topology. Third, once we map the surface geometry onto a 2D image we can utilize proven processing methods for 2D images. More importantly, this gives us the opportunity to compare the performance of 2D and 3D modalities under the same set of algorithms, so that their only difference is the way the data is captured, that is, whether with a light camera or with a depth camera.

The majority of 2D and all the 3D methods in the literature for expression recognition are based on model-driven techniques (landmark detection, AAMs, etc.). Prior information in these methods in the form of face models is advantageous since it simplifies the learning and analysis processes, they can help to better cope with adverse conditions such as pose variations, and require in general smaller amount of data for learning.

In contrast, we follow a data-driven approach instead of model-driven methods due to several reasons. Our first motivation is to focus on comparison of the two data modalities, and we do not want any bias coming from model design to influence these assessments. Assumptions made for modeling can unwittingly favor one of the modalities. Second, via a data-driven analysis we bypass the intermediate step of model

fitting, which usually involves detection and tracking of a high number of landmarks. These intermediate steps have to deal with robustness issues. Third, model construction is a tedious process, for instance statistical models require manual landmarking of a dataset, or parametric models require quite an amount of time and expertise for preparation. Also, some of the most popular model-based AU detection techniques like AAMs are influenced by the subjects' facial texture, and do not permit person-independent systems.

4.1. Overview

The flowcharts of our 2D AU and 3D AU detectors are depicted in Figure 4.1. We basically adapt the state-of-the-art technique of Bartlett et al. [25] who apply AdaBoost and Gabor analysis over 2D luminance images. However, as shown in Figure 4.1, we additionally compare various feature extraction and classification algorithms. The 2D AU detection method is straightforward, and involves the registration step followed by the feature extraction and classification stages (Figure 4.1(a)). In the case of 3D AU detection (Figure 4.1(b)), first, a 3D surface is reconstructed by fitting piecewise planar surfaces to the depth image. Also, texture mapping onto reconstructed surface is performed from the luminance data, if available. The reconstructed 3D surface is registered for pose normalization, whether manually based on landmarks or automatically using a rigid registration algorithm. Surface reconstruction, smoothing and 2D and 3D face registration techniques are explained in Section 4.2. The next step is 2D mapping onto the image plane which is described in Section 4.3. The surface geometry and the available surface luminance data are resampled on the image grid according to this mapping. Several representations of surface geometry are discussed in Section 4.4. Having converted the 3D face to a 2D image, we apply several feature extraction and classification techniques as detailed in Section 4.5 and Section 4.6 respectively. The multitude of methods permits us to achieve a more thorough comparison of the 3D and 2D data modalities. At the classification stage, there is the opportunity to fuse the information coming from luminance and that coming from geometry. Section 4.7 explains the fusion of the two modalities. Notice that, by canceling the luminance channel we are able to compare pure 3D versus pure 2D luminance, and on the other

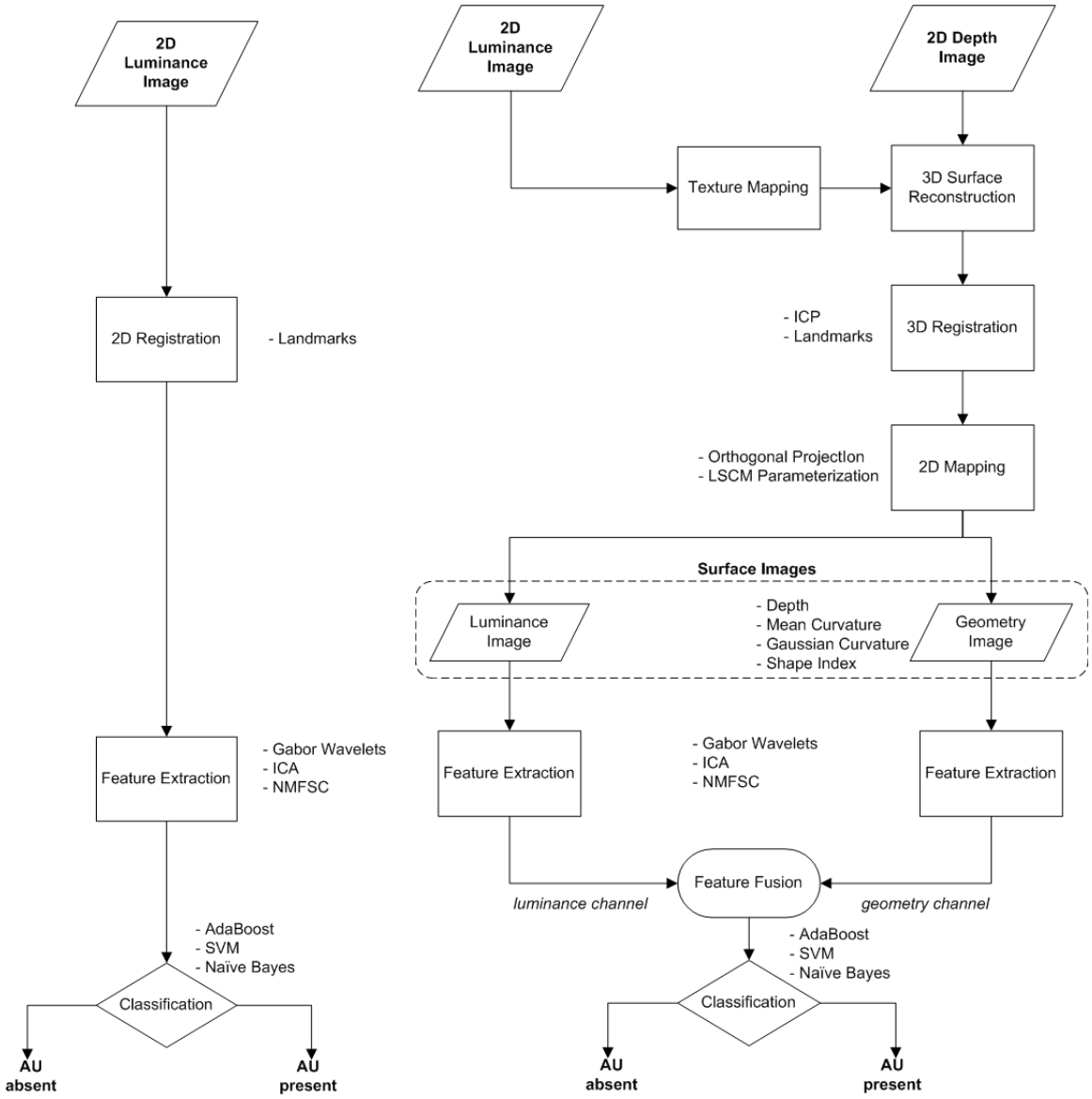


Figure 4.1. Flowcharts of (a) 2D and (a) 3D AU detectors where alternative techniques at each stage are indicated.

hand by canceling the geometry channel we are able to compare 2D luminance data versus 3D pose normalized luminance data, as will be made more clear in Section 4.7.

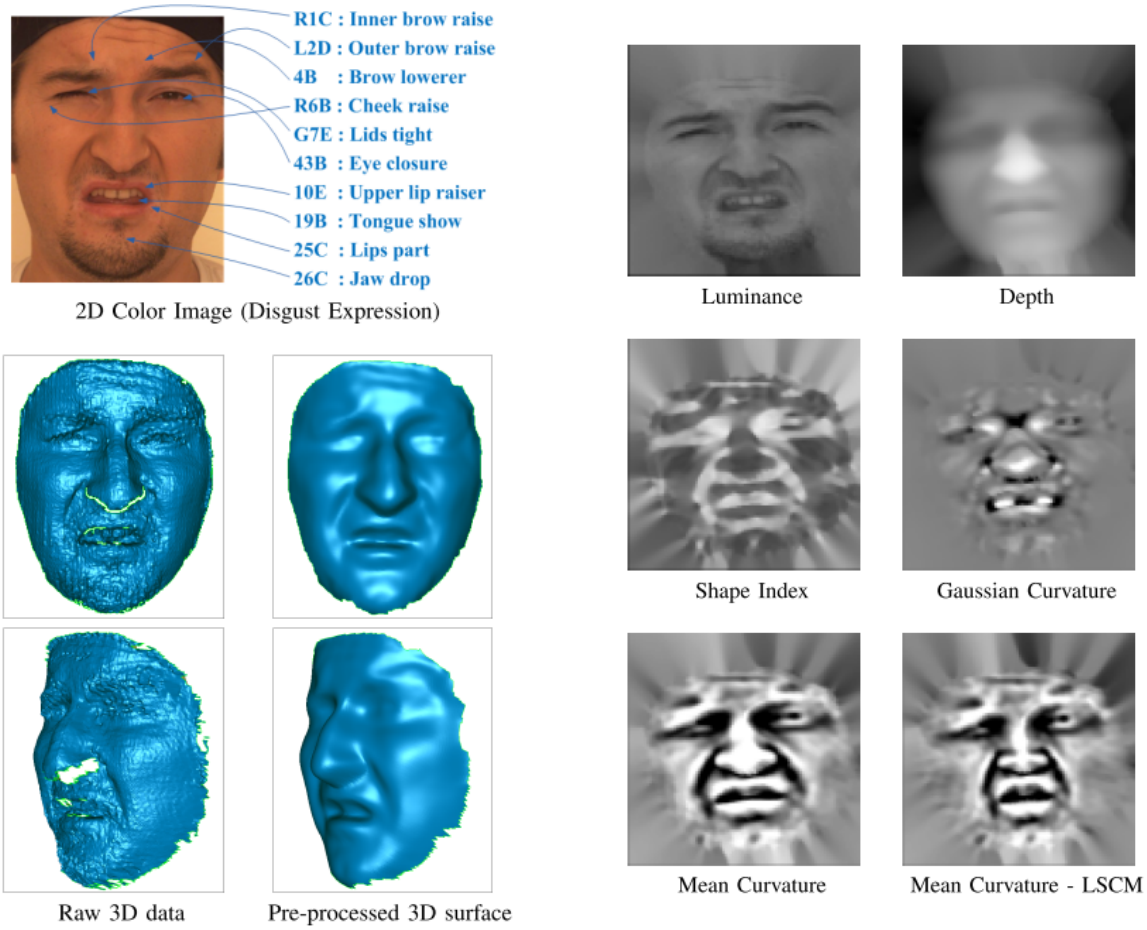
In addition to AU detectors, we have developed person-independent AU intensity estimators. The same feature extraction techniques are employed to predict the intensities, however, estimation is based on regression. We consider regression on SVM margins and on image features as described in Section 4.8.

Section 4.9 is devoted to experiments. The primary goal of these experiments is to investigate expression recognition capability enabled by 3D versus that of 2D data modality. Since our approach is based on 2D maps of 3D surfaces, we first find the best 2D representation among several alternatives, and then proceed to assess various feature extraction and classification algorithms. We also investigate the following five aspects of the problem: i) Fusion of 3D and 2D modalities to make use of any available complementary information for improved performance. ii) Whether moderately non-frontal 2D luminance images can benefit by the pose normalization information provided by the companion 3D facial surfaces. iii) Manual versus automatic pose alignment. iv) Assessment of the detection performance of Action Units at low intensities. v) Assessment of the AU intensity estimation performances.

4.2. Preprocessing

We need to execute certain pre-processing steps to condition the acquired range data of faces. The first step is surface reconstruction by fitting piecewise planar surfaces that results in a triangular wireframe structure. Also, from the luminance data, texture mapping onto reconstructed surface is performed. Since 3D acquisitions are typically noisy, several noise filtering steps are applied to remove the spikes, smooth the data and fill in the holes. Figure 4.2(a) shows pre-processing results of an input face that bears many lower and upper facial AUs.

The final preprocessing is done to register faces. The common approach in 2D facial image registration is to first find eye centers, then align the face accordingly using 2D rotation, translation and scaling. Though a 3D version of eye detection is possible, 3D data offer a more convenient way for registration without using any landmarks. We align faces by the Iterative Closest Point (ICP) [74] algorithm with respect to a reference 3D face model, which is a chosen neutral face in our experiments. The ICP algorithm that we employ matches the surface normals to find correspondences. However, when it comes to compare the methods, in order to preclude any misalignment effects due to automatic registration, 2D luminance images have been normalized using manually determined eye centers and 3D data are normalized with 3D landmarks. In all our



(a) (b)

Figure 4.2. Images of various representations of a sample FACS coded face expressing disgust emotion. (a) 2D color image, raw 3D data and pre-processed 3D surface. (b) Different types of surface images that can be generated for AU detection.

experiments, 2D registered faces (both for 2D and 3D modalities) are all resampled at 96×96 pixel resolution.

4.3. 2D Mapping of 3D Facial Surfaces

Given a 3D surface, various 2D images can be created according to both the mapping function and representation of the 3D geometry. Some of the 2D surface representations are illustrated in Figure 4.2(b) and further explained in the sequel. We apply two different techniques, which are based on projection and surface parameterization, to map 3D surfaces onto 2D planes. The mapping also involves separate scaling of horizontal and vertical dimensions to best fit to the square image analysis

domain. Scaling coefficients are determined by the following procedure. First, a mean bounding box is estimated from the 2D-mapped face meshes, which is then enlarged by one standard deviation from each side. The scaling and translation which maps this final box to 96×96 image grid are calculated, and finally applied to all 3D-to-2D mapped faces.

4.3.1. 3D-2D Mapping by Projection

The simplest means of 3D-2D mapping is projection, and orthogonal projection is a convenient way to analyze frontal faces. With multiple view systems, it may also be beneficial to employ environmental mapping techniques, like cylindrical or spherical projection. In our case we use orthogonal projection since our 3D digitizer is a single camera system and our purpose is to make comparison with conventional 2D single camera images.

4.3.2. 3D-2D Mapping by Surface Parameterization

A drawback of the projection methods is that the resulting maps may not be bijective due to the occurrence of many-to-one mapping instances. This happens in regions where projection direction and surface normals are almost perpendicular to each other, and as a consequence surfaces cannot be represented adequately. For frontal orthogonal projection, lossy mapping occur on the nose wings and on parts of the cheeks receding to the background.

We can utilize mesh parameterization techniques to ensure bijective mapping. Several mesh parameterization techniques have been developed in computer graphics for texture and detail maps of plausible virtual models [66], for face recognition [75] and for tracking of facial surfaces [64]. We have considered an angle-preserving parameterization method, known as Least Squares Conformal Mapping (LSCM) [76]. Angle-preserving property of a mapping yields consistent shapes in 2D images, i.e., topology of the facial features do not differ from one mapping to another, and it is achieved by constraining surface boundaries to fixed coordinates on the plane. The

LSCM method is solved linearly, hence it is a much faster technique than many other parameterizations that require nonlinear optimization. However, the limitations of the LSCM method is that the surfaces must possess disk topology (0-genus mesh with borders). Since this topology constraint can be violated by acquisition noise, i.e. holes can occur, or small nose holes and mouth cavity, we fill in the holes by surface interpolation in the preprocessing stage. In Figure 4.2(b), the orthogonally mapped and conformally (LSCM) mapped mean curvature images are displayed for comparison. In these figures we observe the two most prominent effects of LSCM on the facial surfaces: first, nose wings are better represented and the upper part of nose cone is narrowed to generate this extra surface area; second, the face center is narrowed to allow room for cheek regions. When the entire face surface is represented by conformal mapping on the same 2D domain, some regions need to contract to make room for others, and this is the main trade-off between orthogonal projection and conformal mapping.

4.4. Representations of Surface Geometry

Using any one of the 2D mapping technique the 3D face surface data (geometric or texture) are re-sampled on a regular image grid. This is achieved very rapidly, even for high resolution meshes, by utilizing graphics hardware. As seen in Figure 4.2(b), the final step is the extrapolation of the mapped values outside the 2D domain of the surface. This is needed to smooth the abrupt passage from the delineated region of support of the 3D face and its background. A satisfactory extrapolation is obtained by an efficient image in-painting algorithm [77]. We consider four types of geometry data since representations of surface information other than 3D coordinates can be more suitable for analysis of deformations. Figure 4.2(b) displays these data types in the form of the 2D mapped images of a face surface.

(i) *Depth.* Using orthogonal mapping, we can analyze depth values which are read off directly from 3D surface coordinates.

(ii) *Mean Curvature (H).* Mean curvature at a surface point is the mean of the principal curvatures, i.e., the maximal k_1 and minimal k_2 curvatures, and it is an extrinsic measure of curvature. Principal curvatures are extracted by the local

analysis of the differential structure of the surface.

(iii) *Gaussian Curvature (K).* Gaussian curvature is the product of the principal curvatures, and it is an intrinsic measure of curvature.

(iv) *Shape Index (S).* Shape index has been developed [78] to measure the local shape by a single number in a continuous range, but at the expense of curvedness information. In contrast to mean and Gaussian curvature at a point, shape index directly represents the local shape.

These curvature representations are calculated from the maximal k_1 and minimal k_2 principal curvatures as follows:

$$\begin{aligned} H &= \frac{k_1 + k_2}{2} \\ K &= k_1 \times k_2 \\ S &= \frac{1}{2} - \frac{1}{\pi} \arctan \frac{k_1 + k_2}{k_1 - k_2} \end{aligned} \tag{4.1}$$

4.4.1. Discussion on Curvature Representation

Actually we can categorize surface features according to the differentiability conditions of the surfaces. According to this categorization curvatures are second-order features since they require second-order differentiability. Then depth is a zero-order feature since differentiability is not a required condition. As a first-order feature we can mention surface normal directions. Higher order differentiability conditions can be seen as a disadvantage since we need to fulfill these conditions. In practice curvature estimations are considerably disturbed by the acquisition noise which is unavoidable by many current 3D sensing systems. Moreover, facial surfaces are not differentiable everywhere and existence of facial hair makes the situation worse. Nevertheless, as a remedy to these problems surface smoothing is performed in a pre-processing stage.

However, for expression analysis, there are several inherent advantages of curvature representations over other surface features. First of all, principle curvatures measure the amount of local surface bending in different directions. In this respect, they are good candidates of surface features for the analysis of facial deformations.

Besides, since principle curvatures are invariant to translation and rotation, they are not affected by the head pose variations.

An alternative feature would be the use of surface normal directions since they are translation invariant. However, they are dependent on rotations, which means that they are still susceptible to pose variations. Moreover, 3D normal vectors make the analysis problem more complicated due to the three times higher dimensional data. As an advantage of normals compared to curvatures we can mention the loss of 3D spatial deformation direction when using curvatures; principle curvatures only keep concave-convex direction.

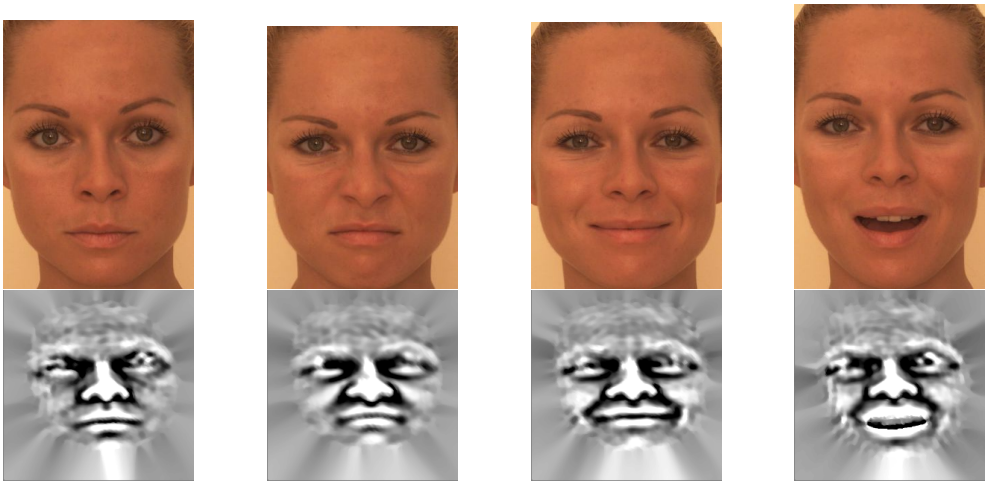
In conclusion we can say that curvature representation is advantageous since it provides surface deformation information in a compact form without being disturbed by the translation and rotation related variations. However, the price is the need for careful preprocessing for conditioning the data and loss of 3D spatial direction of deformations.

4.4.2. Facial Variations Portrayed on Curvature Fields

We can group variations on the faces into two groups, namely as expression and non-expression related variations. Expressions give rise to facial motion, in terms of both deformations and head pose changes. In this context, we can generalize the expression category to include facial changes due to speech articulations, i.e., facial motions called visemes, in order to group temporal surface deformations due to facial muscle activations into one category. Variations not related to any expression can handicap or confound an expression analysis task.

Unlike the luminance data, many characteristics of surface deformations can clearly appear on the mean curvature field without the confounding effects of non-uniform lighting or facial albedo variations provided facial hair is moderate. Variations due to identity and expressions coexist within curvature data. For instance, curvedness of the facial parts, e.g., especially in nose, lips or eye holes varies among the people; sim-

ilarly, expression related transient deformations, such as bulges, furrows and curvedness of lips are well represented. In Figure 4.3 we see such transient features much clearly in the curvature data. Color, 3D surface and surface mean curvature images of various types of expression related deformations are also shown in Figure 4.9.



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

Figure 4.3. Different transient features are shown on the color and curvature field images. Faces of (a) neutral, (b) nasolabial deepening and accompanied bulge on the cheeks, (c) furrows around the mouth, (d) furrows and open-mouth are shown.

The major non-expression related variations originate from differences due to subject identity. While working with 3D surface measurements, factors like pose, illumination, shading and facial hair are minor problems in contrast to luminance images. A case in point is thick beards and moustaches: though many facial hair variations would have insignificant effect on the curvature maps since that amount of detail would be smoothed out and remain only in the surface texture data. Strong facial hair may cause confounding effects for some expression related variations, especially for the subtler ones. Figure 4.4 shows the effect of different levels of facial hair on the curvature fields. We see that if the thickness is not much, the curvature field is not affected significantly. Influence of thick moustache on the curvature values is clearly seen on the bottom-right face.

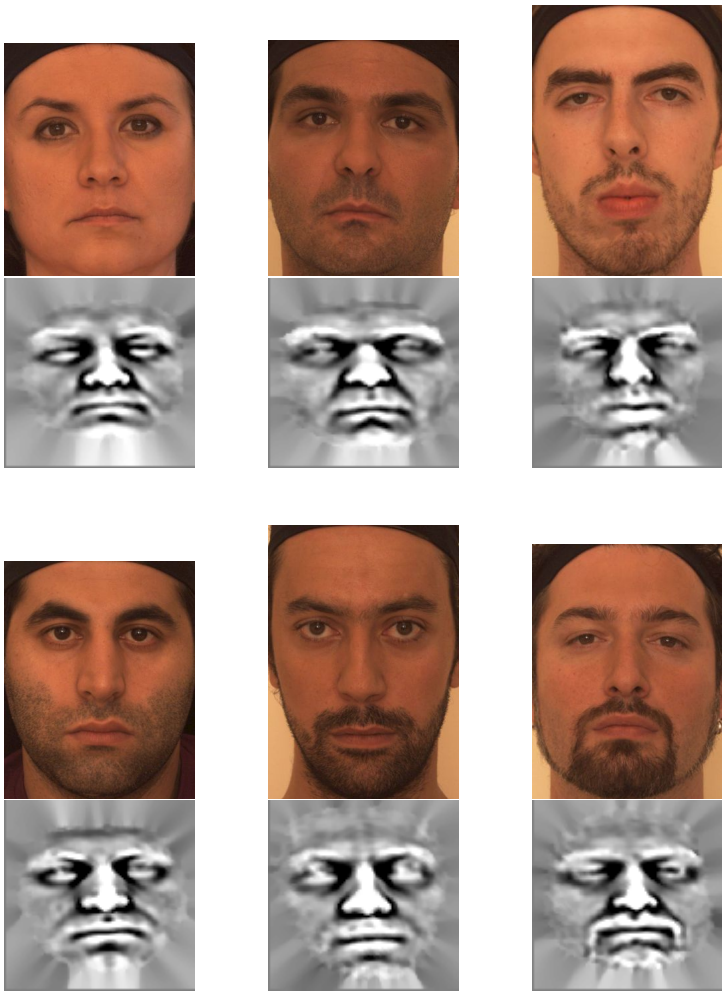


Figure 4.4. Different levels of facial hair. If the thickness is not much, curvature fields are not affected significantly. Influence of thick moustache on the curvature values is clearly seen on the bottom-right face.

4.5. Image Features

We can classify feature extraction techniques as model-driven and data-driven techniques. Notice that in this feature extraction context, model-driven does not mean use of face models, but it implies a predetermined form of feature measurement rather than applying a training procedure to extract features. In other words, we exclusively perform data-driven expression analysis with both data-driven and model-driven feature extraction. In our study we employed one model-driven feature extraction technique, namely, Gabor Wavelets, and two data-driven feature extraction methods, namely, Independent Component Analysis (ICA) and Non-Negative Matrix Factorization with Sparseness Constraints (NMFSC). All three methods perform local analysis,

which is necessary since AUs happen locally. Our choice for Gabor Wavelets stems from a recent study [40] on smile detection where it was shown that Gabor Wavelets were one of the best among other model-driven features such as box filters, edge orientation histograms, and local binary patterns. Among data-driven techniques, i.e., methods that construct subspaces from data and select features, ICA has been used for classifying facial actions [22]. In addition, we employ a more recent method, NMFSC, for our experimentation. We believe that data-driven feature extraction methods can play an important role when comparing data modalities as they may reveal better their intrinsic nature. Finally, we also evaluate use of difference images by subtracting the neutral image of a subject in order to reduce ambiguities due to subject identity. The details of these three feature types are as follows:

- *Gabor Wavelets.* Gabor wavelets have a proven history of success in facial analysis tasks [79]. Since the effective support of a Gabor basis is limited and inversely proportional to its frequency, the resulting analysis remains local.
  In our implementation of Gabor analysis we have used eight directions and nine scales so that their Gabor wavelengths vary in the range of 2 to 32 pixels in half octave intervals, as in [25]. Although the resulting feature vector has $9 \times 8 \times 96 \times 96 = 663,552$ components, not all of them are informative and in fact only a very small portion is selected as described in Section 4.6. This method together with AdaBoost feature selection has been first applied by Bartlett et al. [25] and is the state of the art for 2D AU detection.
- *Independent Component Analysis (ICA).* can be realized in two different architectures, called ICA1 and ICA2 [80]. While subspaces constructed by ICA2 are convenient for holistic analysis, since it finds global basis images, ICA1 architecture generates basis images that are local and sparse. Therefore, ICA1 is employed for local AU detection. ICA has the following model: $\mathbf{X} = \mathbf{AS}$. In ICA1 setting the image dataset is organized as a matrix $\mathbf{X}$, where each row is a sample image of a face, e.g., illuminance or surface geometry image. $\mathbf{A}$ is the mixing matrix and $\mathbf{S}$ is the matrix, whose rows represent the independent source images. Thus, ICA1 models images as a linear mixture of independent source images, and rows of $\mathbf{A}$, i.e., mixing coefficients are used as our features.

In Donato et al. [22] ICA has achieved the same AU recognition performance as the Gabor analysis. These authors estimated 200 independent components separately on six lower and six upper facial AU image. Also, they used only difference images obtained by subtracting the neutral images of the subject. In our work, we have tested both actual images and difference images. The ICA algorithm we apply maximizes the mutual information as in [22] which results in 200 independent components.

- *Non-negative Matrix Factorization with Sparseness Constraints (NMFSC).* Non-negative Matrix Factorization (NMF) is a matrix factorization technique capable of learning parts of images which can be combined additively to reconstruct the image. Consider the image dataset organized as a matrix $\mathbf{V}$, where each column is a sample image of a face, e.g., illuminance or surface geometry image. The matrix $\mathbf{V}$ is factorized into two non-negative factors $\mathbf{W}$ and $\mathbf{H}$ so that $\mathbf{V} \approx \mathbf{WH}$. The first factor $\mathbf{W}$ contains the basis images and the components of $\mathbf{H}$ correspond to combining coefficients. Nevertheless, it does not guarantee extraction of local features since sparseness is actually not the goal of NMF solution. We have indeed observed some non-local basis images resulting from our dataset and the resulting classification performance was considerably low. The local NMF method has already been proposed in [81] and applied to emotional expression recognition [82]. However, we experienced convergence problems in our experiments, both for 2D and 3D data sets, i.e., most of the time we obtained quite noisy basis vectors. A more recent extension of NMF, called Non-negative Matrix Factorization with Sparseness Constraints (NMFSC), imposes sparseness constraints on the $\mathbf{W}$ and $\mathbf{H}$ matrices [83]. A very nice property of this method is that it allows explicit control of the sparseness level in the range $[0, 1]$; moreover, resulting representations are observed not to be very sensitive to the chosen sparseness level.

$$sparseness(\mathbf{x}) = \frac{\sqrt{n} - (\sum |x_i|)/\sqrt{\sum x_i^2}}{\sqrt{n} - 1}. \tag{4.2}$$

By setting the sparseness level of the basis matrix $\mathbf{W}$ to 0.8 and that of the combiner matrix $\mathbf{H}$ to 0, we were able to obtain good local bases, and the detection performance was much higher as compared to NMF. Similar to ICA, we estimated

200 NMFSC components and tested them on both static and difference images.

4.6. Classification Methods

To classify and detect AUs we have used AdaBoost, SVMs (Support Vector Machines) and Normal Bayes classifiers. AdaBoost and SVMs are two of the strongest and common discriminative classification algorithms applied on various computer vision problems. In particular, AdaBoost and linear-SVM have been successfully used to discriminate AUs from luminance images [25]. In addition to linear-SVM, we also tested RBF-SVM. Moreover, to find out the potential of generative classifiers in AU detection, we have applied several Normal Bayes classifiers. Notice that for each AU we train a separate classifier. The details are as follows:

- *AdaBoost.* AdaBoost is a powerful machine learning algorithm to construct a strong classifier by adaptively combining many weak classifiers, each one being weighted according to its performance over the training set. In an alternative scheme, AdaBoost is used as an effective feature selector to find out the most discriminative and parsimonious feature set for the classification task, and then a separate classifier like SVM can be trained on these features. In this work, we resort to the nearest mean classifier in the role of a weak classifier. For Gabor analysis 200 features are selected, a separate set for each AU. Thus, out of 663,552 Gabor magnitudes, AdaBoost selects 200 features.
- *Support Vector Machines.* In our work we evaluate the performance of SVM with both linear and RBF kernels. We have observed that the performance of SVM was very sensitive to its chosen parameters, i.e., capacity and RBF kernel spread. Therefore, we searched for the optimal hyper-parameters over the training set of each AU by 10 fold cross-validation. The training procedure where also the optimal SVM hyper-parameters are learnt is as follows: Given an adequately wide search space for the hyper-parameters, first for each training set of each AU, a 10 fold cross-validation, where folds are balanced according to the positive samples, is performed over a training set, and thus the optimal hyper-parameters are obtained. Then, SVM is run on the entire training set using the best parameters

among the 10 cross-validation folds. These two steps are repeated for each AU. Though this makes training time quite lengthy when a large search space size is involved, we believe it was necessary for a thorough assessment of SVMs since the use of fixed parameters, as often done in the literature, has the risk of yielding suboptimal detectors.

- *Normal Bayes Classifiers.* Although discriminative classifiers like AdaBoost, SVMs or neural networks are proven methods in facial expression analysis, their generalization capability may be poor when the number of training samples is limited and consequently not all possible variations are adequately represented. Also, inaccuracies of the ground-truth labeling (label noise) may have more severe impact in discriminative models. Both of these issues are of real concern in actual AU detection since not all AUs are richly represented in the databases under study, and FACS ground-truths are subject to human annotation inconsistencies. Therefore, we have included a generative classifier. Actually, since we carry out the classification on the most discriminative features already selected by AdaBoost, both discriminative and generative characteristics can be said to be employed jointly. Four types of Bayes classifiers are tested for this purpose, where features are assumed to be Gaussian:
  (i) *Quadratic Normal Classifier.* Diagonal covariance matrices are estimated for positive and negative samples separately.
  (ii) *Simplest Quadratic Classifier.* A single global variance is estimated for all features and for each of the two class.
  (iii) *Naïve Bayes.* Linear Normal classifier with diagonal covariance where a single diagonal covariance matrix for the two classes is estimated.
  (iv) *Nearest Mean.* Simplest linear classifier.

4.7. Fusion of 3D and 2D Modalities

We address fusion of the 3D geometry and 2D luminance modalities from two different points of view. The first question is whether or not the two modalities contain complementary information useful for AU detection. We expect complementary information because factors like skin pigmentation and facial hair change the facial albedo,

factors that cannot be captured in 3D modality. Most of the albedo variations occur on lips, eyes and eyebrows, facial features that have high importance in recognition of expressions. Moreover, acquisition noise and the consequent smoothing operations on 3D may cause loss of some details like wrinkles. These concerns motivate us to investigate the fusion of the two modalities. For this purpose we apply AdaBoost feature selection on the combined set of Gabor features of geometry and luminance data.

The second question we address is whether it is possible to make use of 3D information to obtain better luminance data. Since one of the advantages of 3D is more accurate pose normalization, we can adaptively resample luminance images in order to get rid of small out-of-plane rotations that inevitably occur, even when acquisitions were intended to be frontal. We therefore generate a sampling grid to frontalize the luminance image of the face from the pose learned using the 3D facial surface meshes. Here we profit from the correspondences of range and luminance data provided by the software of the acquisition device. Hereafter we call these 3D-aided luminance images as 3D luminance since its domain becomes 3D surface after the texture mapping stage (Figure 4.1(b)). A sample 2D mapped image of 3D luminance is shown in Figure 4.2(b) together with other geometry map images.

4.8. Action Unit Intensity Estimation

We formulate the estimation of intensity levels as a regression problem. The dependent variable is the intensity in ordinal scale varying from one to five. The explanatory variables are SVM scores or certain image features. Since the output of the regressor is continuous, the outputs are quantized into five discrete intensity levels. Note that our intensity estimators work completely in person-independent manner.

4.8.1. Regression on SVM Margins

It was shown in [25] that distances to SVM margins (separating hyperplanes) are correlated with intensity levels of AUs. This indicates that AU detector decision scores can also be used to estimate AU intensities. Figure 4.5 shows the scatter of SVM

decision scores for 2D and 3D modalities as box-and-whisker plots. In these plots the box incorporates lower to upper quartile of score values, while whiskers are bounded by the extreme values within 1.5 times the interquartile range from the ends of the boxes. In Figure 4.5, the number of samples are written next to the letter symbol of the intensity level, e.g., B: 153 means that there are 153 instances of that AU at intensity B. As expected higher AU intensities correspond to bigger SVM scores. These plots show also that the upward trend varies from AU to AU, and more importantly, there are substantial overlaps between some adjacent intensity grades. Note that the medians of the distributions do differ significantly at the 5% significance level if their notches[2] do not overlap. One explanation for these overlaps is that the SVM algorithm was designed to detect an AU, but not necessarily to estimate its intensity. Furthermore, whatever technique is employed substantial overlaps is perhaps unavoidable due to the fact that a strict separation between intensities is difficult to achieve since FACS does not define a quantitative measure between levels. Finally, in person-independent intensity estimation, one is confronted with more of variability since different subjects can enact AUs differently and facial surface and texture vary from subject to subject. These factors make the estimation problem more challenging.

When we observe score distributions as a function of data modality (i.e., 2D vs. 3D), we can see some of the difficulties of the problem. For the scatter of AU 12 - Lip Corner Puller (first row of Figure 4.5), the trends are quite similar for 2D and 3D. In the case of AU 5 - Upper Lid Raise, the overlap of the scores of 3D are quite higher than those of 2D, and surprisingly D-level scores do not follow the upward trend. On the other hand, in AU 22 - Lip Funneler, the opposite happens. These shortcomings will be partly compensated for when we resort to fusion of 2D and 3D in Section 4.9.9.

The AU intensity levels, $f(r)$, are estimated using logistic regression on SVM scores r:

$$f(r) = \frac{1}{1 + e^{-(a+br)}} \tag{4.3}$$

[2]Notches are the first quartile-to-median and median-to-third quartile ranges

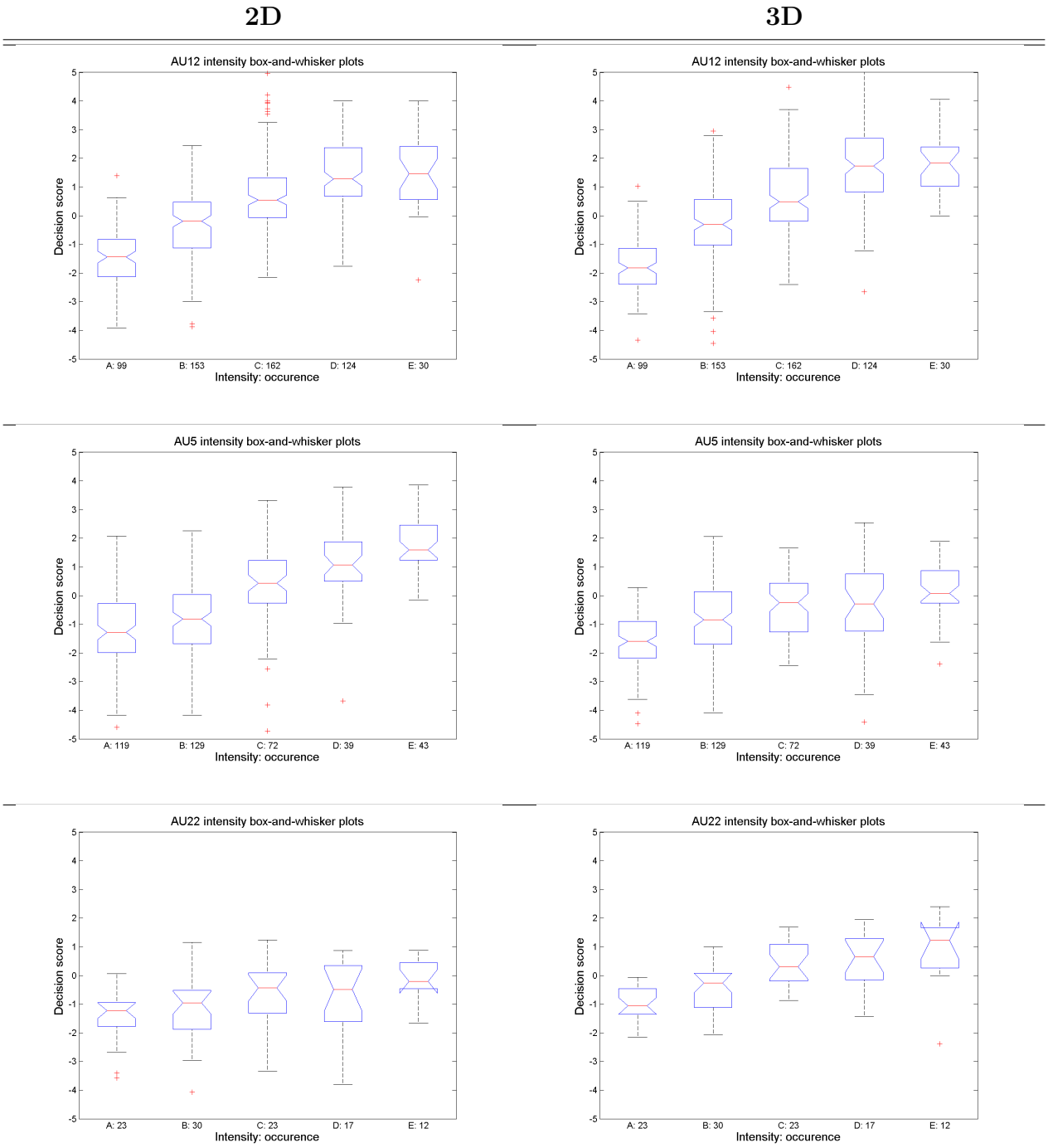


Figure 4.5. Distributions of decision scores (RBF-SVM margins) of three AUs for 2D and 3D data modalities shown as box-and-whisker plots (central mark: median, box: interquartile range, whiskers: extreme values, '+': outlier).

4.8.2. Regression on Image Features

Although the scores, i.e., the distance to the hyperplanes in SVMs designed for AUs imply stronger evidence for these AUs, proportionality of SVM scores to intensities is not guaranteed since the support vectors were chosen for the classification task but not for intensity level estimation. We therefore consider an alternative regression in the feature space of selected Gabor wavelet magnitudes of luminance or of mean curvature field.

This regression problem is not straightforward since we have a high number of explanatory variables (features), and the dependent variable (annotator's scores) are noisy, as there is considerable overlap between intensity grades as discussed in Section 4.8.1. Hence, we apply SVM regression based on Vapnik's ε-insensitive loss function [84]. ε-SVM regression is appropriate because, first, high dimensionality of the input space is not an issue for SVMs, and second, the ε-insensitive loss function is robust and generates a smooth mapping.

Another consideration is the non-linearities between the scale of evidence and intensity levels, as depicted in Figure 2.2. This relationships points out the possible benefits of non-linear modeling. Notice that, there are also other sources of non-linearities, such as combinations of AUs. SVMs are also great tools for effectively learning various types of complex mappings by means of kernels. The SVM regression function has the form:

$$f(\mathbf{x}) = \sum_i \alpha_i k(\mathbf{x}_i, \mathbf{x}) + b \tag{4.4}$$

where $\mathbf{x}$ is the feature vector, $k(\mathbf{x}_i, \mathbf{x})$ the kernel function, $f(\mathbf{x})$ is the predicted intensity level and $\mathbf{x}_i$ are the support vectors. Recall that $\mathbf{x}$ represents the vector of 200 Gabor features that were also used for AU detection.

In our study we investigated both linear-SVM and SVM with nonlinear kernels of the Gaussian RBF variety. Advantage of RBF is its ability in handling various types of

non-linearities despite having single spread parameter. Depending on this parameter and SVM capacity many non-linearities can be captured. Therefore, we optimize these two hyper-parameters also together with insensitivity range ([-ε,ε]) for each AU by performing cross-validation over the training sets.

4.9. Experimental Results and Discussions

4.9.1. Best 2D Representation for 3D Data

We first explore the best 2D representation for 3D data. Table 4.1 lists the results, that is, ROC analysis as described in Section 3.3.3, for various representations under different classifiers. As described in previous sections, the surface geometry data is acquired in 3D, and then mapped onto 2D either with orthogonal projection or via conformal mapping technique. Finally, 200 Gabor magnitude features are extracted via AdaBoost, separately for each AU. In the left section of this table we see the performances of the geometry data types mapped by orthogonal projection. The first relevant outcome of this experiment is the superiority of the curvature related geometry information compared to depth information. Though the differences are small, the mean curvature seems to be the best among the three curvature types, and this holds consistently for all four classifiers. On the right section of the table, we compare the LSCM mapping and orthogonal projection of the mean curvature values. As a second fact, we can conclude that the performances of the mean curvature under orthogonal mapping and conformal mapping are on a par. A third observation from Table 4.1, where four classifiers under four 2D data types are compared, is that SVM with RBF kernel and Naïve Bayes obtain very similar scores, and they outperform AdaBoost and Linear-SVM.

4.9.2. Best Image Features for 3D and 2D Modalities

Figure 4.6 shows the detection results obtained with RBF-SVM for three types of features extracted from 2D images. The 3D data is represented by the mean curvature field and 2D data by the luminance field (both sampled at 96×96 pixels), from which we

Table 4.1. Comparisons of 2D representations for 3D data with average AuC values and 95% confidence interval estimates. All of the classifiers use 200 Gabor magnitude features that are selected by AdaBoost for each AU.

	Orthogonal Projection				***LSCM***
Classifier	**Depth**	**Shape Index**	**Gaussian Curv.**	**Mean Curv.**	**Mean Curv.**
AdaBoost	92.2 ± 0.4	93.7 ± 0.4	94.9 ± 0.4	94.8 ± 0.4	94.5 ± 0.4
Linear-SVM	92.4 ± 0.5	94.5 ± 0.4	93.8 ± 0.5	95.0 ± 0.4	94.8 ± 0.4
RBF-SVM	**93.2 ± 0.4**	94.7 ± 0.4	**95.0 ± 0.4**	**95.5 ± 0.4**	**95.4 ± 0.4**
Naïve Bayes	93.1 ± 0.5	**95.1 ± 0.4**	94.4 ± 0.5	95.3 ± 0.5	**95.4 ± 0.4**

extract 200, respectively, ICA, NMFSC, and Gabor features. For any type of extracted feature 3D modality is superior, and even the worst 3D result is higher than the best 2D result.

Among the feature types Gabor magnitude is the best, both for 2D and 3D, and independent of classification method (among all the classifiers that we employed). This is in contrast to the work of Donato et al. [22] who reported recognition rates that were similar between Gabor and ICA. This may be due to the differences in the databases since theirs is simpler than ours (20 vs. 105 subjects, 12 vs. 25 AUs) and they perform separate analyses on isolated lower and upper AU image sets; in contrast ours contain many AU combinations in addition to a higher number of AUs and also we do not separate the analyses of lower and upper part AUs. We have not found a clear-cut difference between the two sparse subspace methods, that is, ICA and NMFSC.

When we analyze the performance loss in 2D vis-à-vis 3D modality, the performance gap is much wider (6% drop) for the subspace methods (ICA and NMFSC) as compared to the Gabor method (2% drop) to the disadvantage of 2D modality (Figure 4.6). A possible explanation is that, since subspace bases are learned from data in contrast to fixed Gabor wavelets, subspace techniques may be more susceptible to variations unrelated to expressions found in 2D luminance images. These variations can be caused by differences in skin tone or facial hair, by shadows, or by out-of-plane rotation, which apparently impact 3D less. Figure 4.6 also shows the performance attained with difference images. Taking the difference of images with respect to their neutral

provides small improvement in the curvature feature while it does not gain anything in 2D. Finally, among all the classification methods, RBF-SVM is almost always the best classifier.

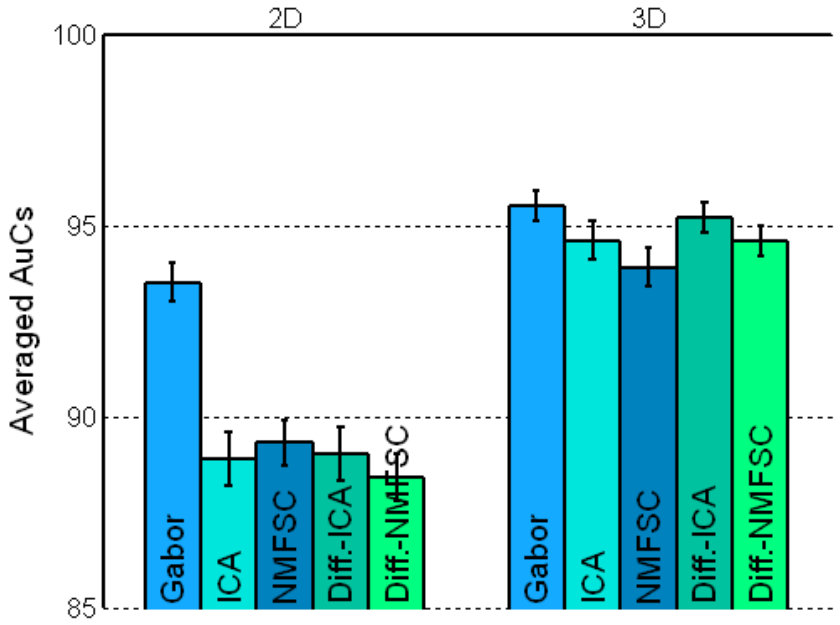


Figure 4.6. Average AuC values and 95% confidence interval estimates of AU detectors for 2D (luminance) and 3D data (mean curvature) under different feature transforms (200 features and RBF-SVM classifiers). "Diff.": neutral face substraction.

4.9.3. State-of-the-Art 2D AU Detection

The AU detection performances from 2D luminance images of the Bosphorus and Cohn-Kanade databases are given in the first two rows of Table 4.2. These two databases yield very similar detection performances: in fact, the score of the best classifier, RBF-SVM, is 93.7% AuC on the Cohn-Kanade and 93.5% on the Bosphorus database. The fact that the confidence intervals are slightly larger in the Cohn-Kanade can be explained with the smaller number of samples (1713 vs. 4219) in that database.

In the bar chart in Figure 4.7, we show the detection performance of individual AUs in these two databases. The scores of individual databases are superposed on each bar. The number of total positive samples is inscribed in the AU bars, the bottom figure

Table 4.2. Average AuC values and 95% confidence interval estimates of AU detectors for 2D and 3D data. All of the classifiers use 200 Gabor features that are selected by AdaBoost.

Data	**AdaBoost**	**Linear-SVM**	**RBF-SVM**	**Naïve Bayes**
Cohn-Kanade - Pose alignment by landmarks				
2D Lum.	92.1 ± 0.8	92.4 ± 0.8	**93.7 ± 0.7**	93.5 ± 0.7
Bosphorus - Pose alignment by landmarks				
2D Lum.	92.2 ± 0.5	92.4 ± 0.5	**93.5 ± 0.5**	91.3 ± 0.6
3D Lum.	93.0 ± 0.5	93.5 ± 0.5	**94.7 ± 0.4**	93.0 ± 0.5
3D Geom.	94.8 ± 0.4	95.0 ± 0.4	**95.5 ± 0.4**	95.3 ± 0.5
Fusion	96.1 ± 0.3	95.9 ± 0.3	**96.6 ± 0.3**	**96.6 ± 0.3**
Bosphorus - Pose alignment by ICP				
3D Geom.	93.6 ± 0.5	94.0 ± 0.5	**94.8 ± 0.4**	94.6 ± 0.5

for Cohn-Kanade and the top figure for Bosphorus dataset. A zero figure signifies that that AU does not take place in that database, e.g., AU 14 is not extant in Cohn-Kanade. The bottom-most number denotes the AU code. Although the average scores of the two databases are quite similar, nevertheless there are instances where individual AU performance differs significantly between the two databases. One possible explanation could be the disparity between the number of positive samples in respective databases; for example in the three AUs (11, 16, 10) where the largest AuC differences occur, the ratios of positive sample populations (Cohn-Kanade / Bosphorus) are roughly, 2:1, 1:7, 1:7, respectively. Also, the confidence intervals are usually wider with smaller amount of positive samples.

Our findings support the results of Bartlett et al. [25] where they have demonstrated the state-of-the-art AU detection performed with Gabor features and AdaBoost, by obtaining 92.6% average AuC on the combined dataset of Cohn-Kanade and Ekman-Hager databases with leave-one-out cross validation. They stated that AdaBoost is only marginally better than linear-SVM. We have also reached similar conclusions, that Adaboost and linear-SVM algorithms perform similarly on the Cohn-Kanade and

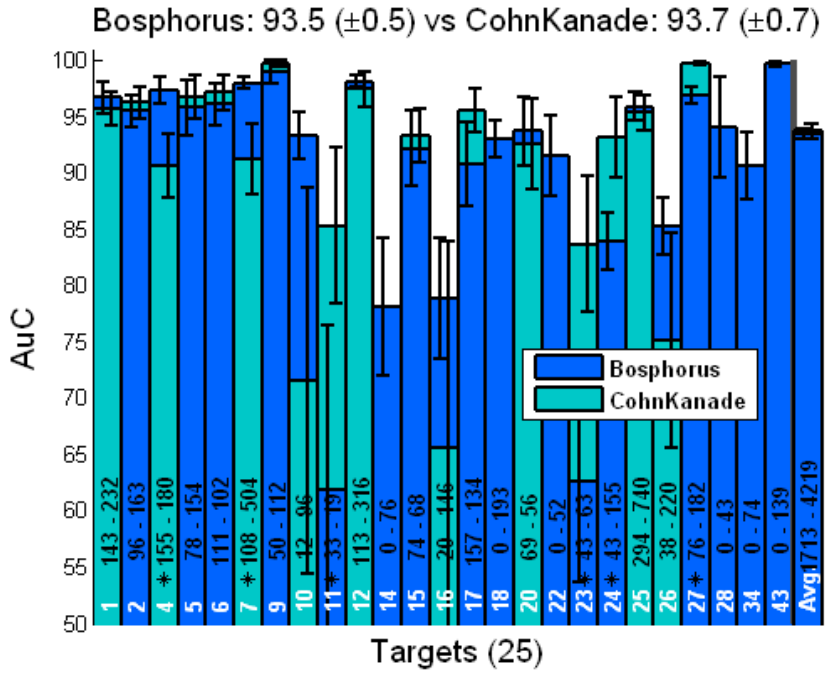


Figure 4.7. Performance comparisons under RBF-SVMs: Cohn-Kanade vs. Bosphorus datasets. On each AuC bar AU code (bottommost), a star if performances are significantly different, the number of total positive samples in the two datasets (bottom is Cohn-Kanade) and 95% confidence intervals are displayed.

Bosphorus databases. On the other hand, we found that RBF-SVM has the potential to improve 2D AU detection. According to the paired t-test under 5% significance level, RBF-SVMs are significantly better than AdaBoost and linear SVMs for eight of the AUs, and when averaged over all 25 AUs the overall score improved by slightly more than 1%.

4.9.4. Pure 3D AU Detection versus 2D AU Detection

In Table 4.2 we see that 3D modality is uniformly superior to 2D modality under all types of classifiers, and for all the various feature transforms as given in Figure 4.6. For example, the best feature and and classifier couple, i.e., the mean curvature - RBF-SVM couple achieves 95.5% average AuC under 3D modality, scoring 2% improvement over the 2D, and also with a higher confidence (smaller interval). The only exception is depth features from 3D whose performance is almost the same as that of 2D images

(Table 4.1 and Table 4.2). Note that in a previous study [48], which is based on landmarks and use of dynamic information via HMMs, the advantage of 3D had been shown for seven AUs (1, 2, 1+2, 4, 5, 15, 20, 27). The average correct classification rate rises with 3D from 80.5% to 87.3%, however the statistical significance of the performance differences are not known. Our average classification rate over the same AUs is almost the same 97.3% rate, however performance differentials vary from AU to AU. Also average AuC values, which measure the performance for all possible threshold values in contrast to correct classification rate, are 95.5% and 95.7% for 2D and 3D respectively. Notice that our higher rates may be due to the superiority of data-driven Gabor-based analysis over landmark-based recognition.

Average AU detection statistics do not show all the interesting aspects and hide big performance differentials for some AUs. When we portray detection rate of each AU in Figure 4.10(a), we observe that the advantages of 2D and 3D modalities differ from AU to AU. As a case in point, consider the bar that belongs to AU 23 and of which there are 63 realizations in the dataset. The darker part of the bar indicates that 2D data achieves only 62% correct detection while 3D data achieves 81%; conversely, consider the last AU bar (AU6 with 102 instances in the dataset). In this case 3D data achieves 87% while 2D is better at 96% correct detection rate. We can conclude that in general 3D data considerably improves the detection of lower facial AUs.

For example, improvements on the detection of AU 23 (Lip Tightener) and AU 24 (Lip Presser) are outstanding. AU 23 was shown to be one of the most difficult AUs both for automatic detection [25] and for human experts [21] since it is often confused with AU 24. 2D color, 3D surface, 3D mean curvature images belonging to four different instantiations of this AU, both alone single appearance and also in combination with other AUs (shown in rows four and five of Figure 4.9). Paired t-tests (with 5% significance level) have found that the following AUs incur significant improvements: AU 23 (Lip Tightener), AU 16 (Lower Lip Depressor), AU 24 (Lip Presser), AU 34 (Cheek Puff), AU 22 (Lip Funneler), AU 18 (Lip Pucker), AU 25 (Lips Part) and AU 2 (Outer Brow Raise), which are all quite significant. Figure 4.9 shows several instantiations of these AUs: AU 23 in (g-j), AU 16 in (n), AU 24 in (k-l

Table 4.3. Average AuC values over lower and upper AUs under the RBF-SVM classifier for 2D, 3D and fusion.

Face Part	#	**2D Lum.**	**3D Geom.**	**Fusion**
Lower AUs	2745	91.5 ± 0.7	95.7 ± 0.4	96.2 ± 0.3
Upper AUs	1474	97.2 ± 0.5	95.3 ± 0.7	97.3 ± 0.5

and o)), AU 34 in (o), AU 22 in (m), AU 18 in (l), AU 25 in (h, m, n and p), and AU 2 in (a,b). ROC curves of 2D and 3D data are also compared in Figure 4.8 for some of these AUs.

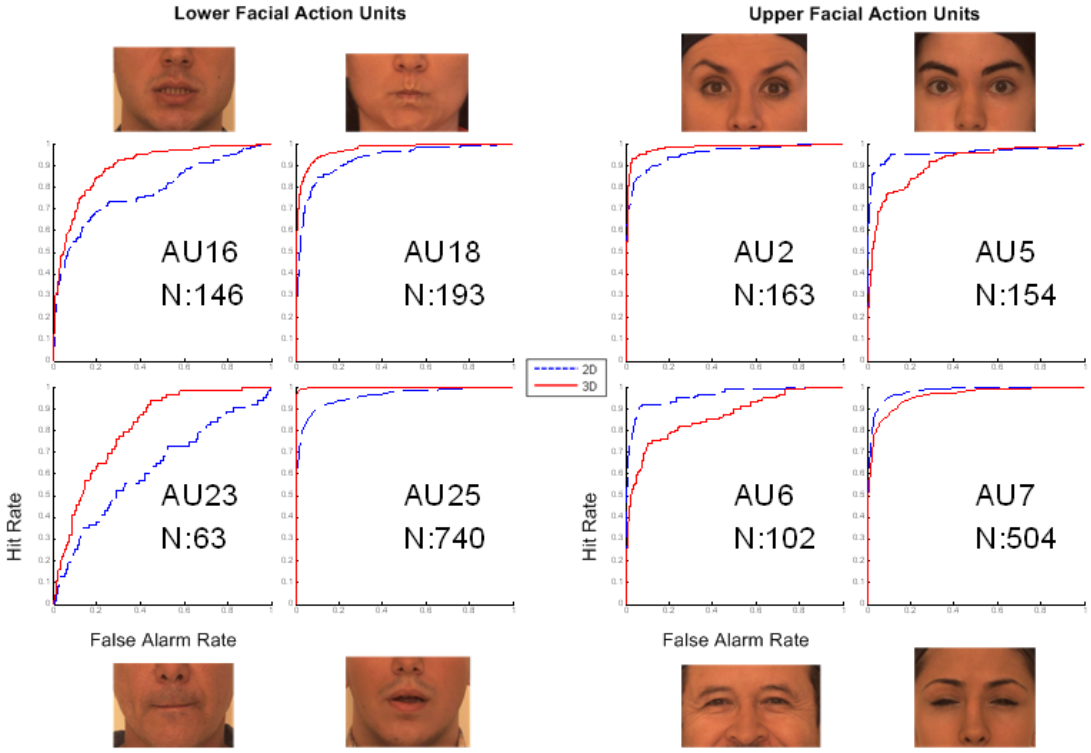


Figure 4.8. ROC curves of some of the AUs on which the two modalities differ significantly. N is the number of positive samples available; red curve denotes 3D results, blue dotted curve denotes 2D results.

3D is not necessarily always more advantageous. Some of the AUs have comparable detection rates in the two modalities. For instance, AU 12 (Lip Corner Puller) and AU 1 (Inner Brow Raise) achieve almost the same detection level in either modality, and also quite high levels of 98% and 97%, respectively. Two instances of AU 12 are shown in the (i) and (p) cells of Figure 4.9, and two instances of AU 1 are in the (a) and (c) cells of Figure 4.9. Moreover, we can see from Figure 4.10(a) that non-negligible

performance degradations occur on eye related upper face AUs in 3D data. According to the paired t-test, degradations on AU 6 (Cheek Raise), AU 5 (Upper Lid Raise), and AU 7 (Lids Tight) are statistically significant. This may be explained by two factors. First, 2D eye texture and especially appearance of pupils are very informative for AU scoring. Second, eye region can be quite noisy with structured light based 3D acquisition due to eyelashes and glitters. This may hide necessary surface detail for detection. ROC curves for 2D and 3D modalities are compared for these AUs in Figure 4.8, and some instances are given in Figure 4.9: AU 5 in (a-b), AU 6 in (d), and AU 7 in (c-e). The differing performances of 2D and 3D modalities with respect to upper and lower face AUs is summarized in Table 4.3. Clearly 3D is better on the average for lower face and 2D is better (though to a lesser degree) for upper face AUs.

4.9.5. Fusion of 2D and 3D Modalities

As seen in Table 4.2, the fusion of 2D and 3D provides additional 1.1% increase over 3D data, achieving 96.6% average AuC (under RBF-SVM). Figure 4.10(b) compares 3D curvature data versus and feature fusion of 3D curvature and luminance modalities on individual AUs. An increase in the AuC values is seen for all the AUs, except for a surprising 7% drop in AU 23. Though paired t-test and overlap of the confidence intervals tell us that this degradation is statistically insignificant, it may be better to completely avoid use of 2D modality for this AU. Recall that for AU 23, best 2D performance was 63% for 2D and 81% for 3D, while after fusion the score regresses to 72%. The improvements brought by fusion over the 3D modality were found to be statistically significant on five of the AUs. All the AUs for which the 3D detection is inferior to 2D (AU 5, 6 and 7), as shown in Figure 4.10(a), are within those seven AUs. Moreover, we observe the improvements by fusion with respect to lower and upper facial regions in Table 4.3, which tells us that greater portion of the performance gain was realized in the upper face. These results obtained by feature fusion indicate first, the existence of useful complementary information between the two modalities, and second, show that the relative weakness of 3D on the upper face region (only for eye region) can be compensated by feature fusion with the 2D data.

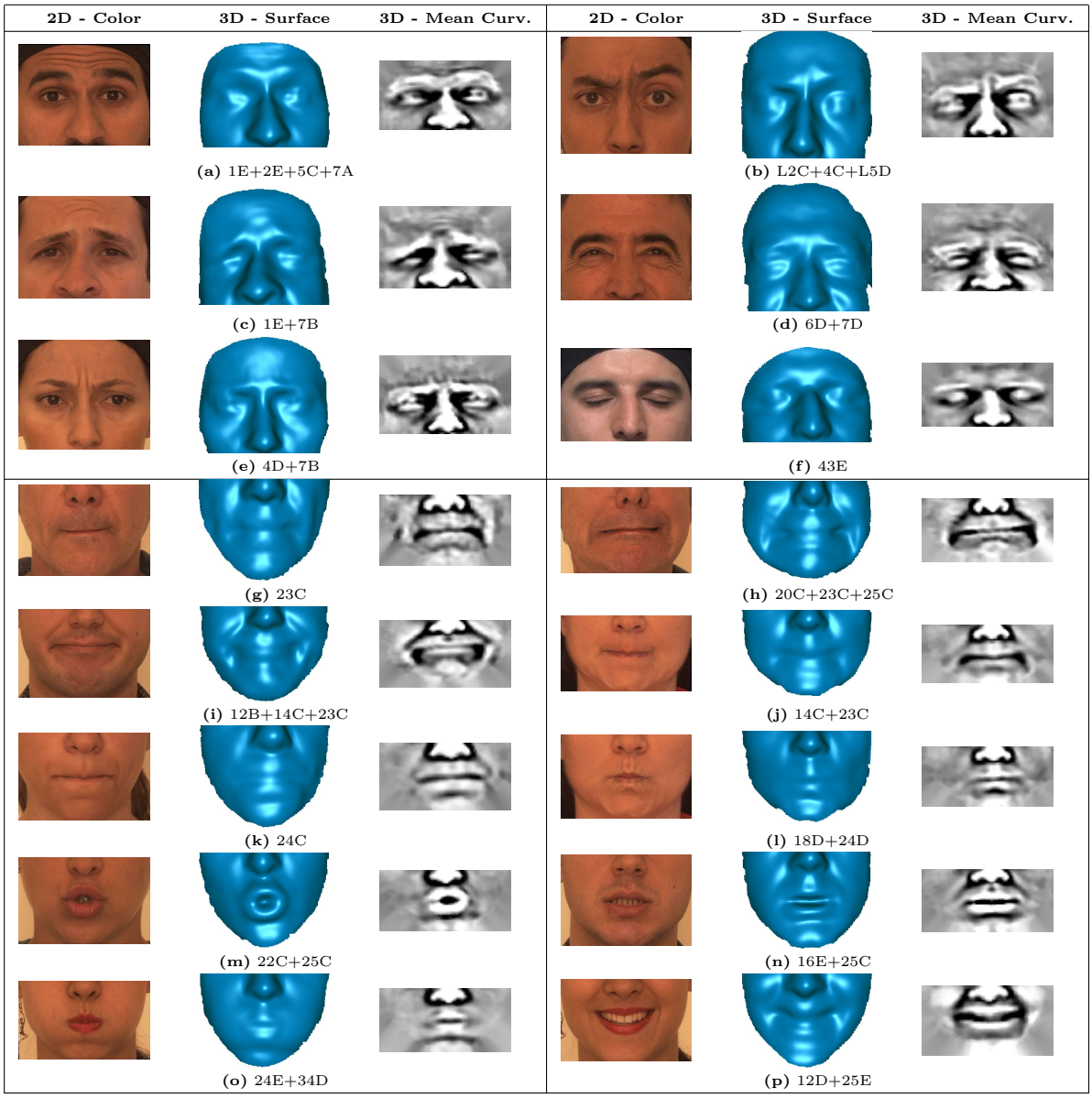
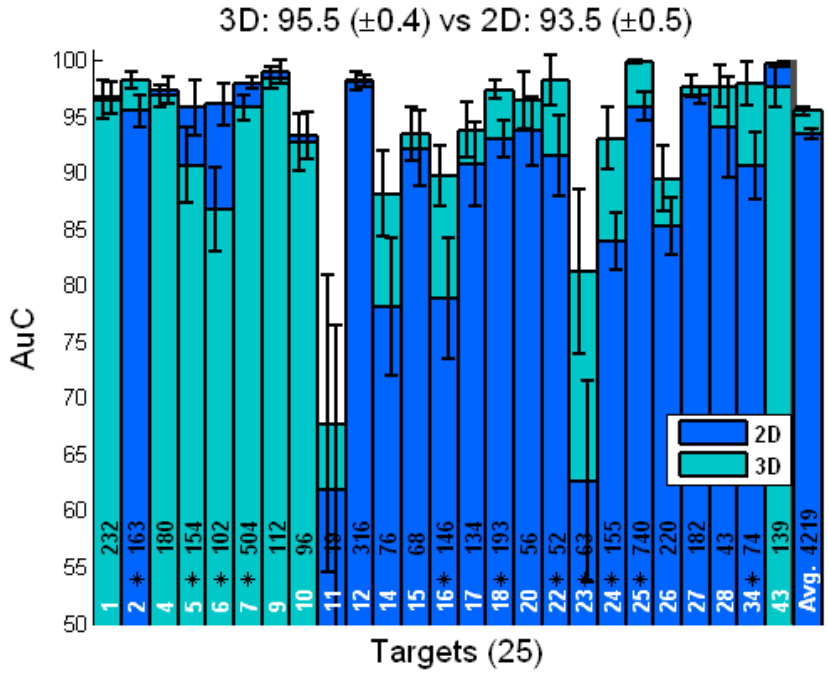


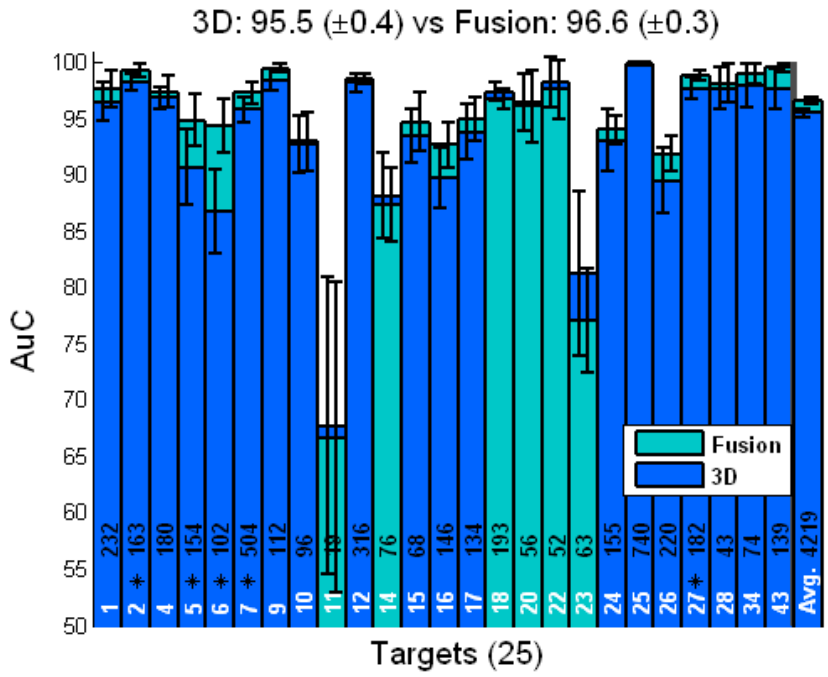
Figure 4.9. Color, 3D surface and surface mean curvature images are shown for some instances of upper AUs (upper three rows) and lower AUs (lower five rows), for which 2D and 3D detection performance differs significantly.

4.9.6. Effect of 3D Pose Normalization

A previous study [44] shows that when recognizers are trained with frontal faces and tested with systematically increasing pitch and yaw head rotations, sudden drop in 2D expression recognition performance is observed while 3D recognition performances remains constant. This is because with 2D we have to confront substantial distortions and occlusions due to the out-of-plane rotations. In our study we also evaluate the



(a)



(b)

Figure 4.10. Performance comparisons under RBF-SVMs: (a) 3D vs. 2D and (b) fusion vs. 3D. On each AuC bar AU code (bottommost), a star if performances are significantly different, positive samples and 95% confidence intervals are displayed.

effect of 3D pose, however, instead of using synthetic (rotated) luminance images we experiment on natural mild pose variations present in our database (Figure 3.4), and trained and test 2D luminance detectors both on original and 3D pose normalized luminance images.

We observe the effect of 3D pose normalization with respect to 2D normalization by comparing the average AuC scores in the Bosphorus database: 93.5% for 2D luminance, 94.7% for 3D luminance and 95.5% for 3D geometry (with RBF-SVM). It is interesting to note that although no 3D geometry related feature is used, we still obtain above 1% performance gain when 2D modality is aided (pose corrected) by 3D modality. The major difference between the direct 2D luminance and 3D luminance images is in the registration phase. 3D data permits better normalization and thus luminance data is compensated somewhat for moderate out-of-plane rotations. Recall that even for frontal images, unintentional slight in-plane and out-of-plane small pose variations always occur.

We also observe from Table 4.2 that automatic 3D pose alignment by ICP causes performance degradations. The drop from 95.5% to 94.8% points out to the possible improvements that may be attained if better registration techniques can be employed for fully automatic 3D systems.

4.9.7. Performance of Action Units at Low Intensity

In real life, facial expressions often occur in a broad range, including low intensities. Naturally, this makes the already challenging AU detection problem even more difficult as differences among AUs or between an AU and the neutral state become more subtle. Bartlett et al. [25] found 21% drop (from 92% to 71%) on average AuC values when their detectors were tested on a spontaneous expression database instead of posed expressions.

In order to see detection potential of lower intensity actions, we use only the B level samples as positives. The same subject cross validation partitions as in the

previous experiments are used, that is, training sets are not modified, they still consist of C, D and E level AUs, and the negative test samples consist of all other AUs. Hence, the only difference in the setup is the use of low intensity positive test samples instead of higher intensity samples. Table 4.4 gives the results for the 2D, 3D, and fusion. The performance ordering for the three 2D, 3D and fusion cases are the same as before, but we see severe 13% to 14% point performance drops on average AuCs: from 93.5% to 79.6% for 2D, from 95.5% to 82.5% for 3D and from 96.6% to 84.2% for fusion. Also, the confidence intervals increase by more than a factor of two. For these difficult low intensity B expressions 3D outperforms 2D by 2.9%. Recall that the improvement of 3D over 2D for higher intensity (C,D,E) AUs was 2%.

Figure 4.11 compares the differences between 3D and 2D modalities for low intensity AUs. Conclusions similar to those given in Section 4.9.4 can be drawn: in general, lower face AUs are more easily detected with 3D as compared to upper AUs. Paired t-tests indicate twelve of AUs as significantly differing between the modalities, 11 AUs in favor of 3D and 1 AU in favor of 2D (Figure 4.11). Low and high intensity samples of some of the, including the significantly different AUs, are illustrated for in Figure 4.14 and Figure 4.15 in both 2D and 3D modality. We want to point out some interesting cases. For instance, AU9 (Nose Wrinkler) maintains its high detection rate under either modality, and for AUs 25 (Lips Part), AU 22 (Lip Funneler) and AU 4 (Brow Lowerer) the detection rates drop considerably with 2D data while 3D detectors are still performing reasonably well. On the other hand, degradation in the detection of AU43 (Eye Closure) is much more in 3D than its 2D counterpart, which may be related to eye texture and 3D acquisition noise as mentioned in Section 4.9.4.

4.9.8. Assessment of the Classifiers

In all the experiments we employed AdaBoost, Naïve Bayes, linear and RBF-SVMs. In general, these different classifiers almost never affect the relative ranking of the classifiers for any of investigated factors. It even seems that there is a common trend between the classifiers in the AU detection problem. First, the most prominent observation is that RBF-SVM is almost always superior to the other three classifiers.

Table 4.4. Average AuC values and 95% confidence interval estimates for B level low intensity AUs. All of the classifiers use 200 Gabor features that are selected by AdaBoost.

Classifier	**2D Lum.**	**3D Geom.**	**Fusion**
AdaBoost	78.6 ± 1.1	81.7 ± 1.1	83.4 ± 1.1
Linear-SVM	77.9 ± 1.2	81.9 ± 1.1	83.1 ± 1.1
RBF-SVM	**79.6 ± 1.0**	**82.5 ± 1.0**	**84.2 ± 1.1**
Naïve Bayes	79.2 ± 1.4	82.1 ± 1.2	**84.2 ± 1.2**

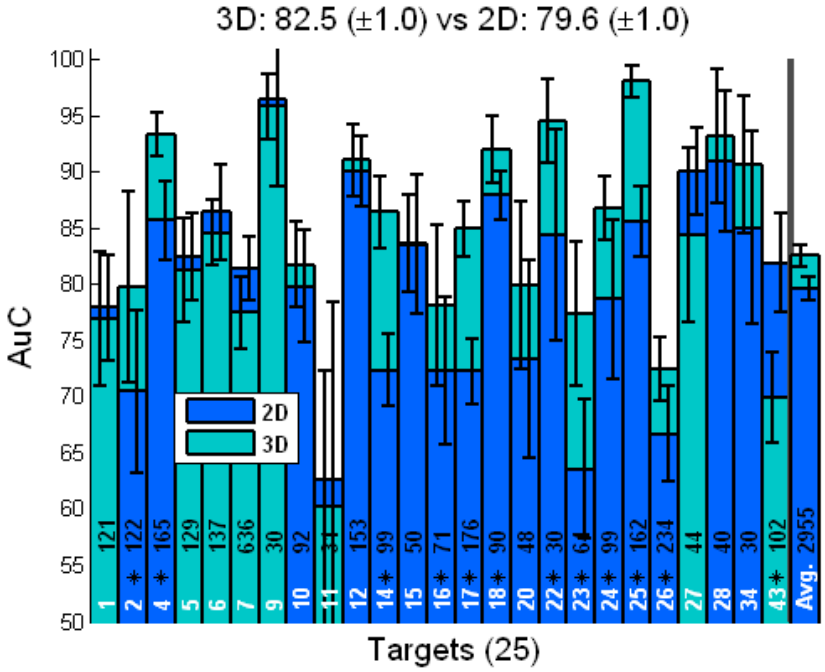


Figure 4.11. Low intensity (B) detection performance comparisons under RBF-SVMs: 3D vs. 2D. On each AuC bar AU code (bottommost), a star if performances are significantly different, positive samples and 95% confidence intervals are displayed.

Naïve Bayes is a close runner-up in 3D modality; AdaBoost and linear SVMs obtain clearly lower detection results.

The superiority of RBF-SVM stems from its ability in handling various types of non-linearities. It is known that depending on the Gaussian spread and SVM capacity parameters non-linearities in a quite wide range can be captured by RBF-SVMs. Therefore, we optimize these hyper-parameters for each AU. The main reason that

explains its advantage may be the high degree of variability found in AU instances. Our database involves many co-articulated AUs, asymmetric occurrences, instances with varying intensity levels which in turn may cause irregular dispersion of samples in the feature space. It is expected that especially the co-articulations may cause some cluster in the feature spaces. These variations may not be handled by linear discrimination as effective as non-linear discrimination provided by RBFs. Second factor may be the high degree of variations on luminance data in Bosphorus database. Though we wanted to eliminate the effect of lighting variations for 3D vs. 2D comparison by making acquisitions in controlled environment, we still have high degree of textural variations (albedo), especially due to facial hair. Maybe because of these variations RBFs demonstrate clear advantage with luminance data. However, for the Cohn-Kanade database Nave Bayes still achieves similar average performance. According to our conjecture this may be because Cohn-Kanade database does not involve facial hair (except a slight moustache for one subject) out of 97 subjects, in contrast to our database which involves 35 men with beard/moustache (19 intense, 16 moderate facial hair) out of 105 subjects.

Occasionally, Naïve Bayes may outperform RBF-SVM in an individual AU. Naïve Bayes is one of the simplest classifiers and is based on strong assumptions such as Gaussian distribution, uncorrelated samples and linear separability, and yet it is successful with 3D, though not as much with 2D. We conjecture that Naïve Bayes owes its good performance to the suitability of 3D data features vis-a-vis 2D data features. These outcomes suggest that RBF-SVM must be used for 2D modality, while with 3D modality either Naïve Bayes or RBF-SVM can be used. Recall that the Naïve Bayes alternative is extremely simple in training compared to SVM and it does not suffer from overlearning problem of the discriminative classifiers.

4.9.9. Assessment of the Intensity Estimators

The same experimentation setup as in the detection problem, i.e., 10-fold subject cross validation for each of 25 AUs, is also used for intensity estimation experiments. We employ 200 Gabor features. The correlations calculated over all AUs are listed in

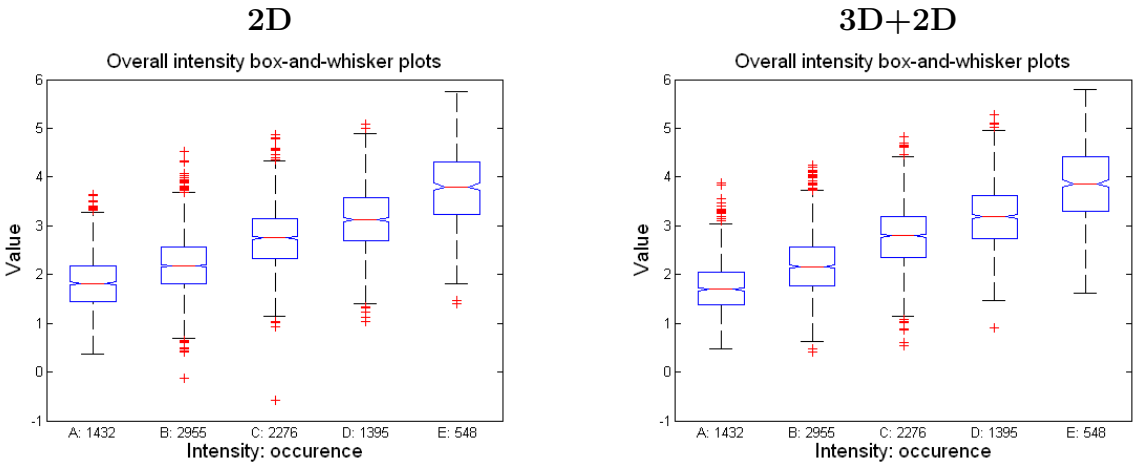


Figure 4.12. Distributions of estimated intensity values (via ε-SVM regression on 200 Gabor features) over all AUs for 2D and 3D fusion are shown (central mark: median, box: interquartile range, whiskers: extreme values, '+': outlier).

Table 4.5. Correlation of the estimated intensities with the scores of the FACS annotator.

	SVM Margins		***Image Features***	
Data	**Direct**	**Logistic**	**ε-SVM-Lin**	**ε-SVM-RBF**
2D	0.51	0.53	0.54	0.58
3D	0.50	0.51	0.53	0.56
3D+2D	0.53	0.55	0.59	**0.62**

Table 4.5. In the first column we see the performance of SVM-margins method, and in the second column the performance of the logistic regression on SVM-margins. Notice that in a previous study Bartlett et al. [25] have obtained a correlation performance 0.53 with 2D luminance images over six AUs using linear-SVM-margins. Using 2D data and RBF-SVM margins over the same set of AUs (1, 2, 4, 5, 10 and 20) we have obtained 0.62, however, over 25 AUs the average performance is 0.51. When we apply logistic regression, we obtain 0.53 (Table 4.5). On the other hand, we see that the improvements with ε-SVM regressor with image features yield higher correlations, and the non-linear RBF modeling achieves 0.58 correlation. The best overall result, 0.62, is obtained by fusion using ε-SVM with RBFs. Figure 4.12 shows the distributions of estimated intensity values by ε-SVM regression on 200 Gabor features over all AUs. We observe upward trend, and distribution medians of the consecutive intensity levels

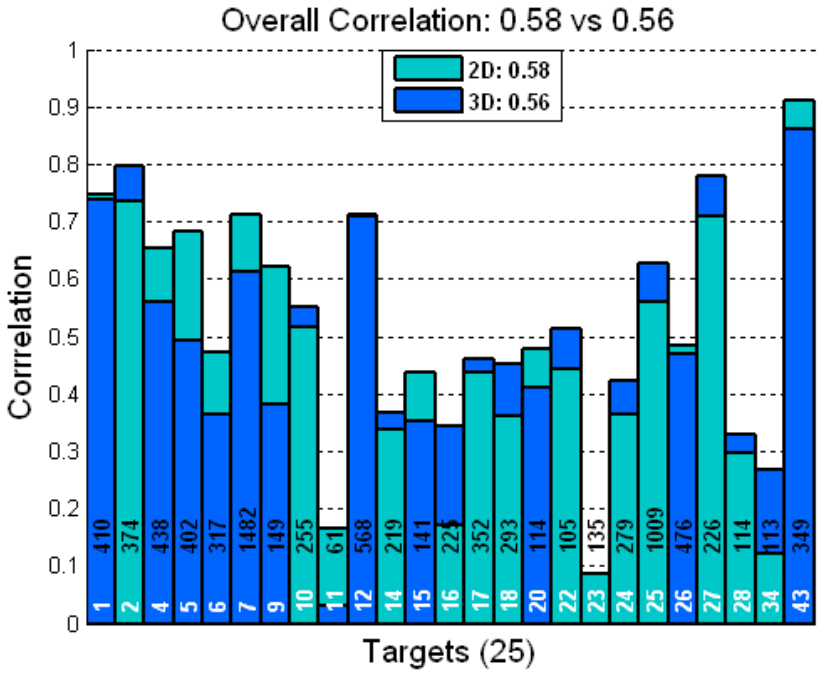
are significantly different at 5% significance level since their notches[3] do not overlap, i.e., their intervals do not overlap.

We also experimented with different number of features and found that using more than 200 features does not improve: with 400 features the results are 0.59, 0.57 and 0.62 for 2D, 3D and fusion respectively.

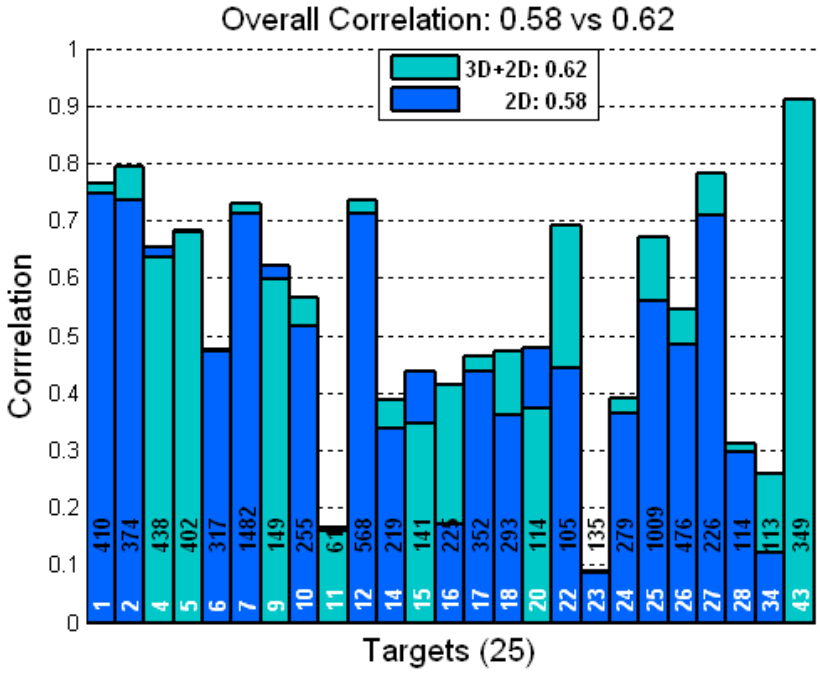
When we compare the performance of data modalities we see that 3D brings some improvement on averaged results only when used in conjunction with 2D data. Overall averaging may hide some important information, hence we compare intensity estimation for each AU over 2D and 3D data modalities in Figure 4.13(a). The number of the available AU samples are inscribed on each AU bar. We see that, with 2D data, most of the upper face AUs, AU 4 - Brow Lowerer, AU 5 - Upper Lid Raiser, AU 6 - Cheek Raise, AU 7 - Lids Tight and AU 43 - Eye Closure, as well as AU 9 - Nose Wrinkler achieve noticeably higher correlation than 3D data. On the other hand, 3D data seems to be more convenient for many lower face AUS, especially for AU 16 - Lower Lip Depressor, AU 18 - Lip Pucker, AU 22 - Lip Funneler, AU 25 - Lips Part, AU 27 - Mouth Stretch and AU 34 - Puff, as well as for AU 2 - Outer Brow Raise. Samples of some of these AUs are shown in Figure 4.14 and 4.15.

In contrast to intensity estimation, the improvements in overall AU detection performances by 3D data are much more substantial. The AuC detection results averaged over 25 AUs are 93.5%, 95.5% and 96.6% for 2D (luminance) data, 3D data (mean curvature) and for their fusion, respectively. Nevertheless, from Figure 4.13(a) it is understood that the this overall performance contrast between detection and estimation is actually not due to the inferiority of 3D modality. In fact, the advantages and disadvantages of the 3D modality for intensity estimation conforms to the results of detection for most of the AUs, however higher performance drops on certain AUs that have much more samples, such as AU 7, inverts the overall performances. One expects normally that 3D data would be more informative for he intensity estimation problem. Explanation for why this promise is not fulfilled are as follows. 3D sensing noise is

[3]Notches are the first quartile-to-median and median-to-third quartile ranges



(a)



(b)

Figure 4.13. Intensity estimation performance comparisons under ε-SVM (RBF) regression: (a) 2D vs. 3D and (b) 2D vs. fusion. The bar heights indicate correlation values. The AU code and the total number of occurrences are inscribed in the bars.

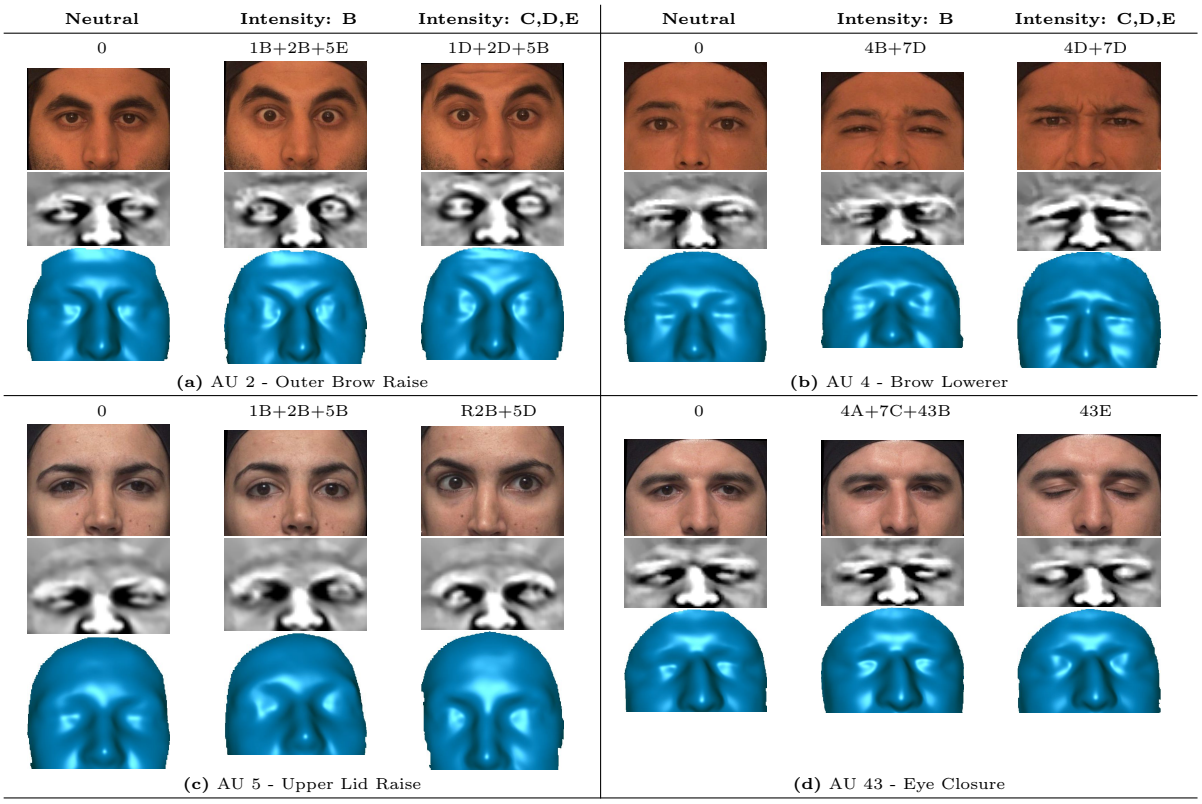


Figure 4.14. Color, surface mean curvature, and 3D surface images are shown for low (level B) and high (level C, D or E) intensity instances of several upper face action units together with the neutrals from the same subject and with the FACS codes.

excessive in the eye region and 3D misses the eye texture information. Moreover, the ground-truth data in manual FACS scoring is generated based on the observation of 2D appearances, which may generate a bias in favor of 2D. It is imaginable that the FACS annotator could have defined the intensity labels slightly differently using 3D data.

From Figure 4.13(b) we see that by means of modality fusion we are able to preserve the highest correlations of 2D and 3D modalities in general. However, interestingly, even though the correlation values of AU 22 are around 0.5 for 2D and 3D, it is boosted to 0.7 with fusion. These results shows the importance of fusion with 3D, as in the detection problem.

Figure 4.15. Color, surface mean curvature and 3D surface images are shown for low (level B) and high (level C, D or E) intensity instances of several lower face action units together with the neutrals from the same subject and with the FACS codes.

5. NON-RIGID REGISTRATION OF 3D SURFACES

Non-rigid registration of 3D surfaces is encountered in a variety of applications. For human face processing it is an especially crucial intermediate step. Among applications of non-rigid registration we can mention one-to-one correspondence between faces with the purpose statistical face model construction that can be used for instance in face reconstruction or recognition [54]. In computer animation, 3D characters can be animated using available expression faces, or from recorded 3D facial videos that are captured by 3D sensors [85, 64, 86]. Other applications are transfer of textures, surface details and animation controls between objects as well as generating morphing animations. The application of non-rigid registration in our work is the novel approach we propose for human facial expression analysis. Naturally, most of the expression analysis applications demand a fully automatic and fast registration algorithms.

In this study we carry out the registration task in 2D instead of 3D space since 2D processing can be more efficient. Also, 3D mesh-based methods have some disadvantages like sensitivity to mesh resolution and topology, and difficulties in multiresolution implementation. The block diagram of 3D surface registration implemented on 2D maps is depicted in Figure 5.1.

The first step is to map the reference surface Ω_A and the target surface Ω_B onto respective 2D image planes. Their planar parameterizations yield 2D domains $D_A, D_B \subset R^2$. This can be achieved by projection (pseudo-parameterization) or by LSCM method as explained in Section 4.3.2. Hence, if we can find a mapping from D_A to D_B (a 2D mapping), this would indirectly lead to the mapping between the surfaces in 3D space. To this effect, deformation of the reference map to match it to a target map is estimated by our image registration technique based on deformable 2D triangular meshes in the 2D parametric space. The mapping resulting from this deformation completes the transformation chain from the reference surface towards the target, and thus we succeed in registering the 3D target by resampling it over the domain of the reference surface. However, this 3D-3D mapping is not bijective

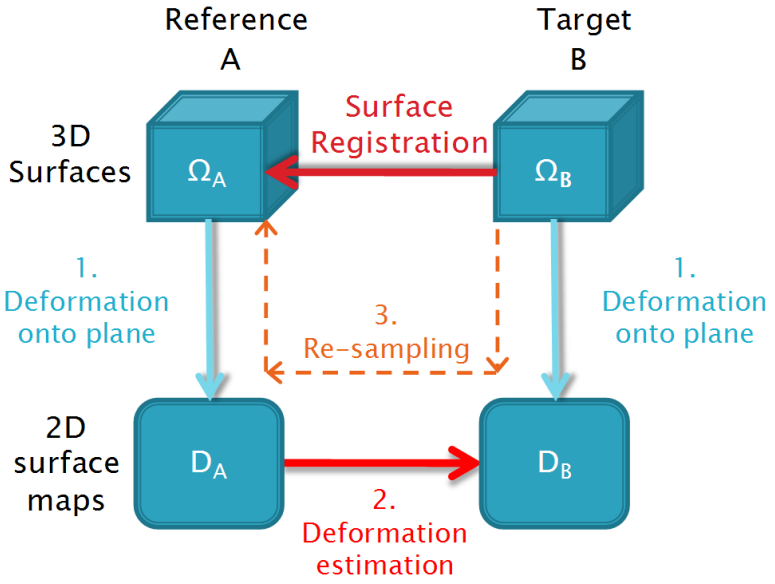


Figure 5.1. Overview of the 3D surface registration realized on 2D maps.

when working with surfaces which are cropped arbitrarily, and thus may not overlap everywhere. To alleviate this situation surface extrapolation has been proposed in Section 5.4.

In this scheme, the most critical component is the image registration part where non-rigid deformations are estimated. In our approach we assume that the two surfaces are hyper-elastic membranes. We prefer to use hyper-elastic deformation-based registration instead of elastic deformation in order to better deal with facial expressions, since small deformation theory becomes insufficient when dealing with soft tissue of the skin. However, computational load becomes an important issue when working with hyper-elastic models as in also other models of large deformation. This is especially not acceptable if the registration algorithm would be used for real-time facial expression analyses. We cope with this issue by developing an efficient 2D triangular mesh-based technique. Our method is described in the following section.

5.1. Deformable 2D Triangular Mesh-based Registration

A mapping from D_A to D_B (Figure 5.1) can be expressed via a vector field

$$\boldsymbol{\phi}(\mathbf{p}) = \mathbf{p} + \mathbf{d}(\mathbf{p}), \tag{5.1}$$

where $\mathbf{p} = (u, v) \in D_A$ denotes the 2D image coordinates and $\mathbf{d}(\mathbf{p}) = (du, dv)$ is a constant displacement. Thus we can express the image constraint (Equation (2.1)) as

$$I_A(\mathbf{p}) = I_B(\boldsymbol{\phi}(\mathbf{p})). \tag{5.2}$$

The correspondences can then be obtained if we can solve the deformation field $\boldsymbol{\phi}(\mathbf{p})$ by minimizing the image matching energy

$$E_M(\boldsymbol{\phi}) = \frac{1}{2} \int_{\mathbf{p} \in D_A} \Big(I_B(\boldsymbol{\phi}(\mathbf{p})) - I_A(\mathbf{p}) \Big)^2 \, \mathbf{dp}. \tag{5.3}$$

However, ill-posed nature of this problem requires additional constraints to be imposed for regularization. For instance, linear elastic models are one of the common regularizers in image registration [87] and force the mapping to be injective. These models treat the reference image as a linear elastic solid which is deformed by forces derived from an image similarity measure. Alternatively, viscous fluid models [88] in place of elastic models or the demons algorithm [89] have been proposed to handle large deformations. Although they provide, higher flexibility, these models need much more computations and this also increases the risk of misregistration and violation of injectivity.

In this study we propose an efficient algorithm which employs a non-linear elastic model based on the hyperelastic St. Venant Kirchoff material. This model is in-between the linear elastic and the viscous fluid models according to the deformation extent it can handle. Although this non-linear elasticity is much more complex as compared to linear models, we show via finite element discretization with triangular elements that it is still efficiently and exactly computed without any approximation since the Jacobian matrix of the mapping over a triangular element is constant. This

is in contrast to recent work of Yanovsky et al. [90] where they have to approximate deformation gradients to keep the run time feasible via finite difference discretization. Also, our algorithm allows adjustment of the computational load of the image matching term. Moreover, an image adaptive mesh generation technique is introduced in order to produce a reasonable number of mesh triangles needed for efficiency of the algorithm.

5.1.1. Image Matching with Triangular Meshes

We can imagine the triangular mesh overlayed on a reference image as an elastic membrane or the finite element discretization of that membrane. Then minimization of the image matching energy term in Equation (5.3) over the reference mesh yields attraction forces at the vertices and drives its deformation. For each triangle t of the reference mesh, we have a mapping function $\mathbf{q} = \boldsymbol{\phi}_\mathbf{t}(\mathbf{p})$ that maps a point $\mathbf{p}$ in a triangle of the reference to the point $\mathbf{q}$ in the target, as illustrated in Figure 5.2. This function interpolates the mapped coordinates of the triangle vertices $\mathbf{q_k}$ $(k \in 1, 2, 3)$ by

$$\mathbf{q} = \boldsymbol{\phi}_\mathbf{t}(\mathbf{p}) = \sum_{k=1}^{3} b_k(\mathbf{p})\mathbf{q_k} \tag{5.4}$$

where barycentric coordinates are obtained by

$$\begin{aligned} b_1(\mathbf{p}) &= Area(\widehat{\mathbf{pp_2p_3}})/A_t, \\ b_2(\mathbf{p}) &= Area(\widehat{\mathbf{pp_3p_1}})/A_t, \\ b_3(\mathbf{p}) &= Area(\widehat{\mathbf{pp_1p_2}})/A_t, \end{aligned} \tag{5.5}$$

and A_t is the area of triangle t.

Thus, an image matching energy, E_M, which accounts for the total square match-

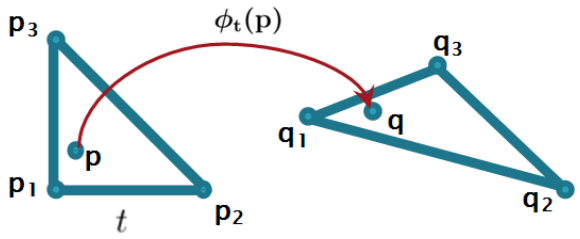


Figure 5.2. Transformation of triangle t by $\boldsymbol{\phi_t(p)}$.

ing error over domain D_A becomes

$$\begin{aligned} E_M(\boldsymbol{\phi}) &= \frac{1}{2}\int_{\mathbf{p}\in D_A} \|\mathbf{I_B}(\boldsymbol{\phi}(\mathbf{p})) - \mathbf{I_A}(\mathbf{p})\|^2_{\mathbf{W}}\, \mathbf{dp} \\ &= \frac{1}{2}\sum_{t\in T}\int_{\mathbf{p}\in D_t} \|\mathbf{I_B}(\boldsymbol{\phi_t}(\mathbf{p})) - \mathbf{I_A}(\mathbf{p})\|^2_{\mathbf{W}}\, \mathbf{dp} \end{aligned} \tag{5.6}$$

where T is the set of triangles, D_t is the domain of triangle t, $\mathbf{I_B}$ and $\mathbf{I_A}$ represent multi-modal image values, and $\mathbf{W}$ is a weighting matrix. Note that the matching energy in Equation (5.3) is written in a more generic form to include multi-channel images via weighted norm, so that we can exploit different surface attributes during matching. In this study diagonal weights, which are determined according to both variance and importance of the channels, are employed.

Having discretized the energy term over triangles, we can obtain the driving forces at the mesh nodes that deform a reference mesh to a target. These attraction forces result from energy minimization and are obtained by using the gradients at each mesh node n with respect to its mapped coordinates $\mathbf{q_n} = \boldsymbol{\phi}(\mathbf{p_n})$. The gradient at vertex n is obtained through the chain rule as

$$\begin{aligned} &\frac{\partial E_M}{\partial \mathbf{q_n}} = \sum_{t\in T_n}\int_{D_t} b_{k(t,n)}(\mathbf{p}) \left(\frac{\partial \mathbf{I_B}(\mathbf{q})}{\partial \mathbf{q}}\bigg|_{\boldsymbol{\phi_t}(\boldsymbol{p})}\right)^T \mathbf{W}\ \mathbf{e_t}(\mathbf{p})\, \mathbf{dp} \\ &\textit{where}\ \ \mathbf{e_t}(\mathbf{p}) = \mathbf{I_B}(\boldsymbol{\phi_t}(\mathbf{p})) - \mathbf{I_A}(\mathbf{p}). \end{aligned} \tag{5.7}$$

Here, T_n is the set of triangles connected to the node n, $k(t,n)$ is the k^{th} vertex of the triangle t that corresponds to node n, and $b_{k(t,n)}(\mathbf{p})$ is thus the k^{th} barycentric coordinate for the point $\mathbf{p}$ ($k \in 1,2,3$).

The gradients are evaluated at each node of the mesh to update the mapping vectors, $\boldsymbol{\phi}(\mathbf{p_n})$, in a gradient descent scheme. The image gradients are evaluated using 3x3 Scharr masks [91]. The integrals in equations (5.6) and (5.7) are approximated by sampling at the recursively subdivided triangle centers. This sampling procedure, however, is adapted to the area of triangles since mesh triangles can differ largely in area. Thus the recursive subdivision of the triangles is made proportional to their area. In this way, while avoiding unnecessary computations for small triangles, we can accurately approximate integrals over the larger triangles. Bilinear interpolation is used for resampling from discrete images. From the computational load point of view, evaluation of matching energy gradients can be the most time consuming part depending on the number of mesh triangles.

5.1.2. Nonlinear Elastic Deformation

Linear elastic modeling of deformations is often preferred in image registration due to its simple linear solution. However, many deformations cannot be accounted for by linear elastic models, which can account for only for small deformations. The Use of the St. Venant Kirchoff material model enables more flexible deformations since it is a simple hyperelastic material model. The potential energy of the St. Venant Kirchoff material is given by

$$W(\mathbf{E}) = \frac{\lambda}{2}(tr\mathbf{E})^2 + \mu tr\mathbf{E^2} \tag{5.8}$$

where λ and μ are the Lame material constants and $\mathbf{E}$ is the Green-Lagrange strain tensor. In the sequel we will also discuss the linear approximation of the Kirchoff material model, namely, the Hookean material. Notice that this model has linear stress-strain relationship, since the stress, $\partial W/\partial \mathbf{E}$, is linear in terms of the strain $\mathbf{E}$. However, strain-displacement relationship is non-linear, as shown below. Therefore, it becomes the simplest non-linear hyperelastic material model. The Green-Lagrange

strain tensor is defined in terms of deformation gradient tensor, $\boldsymbol{\nabla\phi}$, as

$$\begin{aligned}
&\mathbf{E} = \frac{1}{2}\left(\boldsymbol{\nabla\phi}^{\mathbf{T}}\boldsymbol{\nabla\phi} - \mathbf{I}\right) \\
&where\ \ \boldsymbol{\nabla\phi} = \frac{\boldsymbol{\partial\phi}}{\boldsymbol{\partial}\mathbf{p}} = \begin{pmatrix} 1+\dfrac{\partial du}{\partial u} & \dfrac{\partial dv}{\partial u} \\ \dfrac{\partial du}{\partial v} & 1+\dfrac{\partial dv}{\partial v} \end{pmatrix}\ \ and\ \ \mathbf{d}(\mathbf{p}) = \begin{pmatrix} du \\ dv \end{pmatrix}.
\end{aligned} \tag{5.9}$$

$$Defining\ \ \boldsymbol{\nabla}\mathbf{d} = \begin{pmatrix} \dfrac{\partial du}{\partial u} & \dfrac{\partial dv}{\partial u} \\ \dfrac{\partial du}{\partial v} & \dfrac{\partial dv}{\partial v} \end{pmatrix}\ \ displacement\, gradient\, tensor$$

$$\begin{aligned}
\mathbf{E} &= \frac{1}{2}\left(\left[\mathbf{I} + \boldsymbol{\nabla}\mathbf{d}^{\mathbf{T}}\right]\left[\mathbf{I} + \boldsymbol{\nabla}\mathbf{d}\right] - \mathbf{I}\right) \\
&= \frac{1}{2}\left(\mathbf{I} + \boldsymbol{\nabla}\mathbf{d}^{\mathbf{T}} + \boldsymbol{\nabla}\mathbf{d} + \boldsymbol{\nabla}\mathbf{d}^{\mathbf{T}}\boldsymbol{\nabla}\mathbf{d} - \mathbf{I}\right) \\
&= \frac{1}{2}\left(\boldsymbol{\nabla}\mathbf{d}^{\mathbf{T}} + \boldsymbol{\nabla}\mathbf{d} + \boldsymbol{\nabla}\mathbf{d}^{\mathbf{T}}\boldsymbol{\nabla}\mathbf{d}\right)
\end{aligned} \tag{5.10}$$

In equation (5.10) we see the nonlinear relationship between displacement $\mathbf{d}$ and strain $\mathbf{E}$ due to the last term. Omitting this last term results in linear models as given in Section 5.1.3. The advantage of nonlinear model is that, Green-Lagrange strain tensor is independent of rigid body motions, and thus measures pure deformations. To check this let's imagine we apply a rigid transformation by rotation $\mathbf{R}$ and translation $\mathbf{T}$ to the deformation $\boldsymbol{\phi}$

$$\boldsymbol{\phi}_{\mathbf{R}} = \mathbf{R}\boldsymbol{\phi} + \mathbf{T} \tag{5.11}$$

Then the new deformation gradient becomes

$$\boldsymbol{\nabla\phi}_{\mathbf{R}} = \mathbf{R}\boldsymbol{\nabla\phi}. \tag{5.12}$$

Since $\mathbf{R}$ is a rotation matrix,

$$\begin{aligned} \boldsymbol{\nabla\phi_R}^T\boldsymbol{\nabla\phi_R} &= (\mathbf{R}\boldsymbol{\nabla\phi})^T\mathbf{R}\boldsymbol{\nabla\phi} \\ &= \boldsymbol{\nabla\phi}^T\mathbf{R}^T\mathbf{R}\boldsymbol{\nabla\phi} \\ &= \boldsymbol{\nabla\phi}^{\mathbf{T}}\boldsymbol{\nabla\phi}. \end{aligned} \tag{5.13}$$

Hence, Green-Lagrange strain (Equation (5.9) is independent of rigid body motion, since strain remains the same after any rigid motion.

Here, we evaluate a deformation energy which is an internal strain energy based on the Froebenius norm of the Green-Lagrange strain tensor.

$$E_D(\boldsymbol{\phi}) = \int_{\mathbf{p}\in D_A} \|\mathbf{E}(\boldsymbol{\phi}(\mathbf{p}))\|_{\mathbf{F}}^{\mathbf{2}}\,\mathbf{dp}, \tag{5.14}$$

This is energy is a special case of the St. Venant Kirchoff material energy, because the Froebenius norm corresponds to Equation (5.8) with $\lambda = 0$ and $\mu = 1$. Actually for a physical object we have $\mu > 0$ and $\lambda+\mu > 0$, and λ determines deformations in direction orthogonal to external forces, e.g., when we squeeze a rubber in horizontal direction it can elongate in vertical direction. We neglect this term (i.e., setting $\lambda = 0$) to reduce complexity, since our purpose is not correct physical modeling, but regularization.

For triangular discretization, deformation energy can also be expressed equivalently as

$$E_D(\boldsymbol{\phi}) = \sum_{t\in T}\int_{\mathbf{p}\in D_t} \|\mathbf{E}(\boldsymbol{\phi_t}(\mathbf{p}))\|_{\mathbf{F}}^{\mathbf{2}}\,\mathbf{dp}. \tag{5.15}$$

Over a triangular element, Jacobian matrix of the mapping function $\boldsymbol{\phi}(\mathbf{p})$, which is also named as deformation gradient tensor in elastic theory, is constant. This is shown in Appendix A. Thanks to this property we can conveniently evaluate hyperelastic deformations of triangular meshes since Green-Lagrange strain will be constant over triangular elements. However, without finite element discretization or with higher order elements exact calculations would be too complex and some approximations would be

necessary. Fortunately, we can evaluate the exact deformation energy simply by using the constant Green-Lagrange strain tensor of triangle t, $\boldsymbol{E}_t$, as

$$\begin{aligned} E_D(\boldsymbol{\phi}) &= \sum_{t\in T} A_t \|\boldsymbol{E}_t(\boldsymbol{\phi_t})\|_F^2 \\ &= \frac{1}{4}\sum_{t\in T} A_t \left((a-1)^2 + 2b^2 + (c-1)^2\right) \\ where\ \ a &= \left\|\frac{\partial \boldsymbol{\phi_t}}{\partial u}\right\|^2,\ b = \frac{\partial \boldsymbol{\phi_t}}{\partial u}^T.\frac{\partial \boldsymbol{\phi_t}}{\partial v},\ c = \left\|\frac{\partial \boldsymbol{\phi_t}}{\partial v}\right\|^2 \end{aligned} \tag{5.16}$$

where A_t is the mesh triangle area as in Equation (5.5), and the other terms are quantities belonging to triangle t. Analytic expressions for gradients of E_D with respect to vertex coordinate mappings are given in Appendix A.1. By means of these gradients, stresses at each node of the mesh are evaluated during energy minimization to regularize the displacements due to driving forces generated by image matching errors.

5.1.3. Linear Elastic Deformation

We can linearize the Green-Lagrange strain and obtain less complex gradient calculations of the deformation energy. This is achieved by omitting the last term in equation (5.10).

$$\boldsymbol{\varepsilon} = \frac{1}{2}\left(\boldsymbol{\nabla}\mathbf{d}^{\mathbf{T}} + \boldsymbol{\nabla}\mathbf{d}\right) = \begin{pmatrix} \dfrac{\partial du}{\partial u} & \dfrac{1}{2}\left(\dfrac{\partial du}{\partial v} + \dfrac{\partial dv}{\partial u}\right) \\ \dfrac{1}{2}\left(\dfrac{\partial du}{\partial v} + \dfrac{\partial dv}{\partial u}\right) & \dfrac{\partial dv}{\partial v} \end{pmatrix} \tag{5.17}$$

Here, $\boldsymbol{\varepsilon}$ is called Cauchy's strain tensor. Since the quadratic term $\boldsymbol{\nabla}\mathbf{d}^{\mathbf{T}}\boldsymbol{\nabla}\mathbf{d}$ disappears this approximation is valid for infinitesimal displacements. Substituting $\boldsymbol{\varepsilon}$ in place of $\boldsymbol{E}$ in Equation (5.8) we obtain the potential energy of linear elastic model known as Hooke's Law. From the image registration point of view, linear models may not be very appropriate as regularizer for some problems, since they will not allow higher degree of flexibility if necessary. To be applicable, typically, strains should not be greater than 1%. Also, not being independent to rigid body motion unlike the non-linear models,

they penalize rotations as well, though this is not an issue since it is common to perform an initial rigid alignment before non-rigid registration.

Using the Cauchy's strain tensor, the deformation energy of the triangular mesh membrane, E_D, becomes

$$\begin{aligned} E_D(\phi) &= \sum_{t \in T} A_t \|\boldsymbol{\varepsilon}_t(\boldsymbol{\phi_t})\|_F^2 \\ &= \sum_{t \in T} A_t \left(\left(\frac{\partial du}{\partial u} \right)^2 + \left(\frac{\partial dv}{\partial v} \right)^2 + \frac{1}{2} \left(\frac{\partial du}{\partial v} + \frac{\partial dv}{\partial u} \right)^2 \right) \end{aligned} \tag{5.18}$$

The derivation of the gradients with respect to triangle vertex coordinates is given in Appendix A.2. When gradients of the Green-Lagrange strain (Appendix A.1) are compared with Cauchy's, the number of operations needed for their calculation is higher. However, since these evaluations are not done for different coordinates inside triangular elements, which would be necessary to approximate the integrals if higher order elements were used, the computational burden is still not critical.

5.2. Adaptive Mesh Generation

The density of triangular meshes overlaid on the images is quite important because it determines the computational load of the registration. Without computational concerns, we can use very dense meshes. Another drawback of high resolution meshes may be the sensitivity to noise. Here we propose a technique to produce meshes on the 2D domains of mapped surfaces adapted to the actual image. The goal of this adaptation is to provide finer motion estimations only in regions where there is some structure to be matched. Therefore, domain meshes are generated adaptively as a function of image gradient as well as image resolution as illustrated in Figure 5.3.

First, a new domain contour is obtained by a smoothing and down-sampling procedure. This is needed because the contours of scanned face data can contain excessive amount of points and they can zigzag a lot. Smoothing is done by fitting a B-spline curve (third degree) with few control points. The number of control points

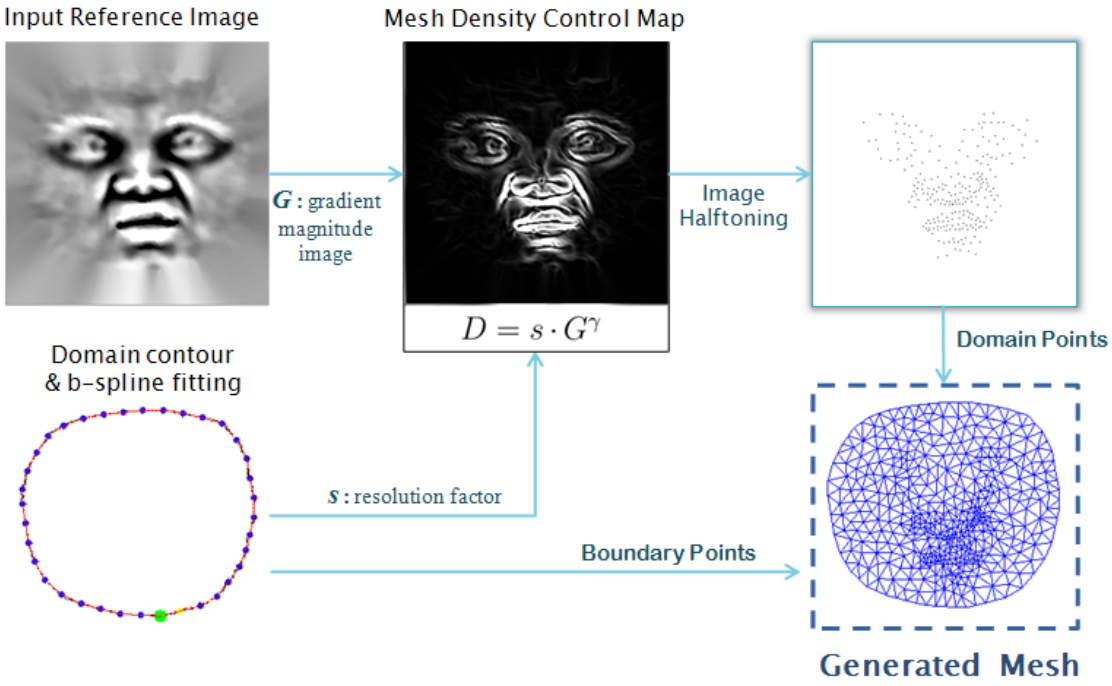


Figure 5.3. Steps of image adaptive mesh generation are illustrated on mean curvature map of a facial surface.

is proportional to the number of contour pixels, and thus smoothing depends on the image resolution as well. The new contour points are obtained by sampling this curve uniformly which defines the new reference domain.

In the second step, constraint points are created for the mesh vertices. These points are generated so that the mesh becomes relatively denser in regions where more features are available for matching, as for example around the lips and the nose. To produce these points efficiently, the method proposed for interactive geometry remeshing [92] is utilized. The idea is to generate a mesh density control map, and then to make image half-toning to produce binary pixels to be used as fixed mesh points. For this purpose, in our work we use gradient magnitude of the curvature images (G) after emphasizing them with power (γ) and scaling by factor s in order to adjust the number of points according to the desired resolution, as in the relation $D = s \cdot G^{\gamma}$. Here s is calculated according to the required number of sampling points N, and I which is sum of the image pixel intensities inside the domain. For eight bit quantization this quantity becomes $s = 255N/I$. Here, N is determined proportional to the square of the number of sampled contour points. Notice that γ is a control parameter that modifies

the relative mesh density, and it is set to 1.5 in Figure 5.3.

In the final step, Delaunay triangulation [93] is applied to the image consisting of half-tone dots and of the contour points, and then the mesh is generated according to the algorithm described in [94]. In our work, the maximum triangle edge length is limited by the average contour edge length so that in order to obtain regular triangles the mesh generation algorithm produces additional vertices.

This mesh generation method provides finer grained estimation where necessary, at lip contours for instance, and at the same time reduces the computational load by placing bigger triangles on featureless regions. As examples, generated meshes from the four different scales of the input image given in Figure 5.3 are shown in Figure 5.4.

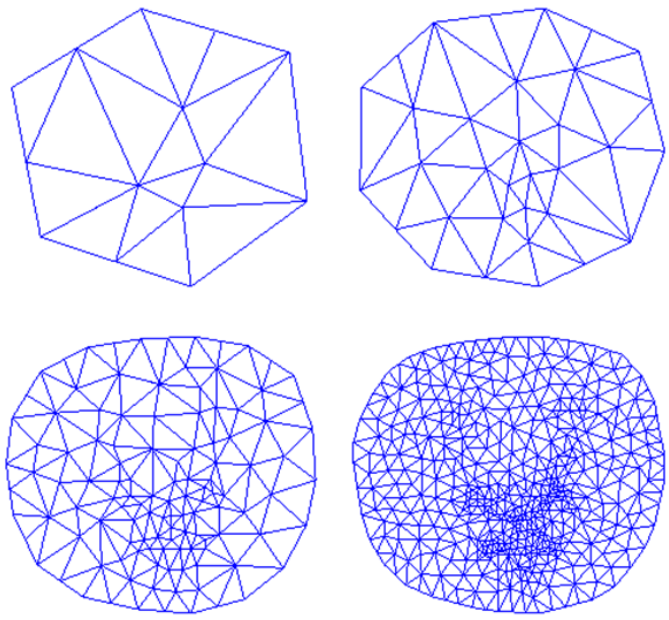


Figure 5.4. Meshes adaptively generated according to the four different scales of the input image given in Figure 5.3

5.3. Multiresolution Registration

A solution for the mapping function $\boldsymbol{\phi}$ is obtained by minimizing of a total energy term over the reference triangulation. The total energy expression is

$$E_T(\boldsymbol{\phi}) = E_M(\boldsymbol{\phi}) + \rho E_D(\boldsymbol{\phi}). \tag{5.19}$$

E_M is the total matching error, as in Equation (5.6), E_D is the strain energy, as in Equation (5.14), and rigidity of the deformable mesh is determined by ρ. If we increase the value of ρ, bigger image forces would be needed to balance the deformation energy and therefore global stretching or shrinking of the meshes would be harder. Thus, while setting to higher values would diminish the effects of forces due to noise and would improve regularization, this could also cause underestimation of the actual deformations.

The energy minimization is carried out via gradient descent in a coarse-to-fine approach. Multiresolution registration is a very common registration strategy since local minima can be avoided and faster convergence can be achieved with less computational burden. We implement this by using Gaussian image pyramids and meshes adapted to each image in a pyramid. Sample meshes at four different resolution levels are shown in Figure 5.4. Registration starts with the coarsest level, and estimated deformations are subsequently transferred to a finer level. The transfer of deformations from one mesh to another is realized via barycentric mapping. For each node of a finer level, its barycentric coordinates at the coarser mesh are calculated, and thus mapping of that node is obtained by a weighted sum of the previous mesh node values. Energy minimization restarts in the next higher resolution level initialized with the values of the previous step. This procedure continues until the finest scale and final deformation estimates are found.

5.4. Surface Resampling

Having obtained the 2D non-rigid registration in the parametric domain, 3D registration is achieved by resampling of the 3D surface. This is realized by resampling at the parametric coordinates of the target, which are calculated from their barycentric coordinates corresponding to reference vertices. However, we cannot guarantee to have all the resampling points inside the target domain. In other words, the correspondence estimation in image registration stage does not produce a bijective mapping. This situation can be disturbing especially in the case of partial correspondences on the forehead, below the chin, or at the cheeks due to rotated faces and single-view acquisition.

In order not to have missing correspondences after the registration, we predict these correspondences based on the estimated ones. This is achieved by finding a function that deforms the reference surface towards the target, and which is measured on the unmatched points. The deformation function can be obtained by fitting TPS [57] to the matched points. However, the use of all these points is not feasible because of the excessive computational load, especially for high resolution surfaces, as in this study. To reduce this computational burden, a subset of node points are chosen. Two criteria are considered for this selection: the global behavior of the deformation and the detailed local trend near the extrapolation points. To make a fast selection, a regular grid is formed in the parametric space and the surface vertices nearest to the grid points are picked. By extending the grid outside the given 2D domain, we can obtain more points in the neighborhood of missing target points. Mesh vertices selected inside and on the boundaries of a registered surface are shown in Figure 5.5. We see that the registered yellow surface does not completely overlap with the reference green surface, hence there would be some missing data points on the reference domain. However, after we extrapolate the surface via TPS using the selected mesh points, we can obtain a plausible extrapolation. In this method, the texture pixels are re-sampled via bilinear interpolation at the barycentric coordinates. Unmatched texture regions are extrapolated using an in-painting algorithm [77] as illustrated in Figure 5.5.

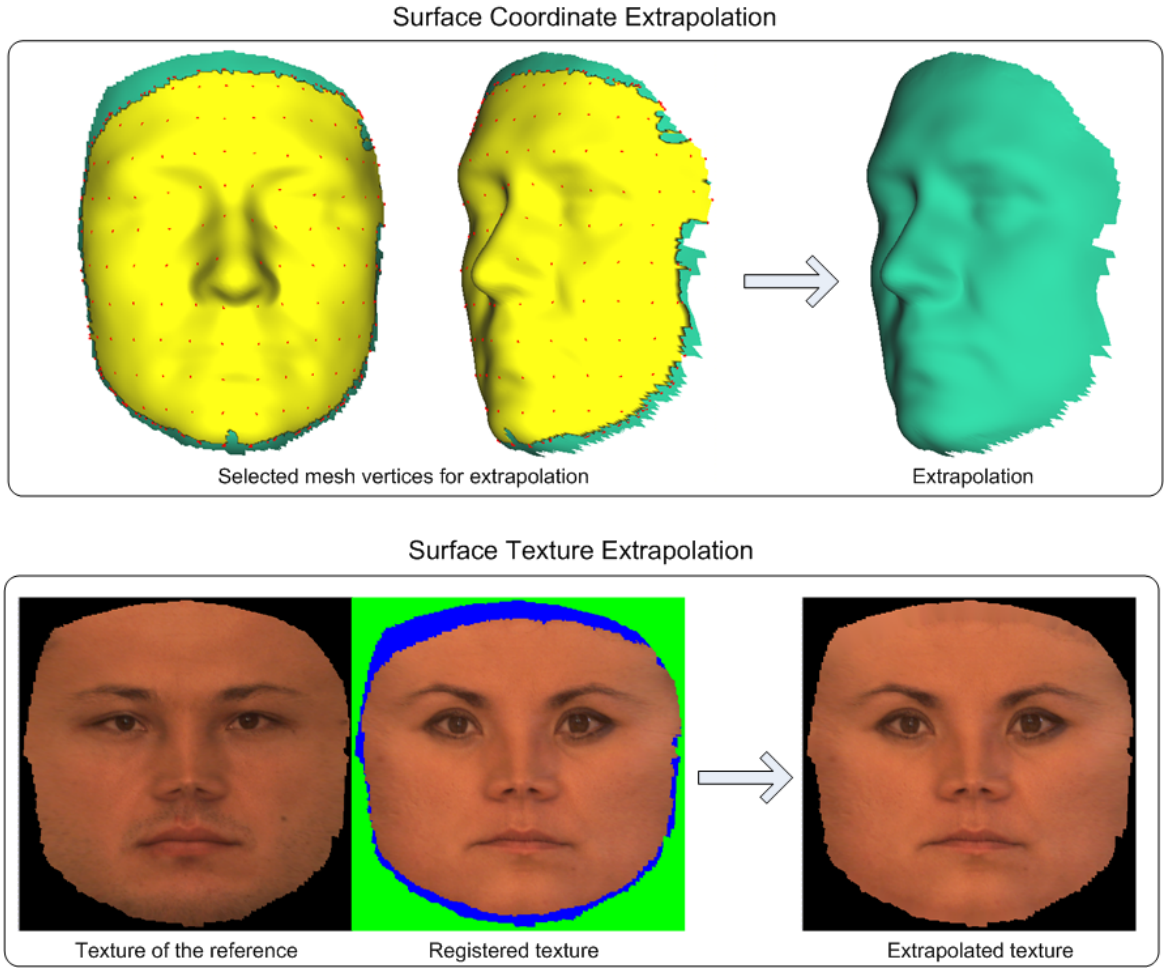


Figure 5.5. Top: coordinate extrapolation of the registered yellow surface and automatically selected control points. Bottom: extrapolation of the registered texture and non-overlapping reference domain is colored in blue.

5.5. Experiments and Discussion

5.5.1. Lucas-Kanade Optic Flow-based Registration

We consider a popular method, the method of Blanz et al. [54], in which optic flow estimation is employed to register range and texture maps of 3D scanned faces to construct 3DMMs. A 3DMM is a generative model of 3D shape and texture very similarly to AAMs (see Section 2.5.1). However, in contrast to AAMs, shape and texture are represented in more detail. To register range and texture maps of 3D scanned neutral faces, Lucas-Kanade [95] optic flow estimation is applied. The Lucas-Kanade algorithm fits an affine model to image patches in order to estimate the flow vectors.

Optical flow estimations are not always accurate everywhere due to the aperture problem and due to the lack of evidence in the homogenous regions like cheeks and therefore Lucas-Kanade method produces sparse motion vectors. Thus one has to perform additional smoothing interpolation to obtain dense one-to-one correspondences. Blanz et al. [54] apply a kind of diffusion reaction where a weighted combination of data and smoothing energy terms are minimized. The smoothing is done so that flow vectors in reliable regions and components of flow vectors orthogonal to edges (at aperture regions) do not deviate from the result of initial flow estimations. However, in regions that are deemed as totally unreliable, a smooth minimum-energy arrangement is made. By thresholding the eigenvalues of matrix $\mathbf{A}$,

$$\mathbf{A} = \begin{pmatrix} \sum \|\partial_u \mathbf{I}\|^2 & \sum \langle \partial_u \mathbf{I}, \partial_v \mathbf{I} \rangle \\ \sum \langle \partial_u \mathbf{I}, \partial_v \mathbf{I} \rangle & \sum \|\partial_v \mathbf{I}\|^2 \end{pmatrix} \tag{5.20}$$

where the summations are over a small neighborhood, one can diagnose whether the solution is reliable or not [95]. If the two eigenvalues of $\mathbf{A}$ are below a threshold the estimate is unreliable, and if only one is above it is due to the aperture problem. Lucas-Kanade optical flow method is implemented in a multiscale framework via image pyramids.

An example application of this method to register a neutral face with a face of smiling expression is illustrated in Figure 5.6. The registration is done on a 512×512 pixels resolution images of texture and mean curvature maps via five-level Gaussian image pyramids. The visualized flow vectors are the ones estimated at the coarsest level (32×32 pixels). The size of the image patches is 5×5 pixels. This figure shows the result of two different threshold values for the eigenvalues, 5×10^{-4} and 2×10^{-4}. We see that when the higher threshold value is used the lip corner puller action cannot be captured, so that the lower part of the source image is not properly morphed to the target image.

Due to the inaccuracies and sensitivity of thresholding, Blanz et al. [54] follow a recursive bootstrapping strategy to reconstruct morphable models of 3D faces. Using

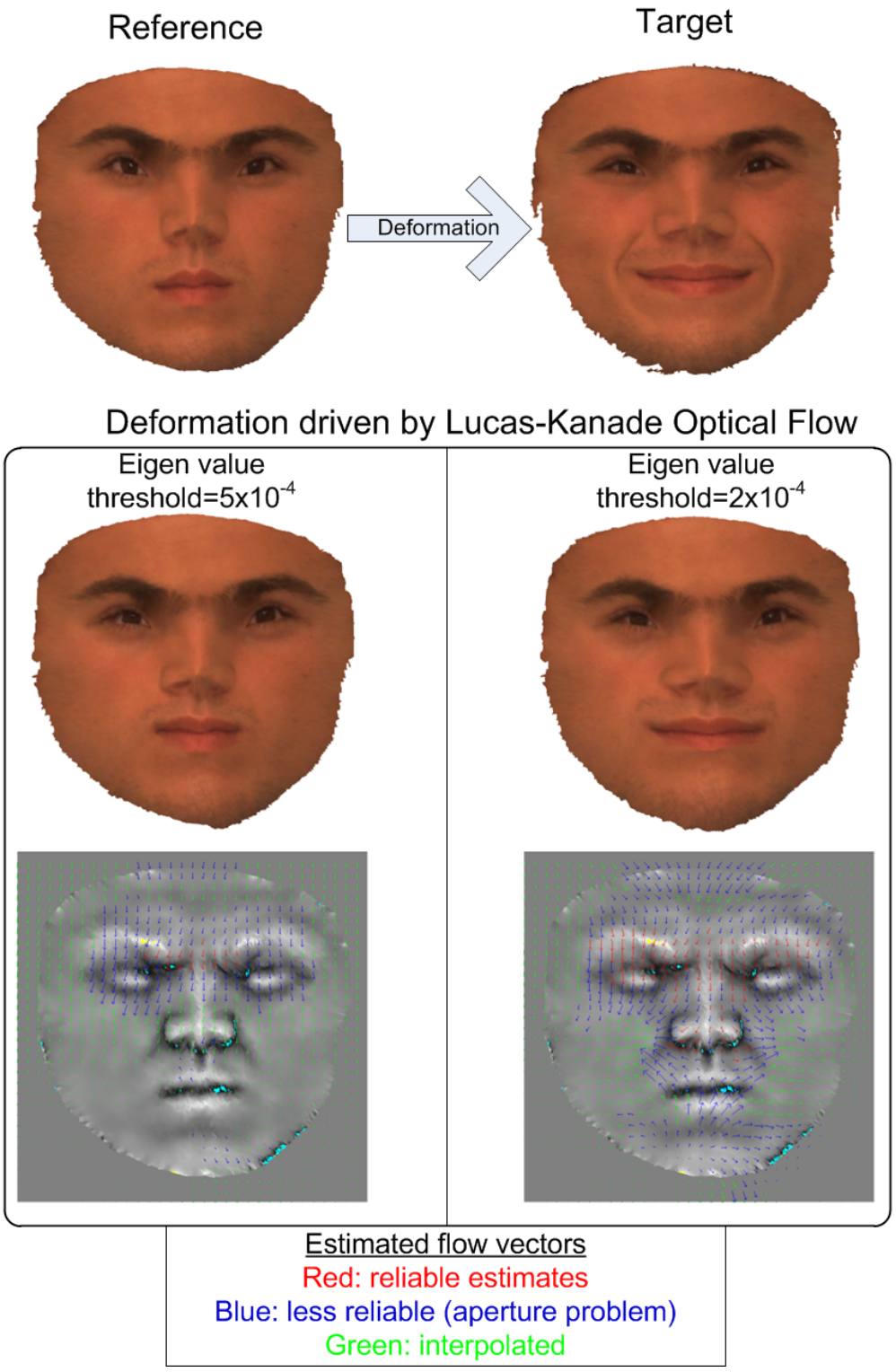


Figure 5.6. Registration of a neutral face via Lucas-Kanade optical flow applied on texture and mean curvature images. Flow vectors, which are estimated at the coarsest level (32×32 pixels) are shown over 512×512 curvature images.

correspondence estimates from the optical flows, they build an initial coarse linear model of 3D deformations and texture variations. This model is then fitted to novel faces to initialize optical flow estimation stage which is used for finer registration. Having obtained new registrations the morphable model is updated and the same steps are repeated until all database faces are processed. However, this is not a generic solution and computationally demanding due to the three stage procedure (optic flow estimation, smooth interpolation, model construction and fitting). Also, it is not fully automatic due to the human intervention which is especially crucial for the initial registrations. Moreover, the Lucas-Kanade algorithm does not impose injectivity (one-to-one) on the estimated mappings, which is necessary for invertible mappings to avoid artifacts, unlike our technique.

5.5.2. Surface Registration Examples

A direct application of surface registration would be automatic 3D face morphing. Some examples are shown in Figure 5.7 and in Figure 5.8. In Figure 5.7 a face belonging to one identity is morphed to another (both from the BU-3DFE database [96]). This figure also shows a neutral face morphed to smiling (both from from the Bosphorus database, Section 3.1). In this figure, textures of reference surfaces is mapped to that of deformed faces for comparison. One can observe that facial features like eyes, eyebrows and lips have good matching accuracy. Also, the overall deformations are smooth. This can be best observed in the middle face which depicts the average of the reference and target faces. Further results are given in Figure 5.8 based on faces from Bosphorus database, where one reference face is registered to different subjects and different expressions. Again one can see that the features are matching well.

In these experiments, surface images were discretized onto 512×512 resolution images, and four-level image pyramids were employed. For the image matching, usually equal importance was given to texture and shape channels; but better expression registration with texture and better identity registration with curvature channels could be obtained for some of the faces. It was observed that after convergence at coarser levels, deformations at the finest level (512×512 pixel resolution) were insignificant,

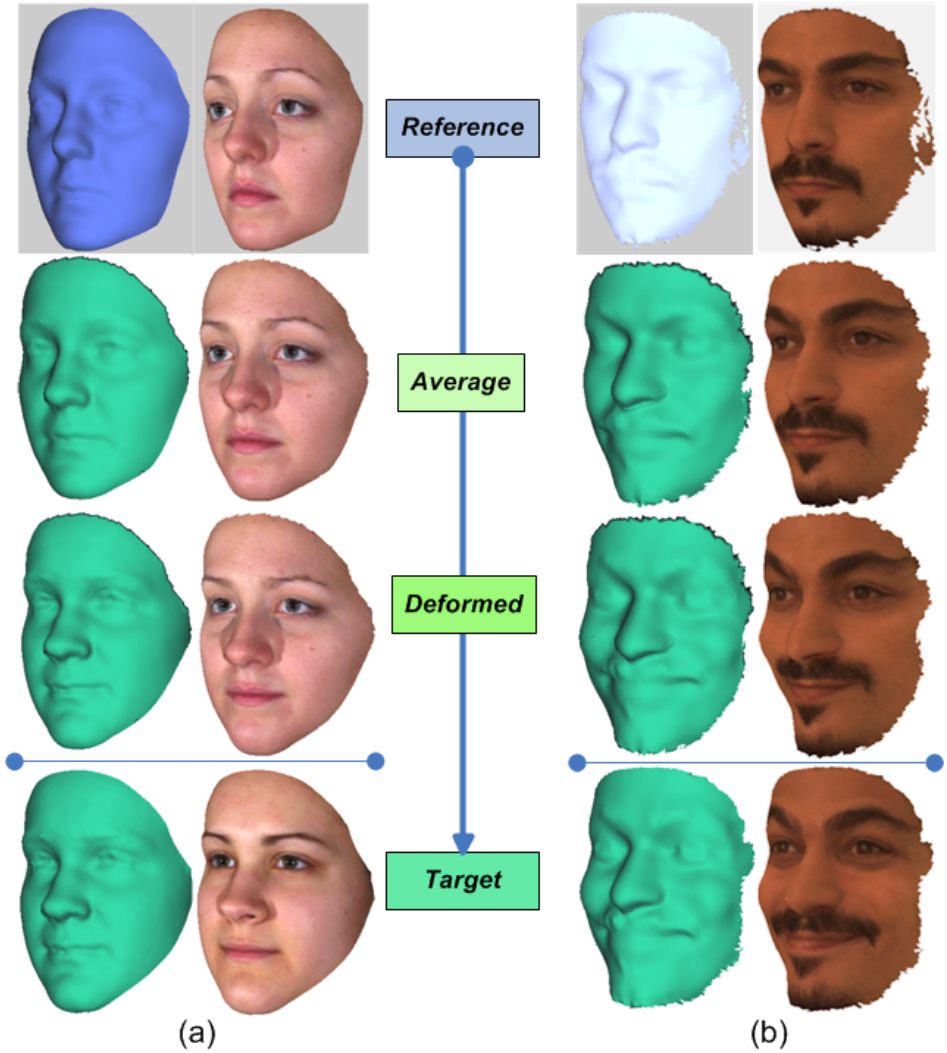


Figure 5.7. Non-rigid registration results of faces of (a) different subjects and (b) of a smiling face using deformable mesh-based registration. Texture of the reference faces are transferred to the deformed faces. Halfway deformation results are displayed as well.

making this finest resolution level redundant. Actually this is a rather expected result since major facial features have low frequency characteristics.

The algorithm is quite efficient. Many of the heavy calculations, like resampling for the reference meshes, are done in the initialization step, and iterations are mostly performed on the coarser resolutions. Except for the mappings between 3D and 2D, the computations do not depend on the 3D mesh resolution. In the experiments with 512×512 resolution images, the number of iterations till convergence were in between 50 and 200, and total registration duration (for mappings and energy minimization)

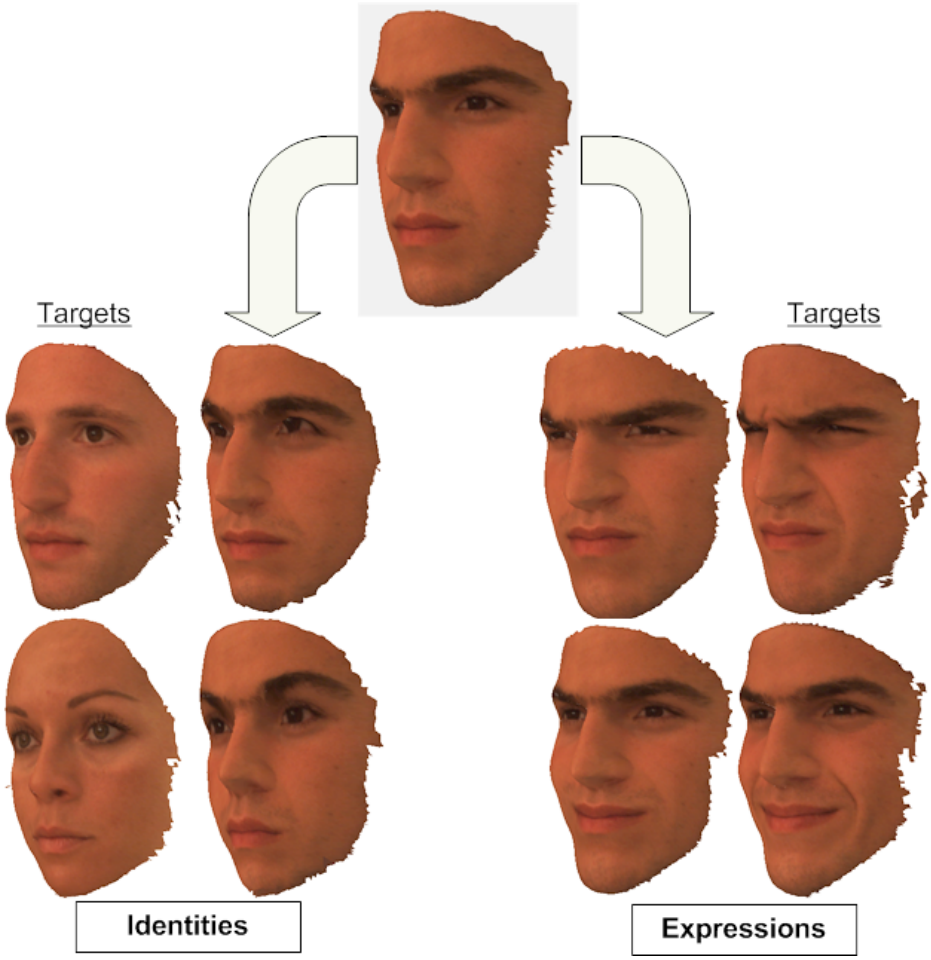


Figure 5.8. Registration results of a reference face to faces of different subjects and with different expressions. The faces at the leftmost and rightmost columns are the target faces. Texture of reference is transferred to the deformed surfaces.

changed from three to seven seconds using single core of 3.0GHz Core2Duo processor, and 2D image registration duration was about one second.

5.5.3. Registration of Different Identity and Different Expression

Identity and expression variations exhibit different characteristics on 2D images (whether curvature or luminance images) such that their registrations are actually not identical problems. First of all, identity differences show themselves on permanent features (see Section 2.2), e.g., nose, lips and eyes. Both their relative positioning (physiognomy) and their individual shapes vary among people. Permanent features are affected from expression related deformations as well, for example, lip corners

are pulled by smiling. However, expressions also give rise to transient features (see Section 2.2): for instance, mouth cavity appears when mouth is opened or furrows are revealed during smiling. Also note that some transient features like furrows and wrinkles can become slightly permanent due to aging. When we compare the effects of the identity and expression factors the obvious observations are as follows:

(i) Expression variations are local while identity variations are holistic;
(ii) Due to the transient features, expressions can cause non-corresponding local regions between the images. On the other hand, when only difference is the subject identity full correspondences exist.
(iii) Many expressions can cause considerably larger deformations on permanent features than the identity variations depending on the expression and its intensity, or conversely can cause considerably smaller deformations due to the same factors. An example of the first case would be deformation of the lips, and to second case would be eye lid opening.

From these observations we can deduce that correspondence estimation between certain expressions of a person could be more difficult than the registration of different subjects having same expression. This is because we can encounter non-corresponding regions due to the transient features, which is in contradiction with the objective of the registration, since objective function do not take into account such differences. Also, large deformations are more difficult to handle.

When we compare luminance and curvature data modalities, we also expect some differences. Because curvature measures the local surface bending, the deformations due to expressions may result in high curvature magnitudes, which in turn yield new surface features on 2D maps, i.e., new transient features. We can mention the bulge of the cheeks to this situation. On the other hand, face texture of the same identity can provide more reliable similarity measures, especially due to lips, facial hair and eyes. Conversely, luminance data can be more problematic for correspondence estimation between different identities and for some expressions that deform permanent features since albedo and lighting variations on the facial surfaces can change a lot

while curvature data exhibit higher degree of similarity.

For these reasons one idea could be to use separate correspondence estimators that are tailored to identity and expressions variations, for instance by using different data modalities or different regularization and image matching cost functions. However, in our problem, we cannot use different estimators. This is because both sources of variations are effective since we want to develop person-independent expression analyzers.

Figure 5.9 shows registration of curvature and texture images of smiling faces under different rigidity values onto a neutral face. These registrations were performed on 96×96 resolution images, and using three-level image pyramids. The image registration durations were about 0.1 seconds. In figure 5.9, we see from the curvature image registration that with decreasing rigidity (or increasing elasticity) lip corner pulling is captured better hence can be normalized. Also, the narrowed eyes of the smiling faces are opening. On the other hand, we see that use of texture images are not good at estimation of the deformations, especially for the example face with facial hair. This is because image structures for the matching objective (Equation (5.6)) are insufficient unlike curvature data, and face textures of different identities cause non-correspondence issues, for instance the moustache.

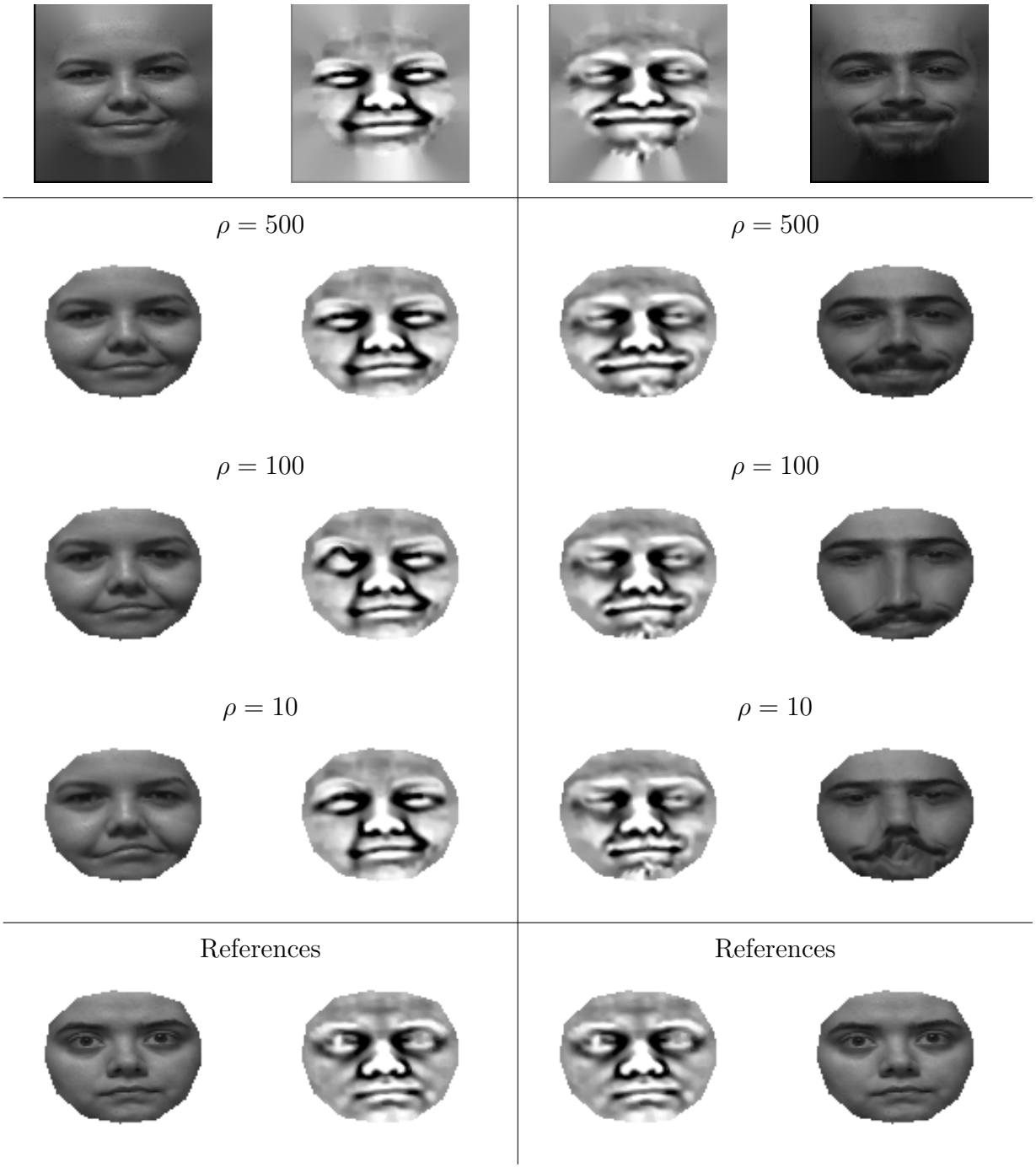


Figure 5.9. Registration of curvature and texture images of smiling faces under different rigidity values (ρ) onto a neutral face.

6. REGISTRATION-BASED DATA-DRIVEN EXPRESSION ANALYSIS

For the problem of person-independent facial action detection, data-driven methods that learn relevant information from available data without relying on any a priori human face model have achieved state-of-the-art results, as reported in Chapter 4. Model-driven methods are still currently effective in handling head pose variations and differences in physiognomy, since they are guided by face models and thus do not need to acquire all the information from data. On the other hand, data-driven methods are advantageous since they bypass tedious model construction stage, eschews errors of the model fitting and are not biased by the assumed face models. Their shortcoming, however, is that they do not achieve as fine face registration as the model-driven methods. Any improvement in alignment accuracy may benefit the outcome of data-driven methods.

In order to profit from the superior registration inherent in model fitting, we could combine both approaches, that is, we could normalize the faces by a model fitting approach and then proceed with data driven methods. However, we would be still confronted with all the drawbacks of model-based analysis. In this chapter we present an improved registration-based on establishing non-rigid matches between pair of faces without recourse to any model information. The main idea is that, instead of aiming at accurate estimation of correspondences, inputs are matched to the expression specific references and the decision is based on the goodness of matching. In other words, we do not find the best registration onto a common reference, but learn the matching patterns of inputs to the predetermined expression references. We show that it is possible to obtain a performance higher than the state-of-the-art methods without using any appearance-based feature extraction methods like the popular Gabors. We apply this registration approach to the problem of facial action unit detection on 3D faces.

This chapter is organized as follows. Since we are working on the 2D surface curvature maps, we first explain 3D facial variations portrayed on the curvature maps in Section 6.1. We then present the proposed approach in Section 6.2. We discuss several methods based on the proposed approach in Section 6.3, and present an extension by using multiple references in Section 6.4. Finally, we give and discuss the experimental results in Section 6.5.

6.1. Analysis of 3D Facial Surface Deformations via 2D Curvature Fields

Before introducing our registration-based detection approach, we want to clarify the effects of surface variations on 2D curvature maps since we are not performing registration in 3D space. We could implement registration-based data-driven expression analysis in 3D space. However, with the goal of analyzing 3D facial surface deformations due to expressions, we work on their 2D maps due to the computational benefits of 2D space and to avoid dependence of surface mesh topology and resolution. To this effect we use mean curvature maps since they reveal local surface bendings and thus render measurement of surface deformations effective. Facial variations portrayed on curvature fields and its advantages are discussed elaborately in Section 4.4. Moreover, its usefulness in the AU detection problem has already been shown in Section 4.9.1 where it was observed to be the top performing one among all types of geometrical representations.

Differences on face surfaces, whether due to permanent or transient face features (see Section 2.2), can be thought as 3D spatial deformations of a surface from a reference face surface. By mapping 3D surfaces onto 2D curvature fields, we can describe the actual 3D spatial surface deformations in terms of 2D spatial deformations and curvature changes. Notice that this conversion is not invertible due to the curvature representation of 3D geometry, though this is not a concern since we do not intend to reproduce the 3D faces, but only to analyze them.

6.2. Registration-based Data-driven Recognition

We aim at non-rigid registration onto references specific to expressions rather than onto a common reference. In contrast to non-rigid registration implemented by model-driven techniques, no facial shape information is assumed in this approach. We discuss the relationship between estimated correspondences and shape variations in the following subsection. This discussion is important to understand the way our approach achieves identity normalization and expression analysis. In Section 6.2.2 we present the details of the proposed approach based on expression specific references, and explain the way to determine these references in Section 6.2.3.

6.2.1. Do Estimated Correspondences Represent Shape Variations?

In Section 5.5.3 we have investigated correspondence estimation between different identity and expression images by non-rigid registration. Correspondences are required so that we can extract shape variations that contain useful information about identity and expression. Although this appears reasonable, we wonder whether estimated correspondences really represent shape variations. Before answering this question, let's review relevant points in registration.

In landmark-free non-rigid registration of an input surface or image onto a reference, generally there are three basic factors that determine the relationship between estimated correspondences and shape. The first one is our reference image, since correspondences explain shape variations with respect to the reference face. The second one is the employed image similarity metric. This metric characterizes the type of similarity accounted for on the correspondences. In this study we use sum of square distances as given in Equation (5.3). This metric actually implies modeling the differences between input image and the reference by a zero-mean Gaussian noise. The last factor is the deformation model. The estimated mapping functions $\boldsymbol{\phi}$ are restricted to given deformation models. In our study we enforce deformations to be hyper-elastic by minimizing the deformation energy given in Equation 5.14. This model provides diffeomorphic mappings since resulting deformation is differentiable, smooth and injec-

tive, but also quite flexible (more flexible than conventional elastic models). There are also other factors related to the implementation of the registration model. Issues like discretization, approximation and optimization techniques can affect the estimation results since we are often confronted with suboptimal solution risks (local minima in cost minimization) in non-rigid registration problems. However, since these issues are related to implementation, we ignore them in this discussion.

A condition to ensure that estimated correspondences represent shape variations is the image matching condition. In our case we hypothesize Gaussian noise model for matching residuals, and as a consequence, our image matching condition is that, pixel intensity differences between the reference and the registered images must be small all over the reference domain. Note that, this condition may not be satisfied well if we use luminance instead of curvature images since brightness of images can be severely modified by illumination effects. In this case there would be more sensible choices, such as the normalized cross-correlation metric or mutual information measures. Physiognomical features of the face satisfy the image matching condition to a large extent, so that extraction of 2D spatial shape variations becomes possible. On the other hand, existence of transient features of expressions like bulges and mouth cavity (see Section 2.2) means we have to confront with missing correspondences since these structures will not always be available. During pairwise registration, missing of correspondences in one of the images can mislead the deformations and therefore can cause local distortions on estimated shape deformations. Nevertheless, even though expression specific non-corresponding regions are problematic for non-rigid registration, they become useful for recognition in our method, as will be explained in Section 6.2.2.

However, even if physiognomical features provide clues essential for correspondence estimation of the shapes, unless the deformations can be adequately modeled by the adopted deformation model, we cannot obtain plausible shape deformations. Hence choice of deformation model is the second important consideration. Our deformation model is highly flexible and its rigidity is adjusted by the parameter ρ given in Equation (5.19). For instance, registration of smiling lips onto a neutral face under varying elasticity is shown in Figure 5.9. We see that lips can be neutralized totally

only when we decrease the rigidity. However, allowing extremely deformable models may also cause implausible estimation of shape variations due to the ill-posed nature of the problem, and second, disturbing effects of small mismatch regions and noise cannot be suppressed adequately. In other words, a more flexible model means a higher risk of unrealistic inconsistencies on the shape estimation. An example to this situation is seen in Figure 5.9, in the left column with curvature images when $\rho < 500$.

We can now answer the question we had posed, "Do estimated correspondences represent shape variations?", as follows: If our reference image resembles to the input image according to the employed similarity metric and the non-rigid deformations can be sufficiently mitigated by the deformation model, then we can achieve accurate estimations of shape variations. However, if the amount of non-similar regions increases and actual deformations do not obey to the deformation model or are larger than the allowable transformation range, the shape estimation will not be reliable.

6.2.2. Registration onto Expression Specific References

When we examine the previous registration-based methods (see Section 2.5), we see that they are all model-driven and they first extract motion describing the change in shape and pose with respect to a reference frame, which is either subject's neutral face or estimated mean face shape, and then use the deformation data in classification and/or for the normalization of the face appearance. Instead of estimating deformations with respect to a reference common to all faces and expressions, in our approach we use references specific to expressions. Our approach is depicted in Figure 6.1. Given a reference image I_k, for a matching similarity metric and a deformation model, we estimate a mapping from the domain of the reference D_k to the domain of the input image D_i. This estimated deformation is denoted by $\boldsymbol{\phi}_{\mathbf{ki}}$. Thus, we warp the input image by resampling according to the inverse transformation $\boldsymbol{\phi}_{\mathbf{ki}}^{-\mathbf{1}}$. The resulting registration onto the reference is denoted by $I_i \circ \boldsymbol{\phi}_{\mathbf{ki}}$. We then perform expression analysis on either the registered image or on the normalized deformation field, as explained in Section 6.3. This approach can also be directly implemented with direct 3D surface registration techniques, i.e., without performing any 2D mapping. Notice that

when directly using estimated deformation fields for recognition, we normalize them in order to refine the pose estimates according to the estimated correspondences, because initial pose estimation methods, like landmarking or ICP algorithm, are actually based on few and/or less reliable correspondences than the non-rigid registration can provide.

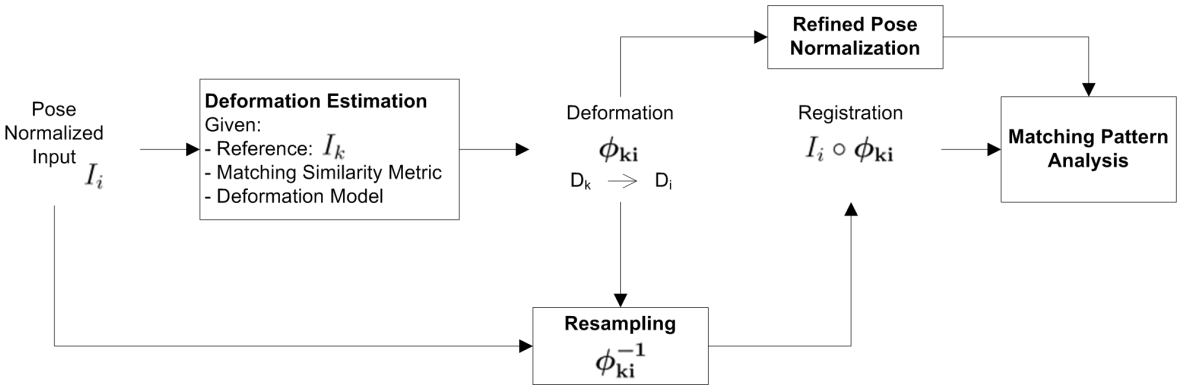


Figure 6.1. Block diagram of non-rigid registration-based data-driven expression analysis. A mapping $\boldsymbol{\phi_{ki}}$ from domain of a given reference D_k onto domain of an input D_i is estimated to realize non-rigid registration for matching pattern analysis.

All deformed, that is, non-rigidly registered input images become more similar to the reference. This means that they resemble in identity and in expression to the reference up to the extent allowed by the chosen deformation model. Compensation of identity differences is helpful since differences in physiognomy can confound expression analysis. On the other hand, expression normalization may cause us loss of some important information. For instance, imagine how we can decide whether a person is smiling or not, if we pull back the smiling lips to a neutral state. Though there would still remain information about the smile deformation in the curvature data, in the absence of spatial deformations the discrimination power of a smile detector may be diminished.

One of the benefits in using expression specific references is to cope with these situations, because if we register pulled lips onto a reference with similar lips, that is shape deformation, we can still preserve the essential lip shape. However, we should also prevent pulling of neutral lips all the way to smiling lips. This can be done

by decreasing the elasticity of the model so that we limit the extent of deformations. Figure 5.9 shows two examples, which are both registered using curvature or luminance data. From the curvature registrations we see that with the lowest rigidity value pulled lips are neutralized while with the highest rigidity the lip shape is preserved. Consequently, by using expression specific reference, we in fact implicitly incorporate shape information in the analysis.

However, a question that may come to mind if do we achieve full subject identity normalization when we limit the allowable deformations. Our experiments show that normalization suffers, however, not very seriously. In contrast to expressions that remain local, identity induces differences over all the face resulting configurational changes (relative positioning of face parts) as well as in the shape of face parts. Since the face parts always have distinctive characteristics with relatively high surface curvatures in larger scale, their registration is achieved even if we make deformations more rigid. For instance, in Figure 6.2 registration of different identity and expression faces onto neutral and smiling references are illustrated. These registrations are obtained with $\rho = 500$ (Equation (5.19)), that is, with a quite rigid model. Notice that the identities of the references is of the same subject. In the last column differences between the warped and original inputs are shown, and these difference images indicate the motion of normalization. The first row illustrates registration of neutral face onto another neutral. We see slight traces of facial parts. However, in the second row which belong to neutral face of another subject, the local deformations are as follows: mouth is moving upward, lips are bending, nose is shrinking in both horizontal and vertical directions, and eyebrows and eyes are moving downward. This is as a result of differences between the subjects due to the changes in physiognomy and facial part shapes. Notice that the same motion exists in the rest of examples as physiognomy of subject C is matched to the physiognomy of A in the rows 3-5. This means that we can still normalize the identity of the subject even in the presence of expressions in the reference or the input. However, when we examine the smiling faces we also see some expression normalization motion. This also means that the effort to compensate for differences in physiognomy could also wipe out spatial deformation component of lower intensity expressions. Therefore, there is a trade-off between suppressing identity effects and

preserving shapes of lower intensity expressions controlled by the rigidity parameter. Setting this parameter to low rigidity values may handicap detection of some lower intensity expressions if we try to recognize expressions only from estimated deformation fields. Actually, this intricacy is also valid for the human brain: without seeing a neutral face of a person, it is often quite difficult to recognize subtle motions on the face, or conversely for instance, some neutral faces appear as if they are smiling. FACS [19] also mentions this situation and it is why human FACS coders are recommended to work in a subject-dependent manner, i.e., by encoding AUs relative to the neutral face. Nevertheless, by using surface curvature data we still preserve substantial deformation information that helps recognition of low intensity expressions.

The second benefit of expression references is related with the transient features of expressions (see Section 2.2) since certain transient features appear only with certain expressions. Expressions give rise to transient features especially in the curvature data; furthermore the extent of these transient features increase proportional to the strength of the expression, as in the example of furrows in smile. Transient features may become a problem for non-rigid registration if they do not exist on the reference. In our approach transients are not problematic even though they effect registration, in fact, resulting local mismatches provide informative clues for expressions and/or AUs. The regions containing expression specific features in the positive samples will fit well to the reference while negatives will not, resulting in low and high residual values respectively.

In conclusion, the approach of using registration to expression specific references can be considered as an unsupervised technique of incorporating shape into the expression recognition since shape is not learned by landmarking. Shape information is employed implicitly by evaluating non-rigid similarities between the input and the reference images. Thus, we can design subject-independent and data-driven expression analyzers which are robust against ambiguities arising from alignment errors and non-rigid shape variations not related to expressions as well as effectively capture transient features of expressions.

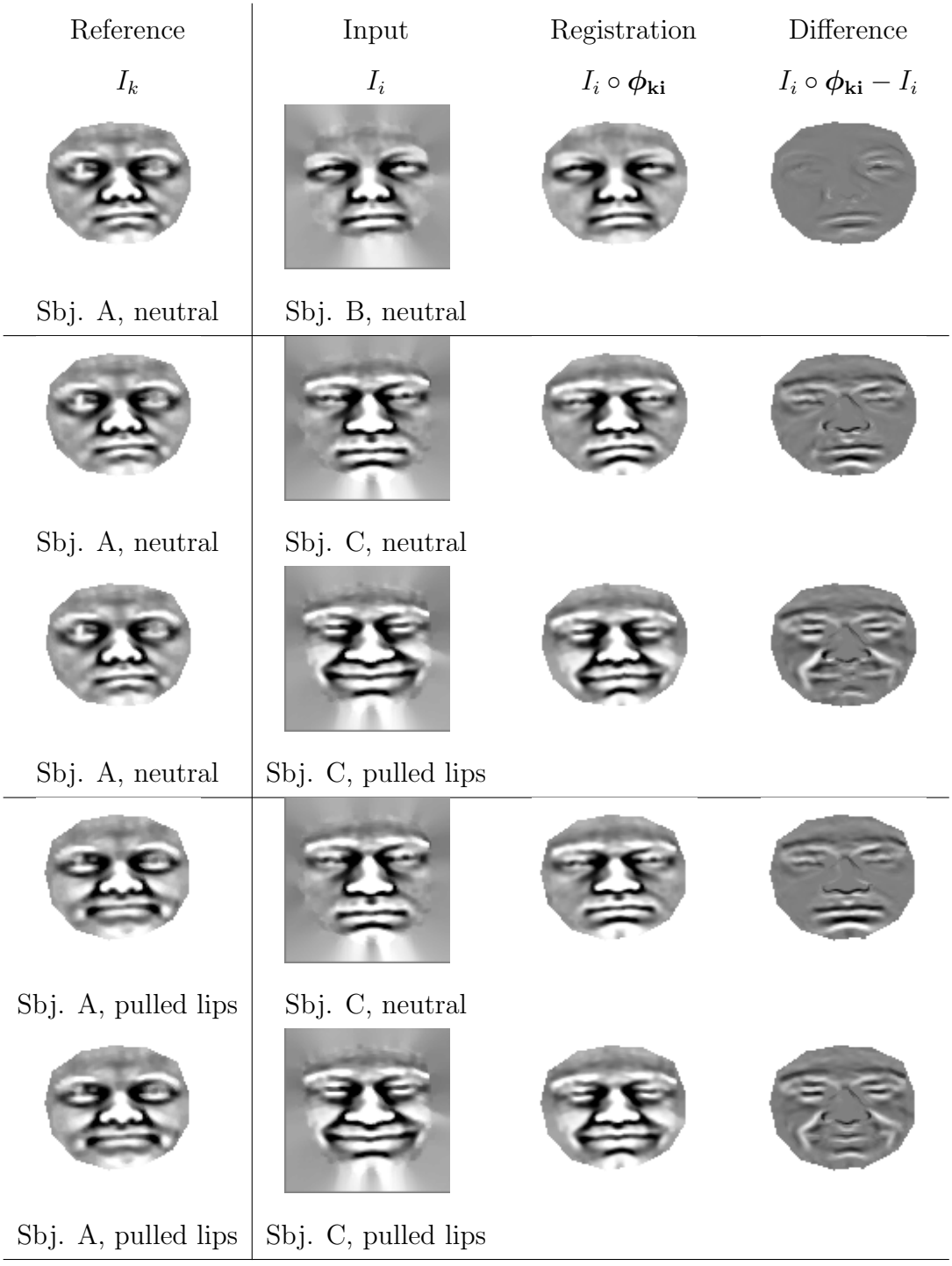


Figure 6.2. Registration of different identity and expression faces onto neutral and smiling references ($\rho = 500$). The images at the last column show the differences between the registrations and inputs where gray tone denotes zero difference. Through these difference images we observe how identity and expression related deformations are removed.

6.2.3. Expression Specific References

Since reference images lead the deformations of input images and thus control the appearance of the registered images, depending on the employed reference we may obtain different success in recognition. Therefore, an important point in this approach is to determine expression specific reference images. We can select a sample from the training set to be used as a reference, or instead, we can estimate a mean image for each type of expression so that the analysis will not biased by the choice of a single sample.

Cootes et al. [97, 98] use an iterative method to estimate the mean face for diffeomorphic statistical shape model construction of neutral faces. However, this approach still requires a manual reference choice for initialization and since the registration method requires that all the other samples deform toward the initially chosen reference, the result will be biased. Charpiat et al. [99] suggest an optimal way by proposing an objective function that involves joint diffeomorphisms of all the images instead of minimizing objective functions of pairwise registrations. Thus the mean is calculated at the last stage rather than by iterative updates during the groupwise registration. Since groupwise registration is not established by pairwise registrations onto a fixed reference, the resulting diffeomorphisms and consequently the mean will not be biased towards the chosen initial reference. A drawback of this method is a heavier requirement of computational resources in proportion to the number of individuals.

An important concern for using the mean images is that, high degree of the variability in expressions may cause blurring effects on the mean images and relevant image structures related to the expression can be smoothed out and crucial details may be lost. However, expressions, especially the AUs, necessitate preservation of local details, and because more than one AU can present different local characteristics due to the interactions with other AUs, averaging over different modes of the AUs will not be appropriate.

In view of these drawbacks, we subjectively determine the reference images from

training samples. We select references from the most typical examples for each expression, i.e., the ones that clearly include the most important characteristics of the expression or expression component. In Section 6.4 we propose the use of more than one reference for expression recognition.

6.3. Recognition through the Matching Patterns

To recognize expressions, we train classifiers which learn matching patterns between the samples and expression references. Images that are warped to match with an expression reference will bear differing patterns depending on being a positive or negative sample image of that expression. We expect to observe positive samples fitting well, and negatives samples poorly. As explained in Section 6.2.2, this is because local expression features are specific mostly to an expression and are not expected to be present in most of the negative images. In the following subsection we describe a method for local analysis of AUs. Section 6.3.2 explains recognition through the matching patterns over curvature fields and rationales behind using different classifiers. Finally, use of deformations fields is described in Section 6.3.3.

6.3.1. Local Analysis by RoI Masking

In this method we use region of interest masks per AU (RoI masks) to test the hypothesis of the presence of that AU. Since AUs happen locally, we filter out irrelevant facial regions by using these RoI masks. Thus, once the face is registered, only face regions under the RoI masks are examined.

We have designated a RoI mask for each monotype AU without any non-additive combination over its reference image. These masks were prepared by manual delineation based on the knowledge of each AU as follows: First, sample face images only containing that AU (other AUs that do not significantly change the target AU deformations are allowed) are registered onto the reference of the respective AU by non-rigid image registration, and an average curvature map is computed. By comparing this averaged AU specific curvature map and the averaged neutral map visually, and based

on our intuitive notion of that AU, we crop the most differing region(s) to generate the RoI mask. Figure 6.3 depicts this process for monotype AU 12 (Lip Corner Puller). First, samples of this expression are registered on the AU 12 reference and then its mean over all the registered samples are estimated. Similarly a neutral reference face is generated. Finally, a mask is cropped by visual comparison of averaged AU 12 and neutral images. Examples for some of the other mean AUs and their masks are shown in Figure 6.4, where the left image for each AU shows the actual image from a subject; the right image shows the estimated mean image obtained from the samples of that AU, viewed through its corresponding analysis mask. The analysis masks delineate AU activation regions. To train an AU detector, all positive and negative class samples for a target AU are first registered onto its reference and then masked, so that the local matching patterns of these two classes can be learned. This procedure is depicted in Figure 6.5. Thus, our AU detector is composed of three stages: First, an incoming face image is registered to the AU specific reference in a non-rigid fashion; next, warped face is masked to perform local analysis; finally positive and negative hypotheses are tested for a decision.

6.3.2. Matching Pattern Analysis Over Curvature Fields

For classification we use four generative and three discriminative classifiers. The generative classifiers are the four types of Normal Bayes classifiers, which are nearest mean, Naïve Bayes, Quadratic Normal and Simplest Quadratic Normal. The three discriminative ones are AdaBoost, linear and RBF-SVMs. These classifiers and the procedure to train them are described in Section 4.6. Here, we discuss their use in the non-rigid registration framework.

For registration of an input facial surface onto reference, we minimize the matching energy E_M between their curvature fields

$$E_M(\boldsymbol{\phi}) = \frac{1}{2} \int_{\mathbf{p} \in D_A} \Big(I_i \circ \boldsymbol{\phi}(\mathbf{p}) - I_k(\mathbf{p}) \Big)^2 \, \mathbf{dp}. \tag{6.1}$$

as described in Section 5.1. After registration, discretized matching energy corresponds

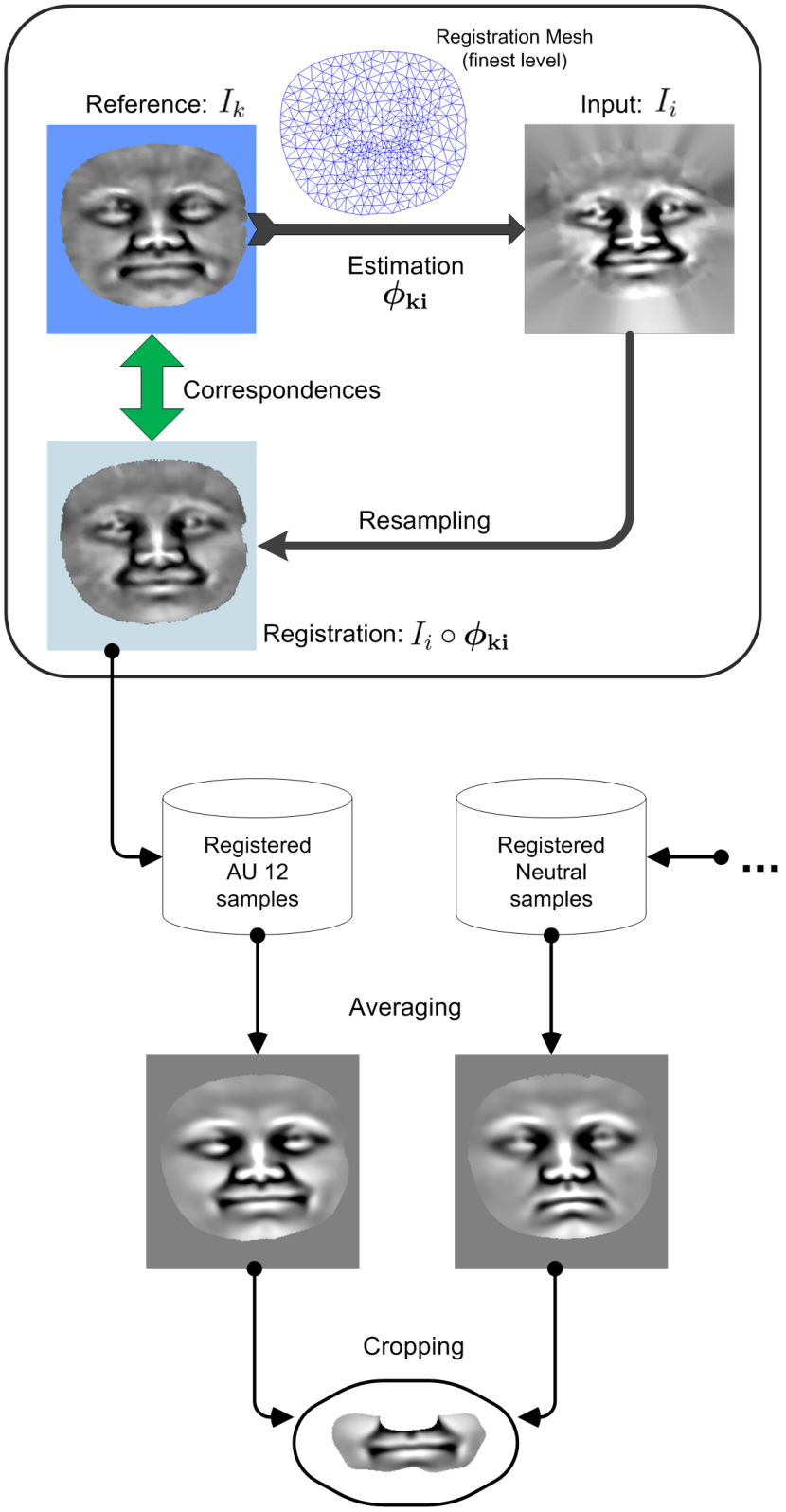


Figure 6.3. Preparation of a RoI mask for monotype AU 12 - Lip Corner Puller. Average of the registered AU 12 samples are computed and then visually compared with averaged neutral samples to crop a mask.

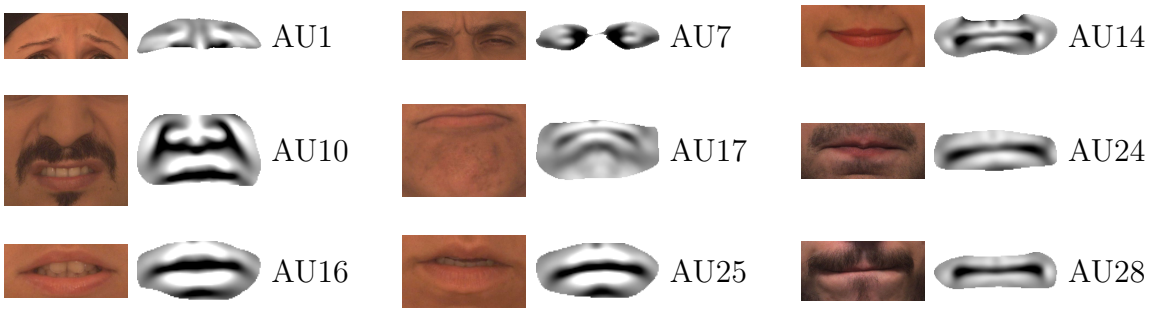


Figure 6.4. Average of curvature images (right) of selected AUs under respective analysis masks; and actual AU instances (left).

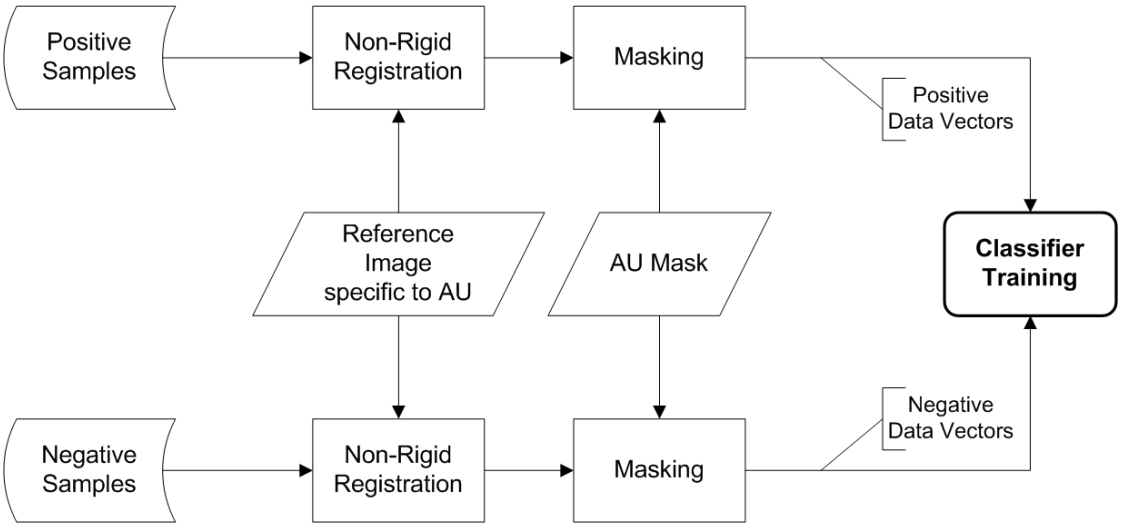


Figure 6.5. Training of a registration-based AU detector. In testing the same registration and masking is applied, and classifier decides whether the AU exist or not.

to matching error over image pixels as given below.

$$E_M = \sum_{x,y} \Big(I_i \circ \boldsymbol{\phi}_{\mathbf{ki}}(x,y) - I_k(x,y) \Big)^2 \tag{6.2}$$

The matching objective assumes that pixel intensities of registered images differ only by zero-mean homoscedastic Gaussian noise. Therefore, if we have N samples that are accurately registered and satisfy this assumption, the statistics below must be such that the sample averages are very close to the reference pixel values and they have

small variance over the image domain.

$$
\begin{aligned}
ave_k(x,y) &= \frac{1}{N}\sum_{i=1}^{N} I_i \circ \boldsymbol{\phi}_{\mathbf{ki}}(x,y) \\
var_k(x,y) &= \frac{1}{N-1}\sum_{i=1}^{N} \Big(I_i \circ \boldsymbol{\phi}_{\mathbf{ki}}(x,y) - ave_k(x,y)\Big)^2
\end{aligned}
\tag{6.3}
$$

If our reference is a typical realization of an expression or of an AU class, we expect that, these assumptions will be satisfied more easily over the related face regions for positive samples than for negative samples. Therefore, we can design classifiers under the Gaussian assumption for the registered images, or equivalently for the registration matching errors. Let's consider the nearest mean classifier on the registered images, which is the simplest Normal Bayes classifier. We can interpret this classifier as follows. It corresponds to using two generative models for the image matching errors: one for the positive and the other for the negative hypothesis. When we use a reference which bears image structures of positive samples, positive samples should generate lower matching error than negatives. Therefore, nearest mean actually classifies the matching errors of the positive and negative samples.

We can also incorporate pixel variances in the classification by estimating a single diagonal covariance matrix from pooled data of positive and negative samples. This corresponds to the use of Naïve Bayes classifier. Hence the net effect will be to weight the matching errors inversely proportionally to their estimated variances, so that less reliable pixels in the RoI mask will be down weighted. This is due to the assumption mentioned above, variances should be smaller for more accurately registered regions.

However, use of linear classifiers in this non-rigid registration-based detection framework may not be the most effective approach. This is because linear discrimination means weighting of the registered pixels will be identical for positive and negative hypotheses, which actually may not conform to our model. In fact, we use expression specific references so that similar deformation types fit properly, but those of other deformation types do not. Since we expect good quality matches to have small variance values, and conversely for the negative samples, we employ quadratic Gaussian

classifiers.

Sample mean and variance estimations (as given in Equation (6.3)) for dimple and mouth opening facial actions are shown in Figure 6.6. In mean maps, gray pixel denotes zero value and lighter the pixel is lower mean value. In variance maps, darker pixel values represent higher variance values. When we look into the mean estimates over the positive samples which do not contain non-additive combinations, we see that these average difference images are well localized, because they are calculated from similar expressions. On the other hand, for negatives differences are more spread. A case in point is the mouth region of the mouth open expression where local mismatches due to negative samples are spread over a larger region. In fact, sharpness of the pixel variance maps indicate the quality of registration. We first see that eye regions yield high variance, which means eye mismatches are frequent or the variations on the eye curvatures are quite high. Second, for dimple case the positive sample variance estimates over the lips are low, and conversely they are high on the negative estimate due to all other shapes assumed by the mouth in other expressions. We see similar effects for the mouth open case, shown on the right.

In Section 6.5 we report the detection performances of the two linear and two quadratic Gaussian classifiers in order to validate the assumptions made by these generative classifiers.

Discriminative classifiers, on the other hand, do not model the likelihood function, instead they try to learn the best discriminant function directly. For instance kernel SVM classifiers choose some of the training samples as support vectors to obtain a linear separator in a higher dimensional space than that of the data vectors. These support vectors take place along the decision boundaries and are assigned weights to optimally separate class distributions. In our problem, these support vectors are the registered samples. The observation that quadratic discrimination can perform better than linear discriminators, points out to the potential of non-linear classification. Since RBF kernels have the capacity to capture non-linearities in a wide range depending on the hyper-parameters, it may find better separators.

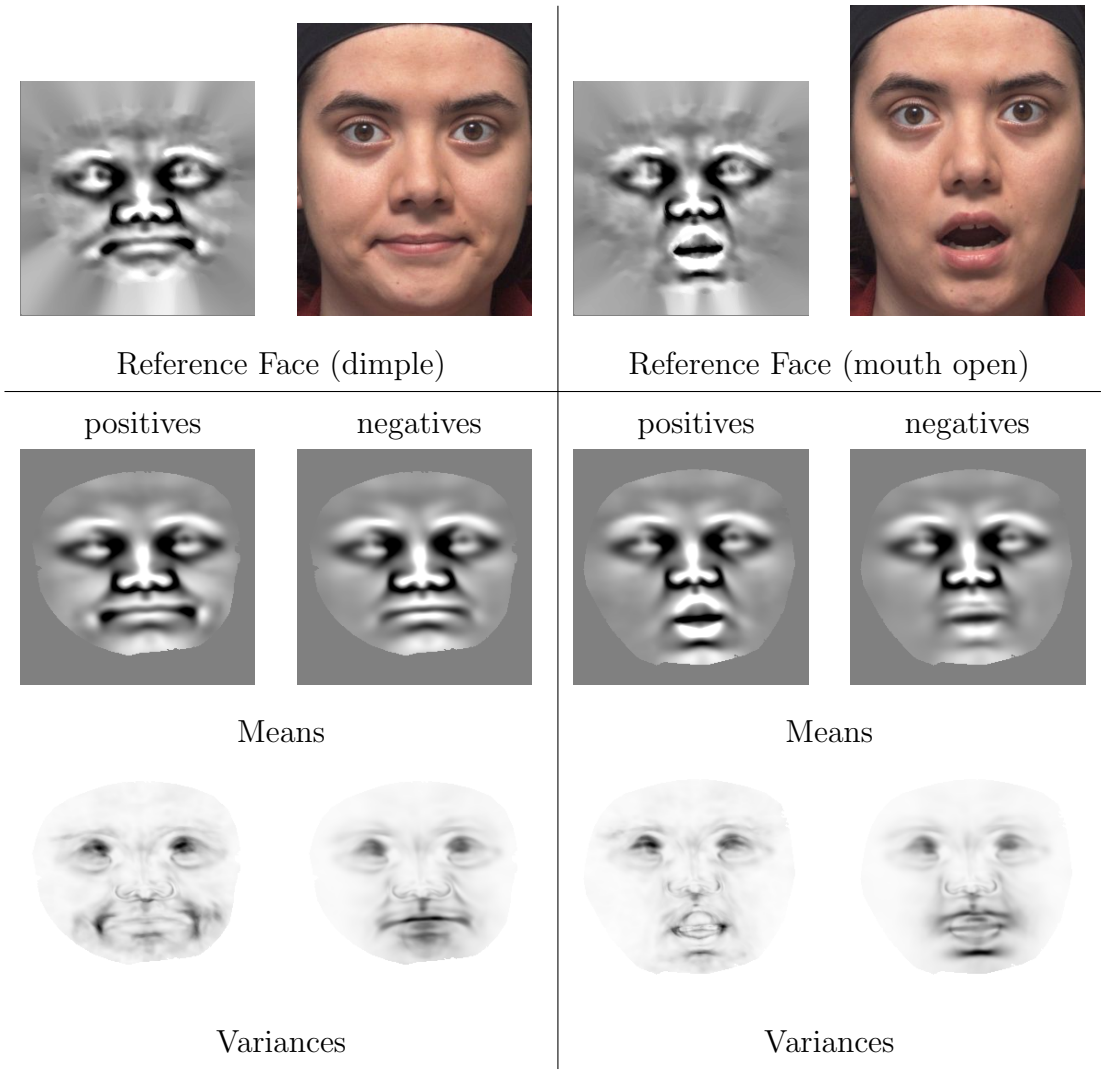


Figure 6.6. Mean and variance estimation over images that are registered onto dimple (left) and mouth open (right) reference face actions. Estimations are done over positive (only ones without non-additive combinations) and negative samples separately. In variance images black regions represent high variance pixels.

The third discriminative classifier is the AdaBoost algorithm, which can also be conceived as a feature selector (see Section 4.6). It uses weak classifiers which are weighted according to their discrimination ability. In our implementation, we have employed the nearest mean classifier as the weak classifier. When used as feature selector, AdaBoost selects the best discriminating curvature pixels over the registration reference domains. This means that we can also avoid the preparation of the AU masks and let AdaBoost select the effective pixels automatically. This allows us to investigate the whole reference domain exhaustively in the training stage rather than

committing ourselves to subjective chosen masks. In the experiments in Section 6.5.1, the performances resulting from the automatic selection of the reference domain pixels and from manual selection of RoI masks are compared.

6.3.3. Matching Pattern Analysis Over Deformation Fields

We can also perform analysis over the estimated deformation fields $\boldsymbol{\phi_{ki}}$, instead of registered images $I_i \circ \boldsymbol{\phi_{ki}}$. Estimated deformation fields contain facial shape information, though the accuracy of shape extraction depends on the registration as discussed in Section 6.2.1. Actually, estimated deformation field $\boldsymbol{\phi_{ki}}$ and the resulting registrations $I_i \circ \boldsymbol{\phi_{ki}}$ carry complementary information since registration means application of inverse mapping $\boldsymbol{\phi_{ki}^{-1}}$. For instance, estimated deformation field between a reference face with pulled lips and a neutral test face will possibly include motion pulling the lips back as well as deformations that removes identity differences. Conversely, after registration by applying inverse of this deformation, the lips of the neutral test face will be pulled and identity differences will be compensated. In our implementation, we can control the extent of deformation employed by the rigidity parameter ρ in Equation (5.19). We increase the value of this parameter, if we want to limit the extent of deformations generated by the hyper-elastic model which is explained in Section 5.1.2.

In our registration technique, the degree of freedom of a deformation is determined by the registration meshes, which are automatically generated according to images as described in Section 5.2. Therefore, rather than resampling deformation fields on the image grids, we use displaced mesh vertex coordinates as features for classification. Thus, dimension of the deformation features are typically much less than the registered curvature features. In our experiments we use 96×96 size images, and typically, while face domain size is about 5000 pixels, the number of mesh vertices is about 100. In Figure 6.7 we show an estimated deformation field example, by registration with $\rho = 500$. We see the deformation of a reference with open mouth toward an input face of another subject with closed mouth, as a motion field making the lips more closer. The estimated motion is better seen by the superposition of the deformed and the initial meshes.

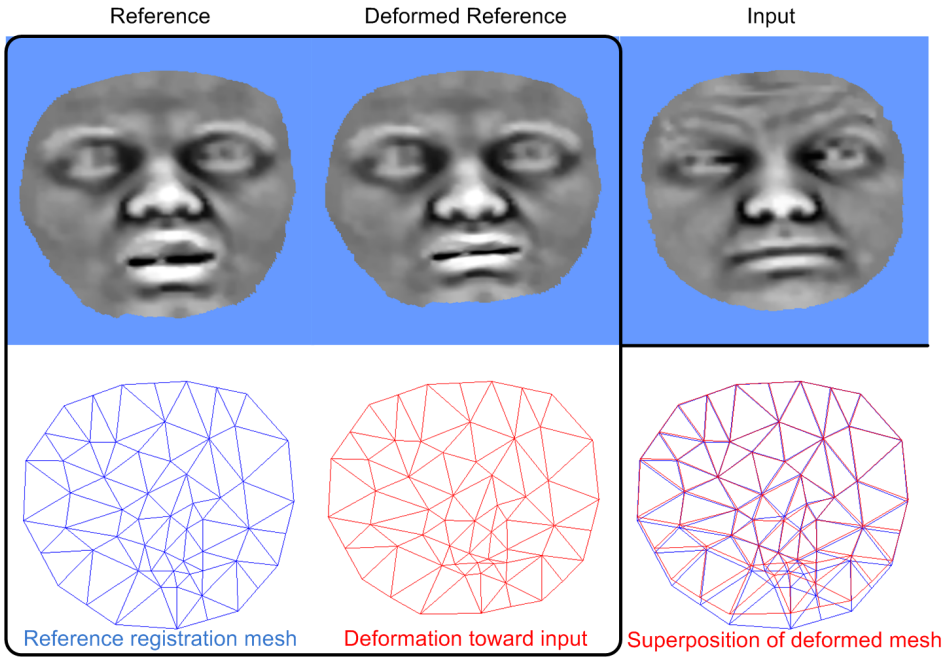


Figure 6.7. Deformation of a reference with open-mouth toward an input face of another subject with closed-mouth. Registration is performed with $\rho = 500$. Superposition of the deformed and the initial mesh shows the estimated motion.

We have to use deformations only after normalizing (similarity normalization) in order to minimize confounding effects of pose. Initial alignment done by techniques like landmarking or ICP algorithm are actually based on few and/or less reliable correspondences than the non-rigid registration can provide. Therefore we are likely to observe significant residual pose effects in the estimated motion field. These effects can disturb the recognition task. For this reason we estimate and compensate for translation, rotation and scale differences. These transformations are solved by linear least squares using the estimated displacements of the mesh vertices.

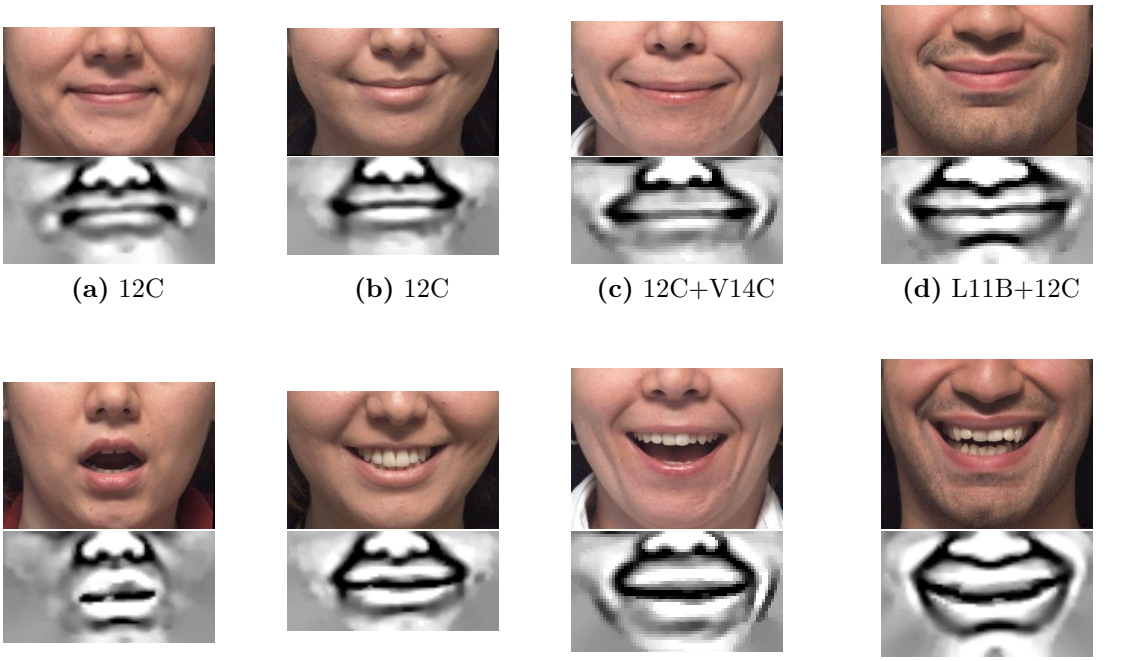
Otherwise, we employ the same seven classification techniques as described in the previous subsection, i.e., feature selection by AdaBoost and then recognition with Gaussian and SVM classifiers. Finally, we consider fusion of the deformation fields with registered curvature fields to exploit any complementary shape information. For this purpose, we apply AdaBoost on the concatenated feature set of both components.

6.4. Multiple Reference Scheme

A difficulty in AU detection problem is high within-class scatter due to the existence of different types of deformations, not only for a negative class but also for a positive AU class. The registration-based approach offers a convenient way to deal with this issue.

The variations among the samples of an expression or AU class can be too much to be adequately modeled by non-rigid transformation to a single expression reference. This may limit the performances of detectors due to increased correspondence problems and registration mismatches. A typical example of this situation occurs when the AUs appear in non-additive combinations. While in additive combinations characteristics of individual AUs are preserved, non-additive combinations can considerably alter the characteristics of the combining AUs. In fact, FACS separates additive and non-additive AU combinations. FACS furthermore defines a new set of features for each such combination. This is why some authors treat certain AU combinations as a separate class [33]. However, this approach is feasible only for a small number of specific combinations. By employing multiple references that bear different characteristics of an AU, we can obtain different registrations each of which may facilitate analysis of different types of an AU, because certain AU samples will locally better fit to the certain references. Thus we may alleviate within-class scatter problem.

Each such reference will be representative of a subset in an AU's instantiations. For example, let's consider detection of AU 12 (Lip Corner Puller), for which in Figure 6.8 four samples without non-additive effects are shown. It is quite likely to encounter combination of AU 12 with AU 25, that is, mouth may be opened in addition to the pulled lip corners as in displayed Figure 6.8. In this case the appearance of AU 12 will be quite different from the former instances. Therefore, when we use a monotype reference the mismatch in AU 12 may deteriorate due to the co-occurrence of AU 25. However, if we use an open mouth reference as in Figure 6.8, we will obtain matching pixels inside mouth.



(a) 12C **(b)** 12C **(c)** 12C+V14C **(d)** L11B+12C

(e) 25C+26C **(f)** 12D+16B+25D **(g)** 12C+25D+27C **(h)** 12E+16C+25D+26C

Figure 6.8. Color and surface mean curvature images of various AU 12 - Lip Corner Puller instances. (a-d): only pulled lip corner deformation, (e): open-mouth, and (f-h): pulled lips and open-mouth together.

Another way to improve our detector can be to extract features that better represent the negative samples as well. So far we were interested in better matching of positive expression samples only, vis-à-vis of negative samples. Actually we could approach to problem in an opposite direction as well, that is, we could use more references representing the negative samples. Since negative samples encompass a much larger set of variations, the negative class can benefit more from multiple references.

The multiple reference scheme for expression recognition is illustrated in Figure 6.9, and is based on selecting the most discriminative and parsimonious set of curvature pixels from different references. To this effect we use the AdaBoost feature selection. The input feature vector for AdaBoost is created simply by concatenating all registered pixels of an input sample image in a vector. As shown in the block diagram in Figure 6.9, first, the input image is registered onto all references. Next, the best performing pixels according to the AdaBoost algorithm are picked up from the different reference domains, and this forms the joint feature vector. Finally, a classifier is trained on the fused feature vector. Some example registrations of two different

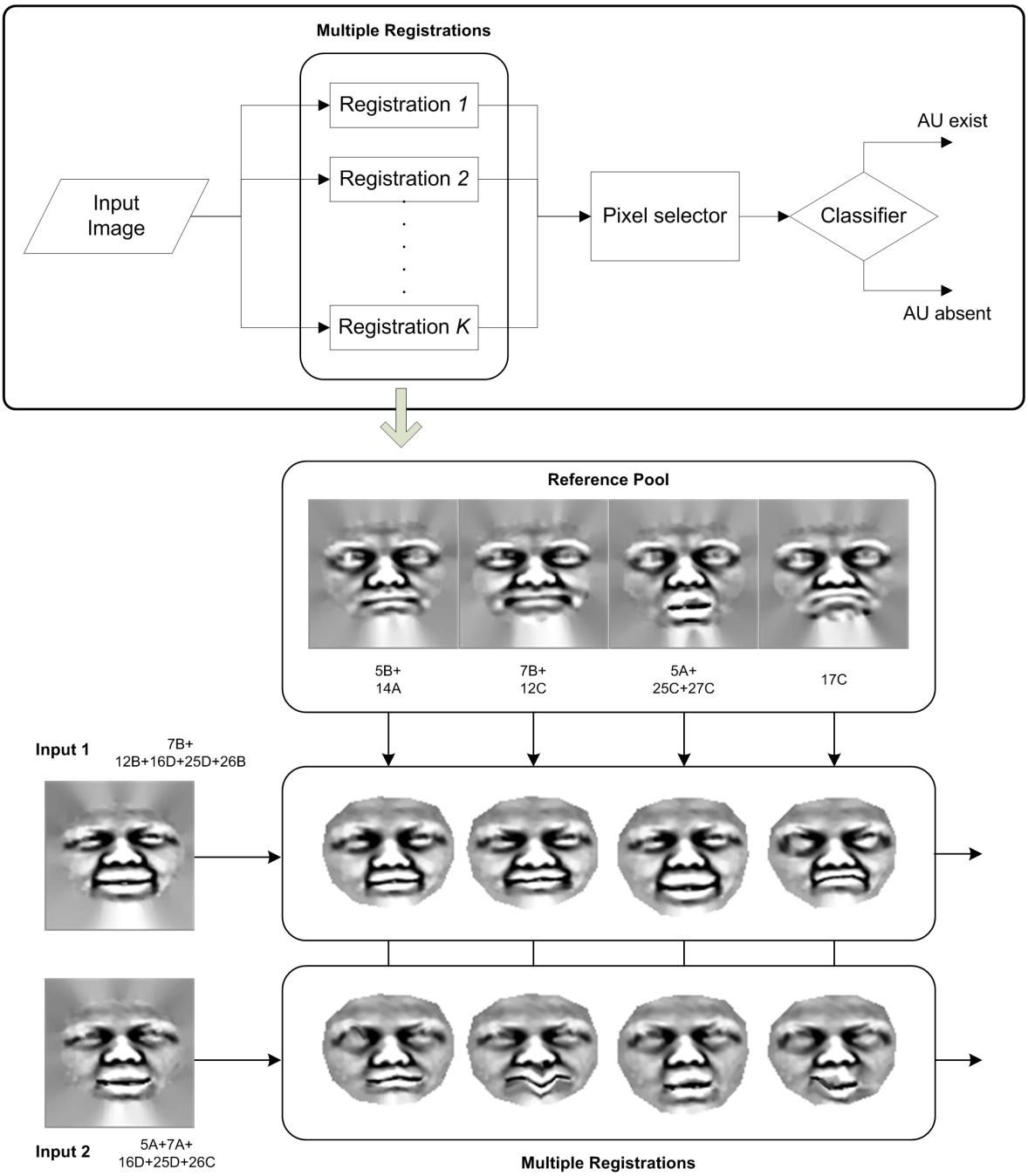


Figure 6.9. Detection of a target AU by multiple registration-based AU detector. Example registrations of two different input expression faces of a subject with a quite elastic model ($\rho = 10$) are shown below.

expression faces of a subject with a quite flexible model ($\rho = 10$) are also shown in Figure 6.9. The reference subject is the same for all registrations. From these registration examples we observe different deformations of the input images due to the expression specific references. We see that while some of the deformations are good estimations

of shape differences, others are locally unrealistic. This is due to the transient features of expressions and use of a very flexible model. Nevertheless, both types of results are informative for recognition.

Obviously, the more references we include in the pool the better will be the selection of AdaBoost and hence potentially the performance would improve. It is important to select the reference sets so as to incorporate many facial features of expressions for the positive classes and of the remaining expressions for the negative classes. In our design we limited the number of references to 23 as a compromise between computational load of registrations and representative thoroughness. We selected a set of 23 reference faces by visually inspecting their database samples, and use this set for testing all 25 AUs. All the expression references are selected from only one subject (shown in Figure 6.6 and Figure 6.9), in order not to use test subjects as references for the person-independent evaluations. Notice that though we employ many references, our reference set does not have to cover all types of facial deformations of AU combinations. In Section 6.5.2 and Section 6.5.4 we evaluate effects of references.

6.5. Experimental Results and Discussions

6.5.1. RoI Masking versus Automatic Coordinate Selection

In order to compare local analysis done by manual RoI masking (see Section 6.3.1) with automatic coordinate selection via AdaBoost (see Section 6.3.2), we performed detection experiments on the Bosphorus-DS1 dataset containing 1771 faces. As described in Section 3.1.3, DS1 dataset is composed of monotype 22 AUs, hence this experiment deals with faces possessing only one type of AU without any non-additive combination.

The experimental setup is as follows. For each monotype AU we used a typical instance of it as a reference and applied registration with the rigidity coefficient set at $\rho = 500$. The mean curvatures of facial surfaces are mapped by the LSCM technique and resampled on 96×96 image grid. The resulting average size of reference image

domains is 5212 pixels. We tried different number of features (curvature pixels) with AdaBoost feature selection and observed that after 200 pixels the average performance does not improve. On the other hand, the AU masks, some of which are shown in Figure 6.4, cover approximately between 400 to 1000 pixels depending on their activation regions.

In Figure 6.12 we see the differences between the two methods from the bars denoted by "RoI Masking" and "Single Registration" under three classifiers. These results show that finding pixels that have the most discriminating capability is better than manually selecting them according to their activation regions. The best performance was obtained with quadratic Gaussian classifier. Quadratic classifier also achieves higher performances than Naïve Bayes and nearest mean for all the registration-based techniques, as shown in the bar chart in Figure 6.12. Since DS1 is a dataset of single occurring AUs, that is, mostly a single deformation exists for each AU, these results validate our assumptions (see Section 6.3.2) that variances are informative in discriminating between positive and negative deformations.

Figure 6.10 compares manually and automatically selected pixels for one upper and two lower face AUs. In this figure 200 AdaBoost-selected pixels are shown on the reference images for each AU while the manually selected regions of interest is portrayed in the insets. Although the density of the selected pixels is slightly higher in the AU activation region, it is interesting to observe that AdaBoost-selected pixels actually take place all over the face. There may be three explanations for this situation. First of all, AdaBoost selects the most discriminative pixels such that in every successive training iteration, the recruited pixel set is the best performing one over the misclassified samples in the previous iterations. Therefore, a parsimonious set of AdaBoost-selected pixels may be sufficient for AU activation regions. The second reason may be the correlations between the AUs occurring on different face parts, which are captured at feature level through the AdaBoost. Since Bosphorus-DS1 dataset is composed of monotype AU samples involving few co-occurrences with AUs at other regions, we may encounter negative correlations; for instance when detecting deformations on lips, actions on eyes are found only in the negative classes and are not co-activated frequently with lip ac-

tions, hence this may be the reason why AdaBoost selects some pixels from the upper face. A third reason may be the estimated deformation fields itself; it may be possible that estimated face deformation due to an AU may spread to an area larger than the designated one.

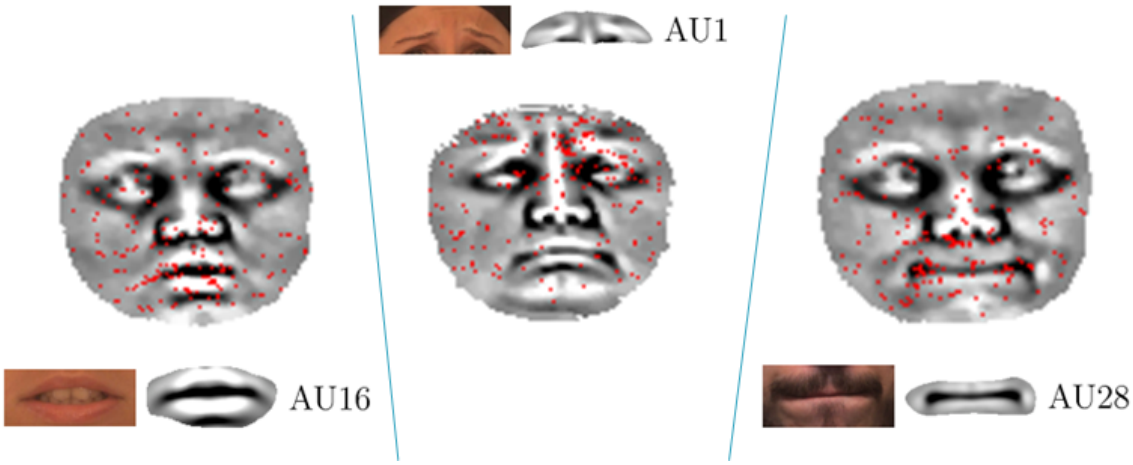


Figure 6.10. 200 AdaBoost-selected pixels are compared with manually selected RoI mask pixels for AU 1, AU 16 and AU 28. Registration with $\rho = 500$ was performed on 96×96 pixel images of surface curvature mapped by LSCM technique.

6.5.2. Multiple References versus Single Reference

With the same setup as in the previous subsection, we compare use of single expression specific reference with the use of more than one reference, on the 22 AUs of the DS1 dataset. For the multiple reference scheme case, we register the inputs on the 22 AU and a neutral face reference. In other words we use the same set of 23 references for every AU. Thus, the feature vector, i.e., the total number of pixels belonging to different references, is about 120,00.

Figure 6.11 compares the performances with multiple references versus single reference under the quadratic Gaussian classifier and different number of features. We see that using more than one reference improves the performance, which is statistically significant since the overlap in confidence intervals are minor. For both of the methods, after 200 features the results did not improve as shown in Figure 6.11. Figure 6.12 shows that this improvement on the DS1 dataset is consistent under three different classifiers. In Section 6.5.4 we examine the effect of references on the detection performance, and

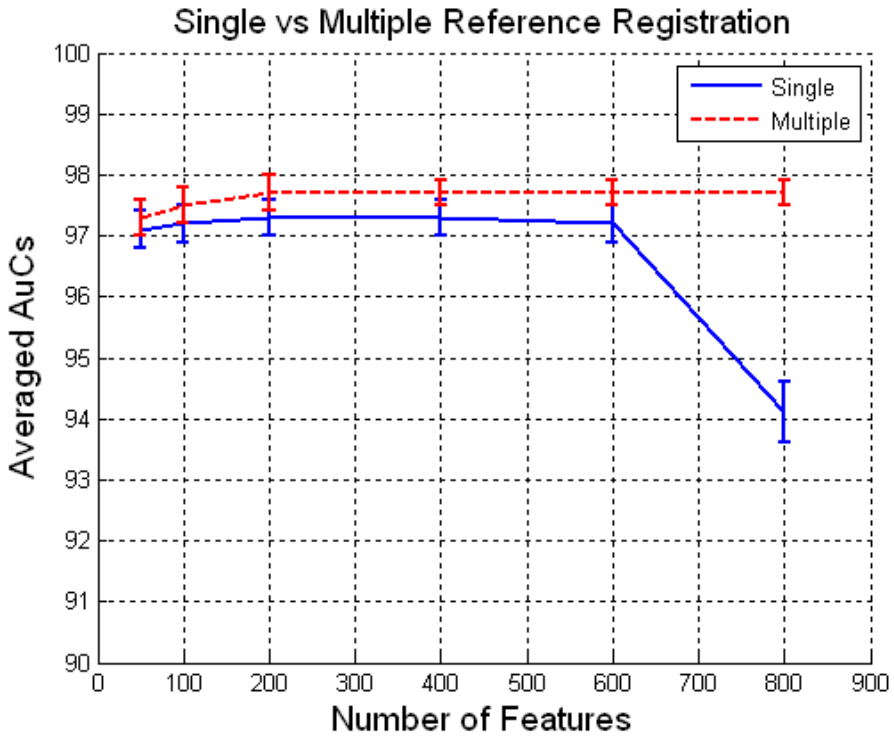


Figure 6.11. Average AuC values and 95% confidence interval estimates of single and multiple registration-based AU detectors on the Bosphorus-DS1 dataset under the quadratic Gaussian classifier with varying number of features.

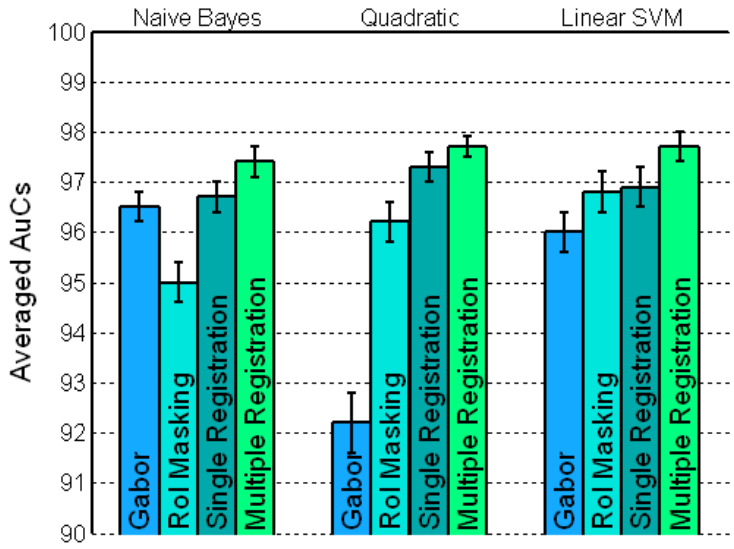


Figure 6.12. Average AuC values and 95% confidence interval estimates of Gabor and registration-based detectors on Bosphorus-DS1 dataset. 400 feature are used, except masking where the number of features is variable dependent upon the RoI mask.

in the rest of the experiments we use the same 23 references in registration-based detectors.

6.5.3. Sensitivity to Elasticity

For registration-based detectors an important parameter is the ρ given in Equation (5.19) that controls the rigidity to the shape deformations. As explained in Section 5.3, ρ controls the regularization, is necessary to estimate plausible deformations, and it is set a priori. As depicted in Figure 5.9 lower values of this parameter allows larger deformations. To determine how to set this parameter, we investigated the detection performances under different ρ values. We tested multiple reference detectors using the 23 references as in Section 6.5.2 with varying ρ values. In this experiment we used the Bosphorus-DS2 dataset (see Section 3.1.3) which is composed of 2902 faces encompassing all the expressions with full FACS codes. This was also the experimentation dataset used in Chapter 4.

Figure 6.13 shows the recognition performance as a function of ρ, obtained with orthogonal projection and 200 features under quadratic Gaussian classifier. The ρ values are logarithmic on the x-axis. We see that the performance is not sensitive to the elasticity within a range from $\rho = 0$ to $\rho = 10^3$. Notice that this range is quite wide since $\rho = 10^3$ strongly penalizes deformations. At the right end of the graph, the performance drops to 92.9% when ρ goes to infinity. This corresponds to no deformation at all, hence no non-rigid registration, which means that we can still recognize AU even though we employ only 3D rigid registration and raw mean curvature values, i.e., no other feature transformation like Gabors.

It is somewhat surprising that the highest performance 96.3% is attained for $\rho = 0$. In other words, for recognition task over the mean curvature fields, elasticity constraint, hence regularization is not necessary. This is surprising because regularization is a critical component in non-rigid registration problems. The role of regularization is to generate plausible shapes. An explanation could be that, for the recognition task, it is not important to estimate the most plausible deformation, but to generate

the most discriminating matching patterns.

Many registration examples with $\rho = 0$ and $\rho = 500$ are illustrated in Figures B.1-B.10. The figures are set as follows: The reference face is at the top, several input faces of different subjects with lower and upper face deformations and neutral faces are shown in the leftmost and rightmost columns, and the registration results are in the middle two columns. Notice that in all these figures, the inputs are the same, while the reference and/or ρ vary. For instance Figure B.1 and B.2 show registration onto a neutral reference with $\rho = 0$ and $\rho = 500$. In general we see that without regularization abrupt or non-plausible shape changes can occur on certain parts of the face, while for $\rho = 500$ registration results in good matches over facial parts for all subjects but with limited expression deformations. The only exception is when there are non-corresponding parts, as in the case of open and closed mouths. However, despite the more plausible registrations with regularization ($\rho = 500$), the performance does not improve, in fact it drop slightly to 96.0%, 0.3% lower. Use of deformation fields instead of curvature fields with $\rho = 0$ and $\rho = 500$ is also compared in Section 6.5.5.

We have found that the quadratic classifier is less sensitive to the elasticity parameter than the linear classifier (Naïve Bayes). The performance of linear classifier declines from a maximum 96.3% AuC for $\rho = 0$ to 95.2% for $\rho = 10^3$ whereas for quadratic classifier it drops less to 95.8%. This may be explained by the variance reduction effect of using a Gaussian model for image similarity metric, as described in Section 6.3.2. Increasing rigidity means deformations will be penalized more so that negative samples will not be able to deform sufficiently and therefore cause higher variances. Therefore, since quadratic classifiers use the information coming from variances while linear classifiers do not, the performance of Naïve Bayes drops more quickly.

Figure 6.14 shows the performance differential of AUs resulting from $\rho = 0$ and $\rho = 1000$. We see that almost all the AUs are negatively affected when we employ a high rigidity model and there are very few instances of performance increase with rigid model for any of the AUs. Therefore, in the rest of the experiments we do not use any regularization.

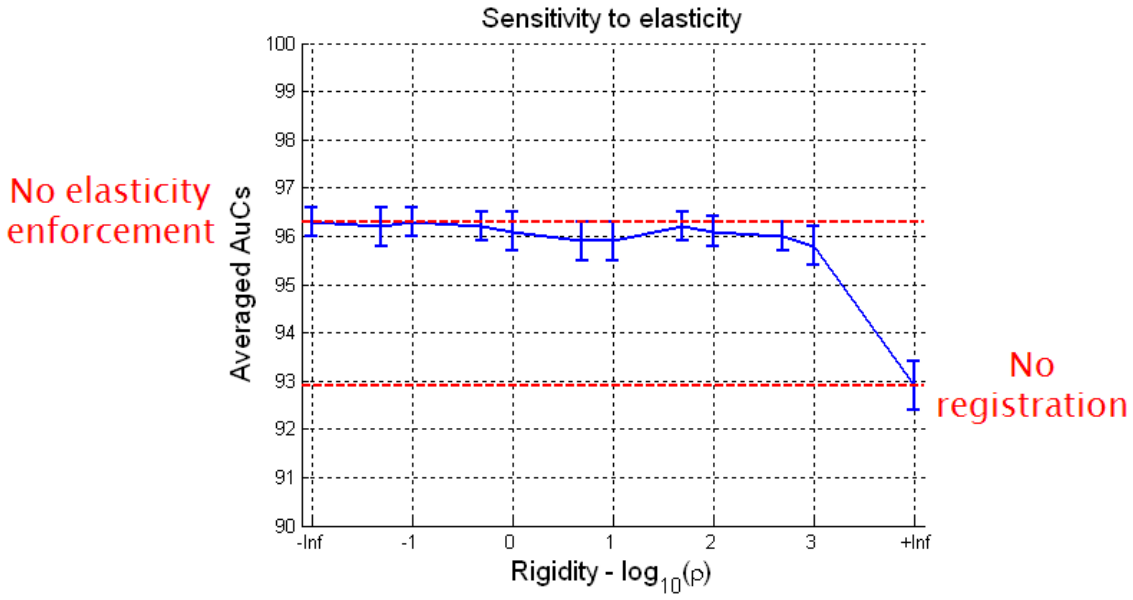


Figure 6.13. Average AuC values and 95% confidence interval estimates for different ρ values when multiple reference registration-based detectors are used with 200 features and quadratic Gaussian classifier.

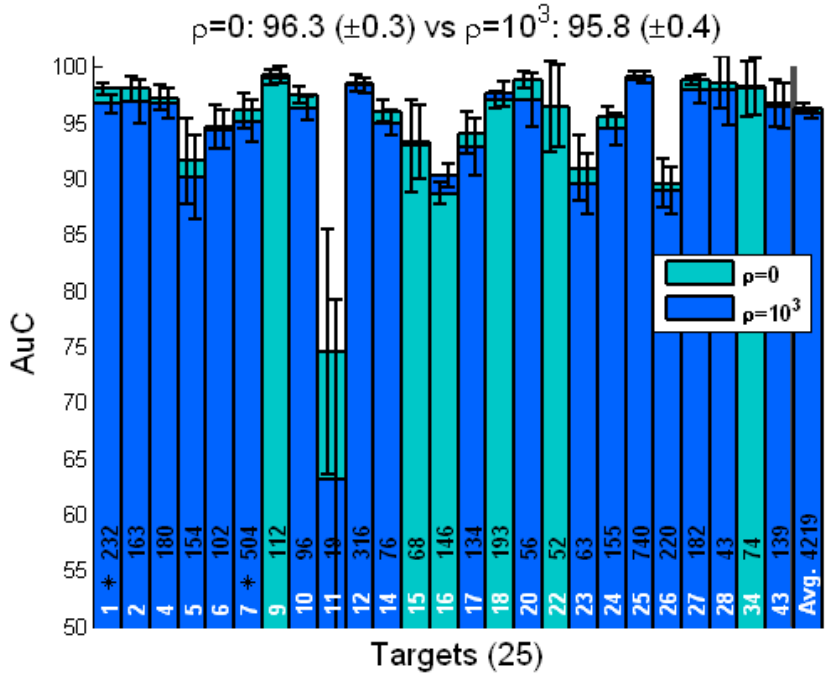


Figure 6.14. Performance comparisons under quadratic classifier (200 features): $\rho = 10^3$ vs. $\rho = 0$. On each AuC bar AU code (bottommost), a star if performances are significantly different, positive samples and 95% confidence intervals are displayed.

6.5.4. Effects of References

In this experiment we examine the effect of the number of employed references on the performance. For every AU the same set of references are employed. We start first with lower face AU references, and then add upper AU references, the selection being arbitrary. Figure 6.15 shows the recognition performance as a function of the number of references. The first reference is of the neutral face and we obtain 94.7% averaged AuC. This value suddenly increases to 95.6% by adding a second reference, which is of AU 10 - Upper Lip Raiser. Notice adding a third or further references the performance increments is very slow. We switch to adding upper facial AU references after the 18^{th} reference. With the introduction of the upper face AUs, the performance curve settles at 96.3%.

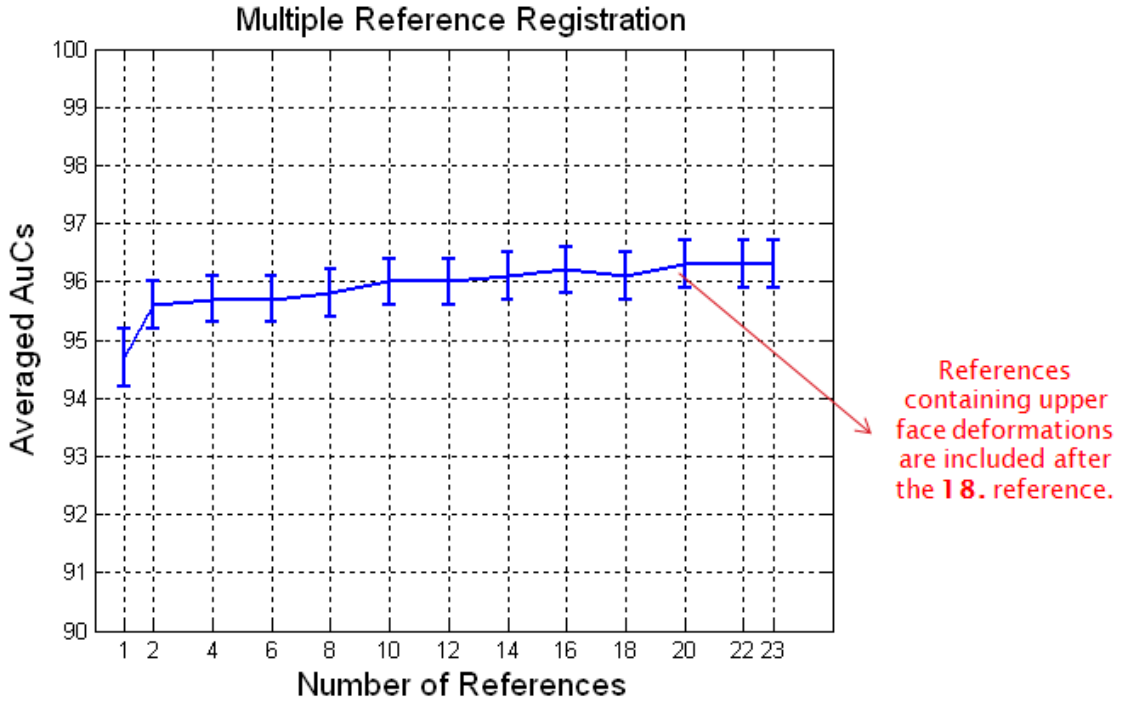


Figure 6.15. Average AuC values and 95% confidence interval estimates for varying number of references under quadratic Gaussian classifier with 200 features.

6.5.5. Deformation Fields versus Curvature Fields

We compare the use of two estimated fields for AU recognition, namely, that of spatial deformation and of registration of the curvature field. The feature length of the deformation fields collected from different reference is about 2,200, much less than that of the curvature fields, which is about 120,000. The performance results are shown in Figure 6.16 as a function of the number of features. We observe again as discussed in Section 6.5.3, use of deformation regularization is not useful, it actually causes a performance drop. When we compare the deformation and curvature fields in Figure 6.16, the curvature field is the winner. The average results for 400 features without using hyper-elastic regularization is 96.7% and 95.5%, respectively.

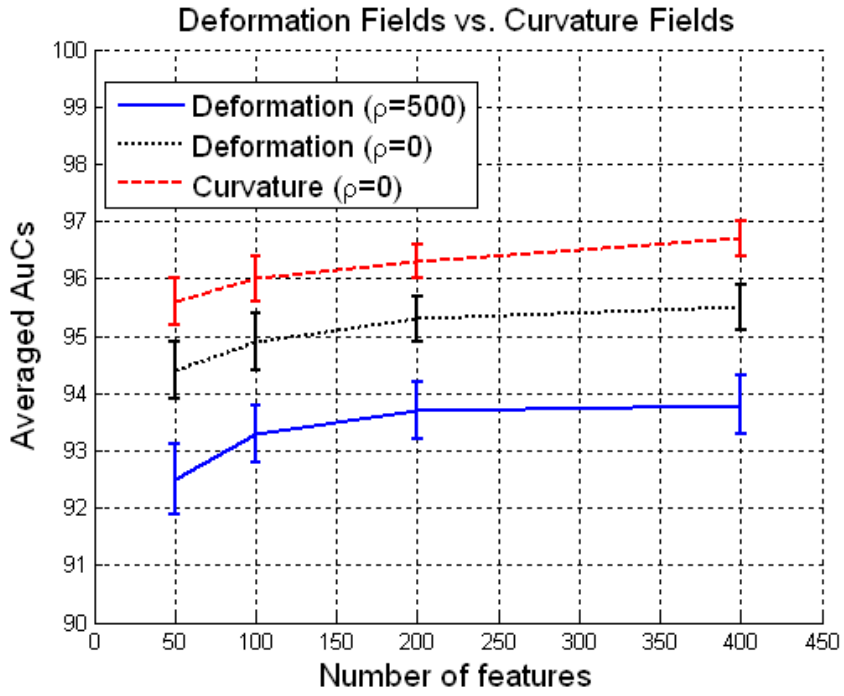


Figure 6.16. Average AuC values and 95% confidence interval estimates of deformation and curvature field-based detectors under SVM classifier with RBF kernel with varying number of features.

We have also found that in contrast to curvature fields, when the deformation field is used, Gaussian classifiers perform considerably worse as compared to SVM classifiers. With 400 features, RBF-SVM, Linear-SVM and quadratic-Gaussian achieve

95.5%, 95.1% and 91.9% averaged AuC values respectively.

The deformation field carries the shape information, which is missing in the registered curvature field since inverse deformations is applied to register. Hence, there may be useful complementary information between them. To test this, we pooled the two types of features, and then selected a subset of them by AdaBoost. However, we did not obtain any gain, the results were exactly the same as with the curvature field alone case. This finding indicates that either complementary shape information found in the deformation field is not helpful or the residual shape data of identity differences found in combination with expression related deformations does not allow effective use of this complementary information. Therefore the use of deformation becomes unnecessary.

6.5.6. Gabor Wavelets-based versus Registration-based Detection

We compare Gabor wavelet-based and registration-based AU detectors on Bosphorus DS1 and Bosphorus DS2 datasets. Figure 6.12 compares Gabors with three different registration-based detectors, namely masking, single reference and multiple references, on the Bosphorus-DS1 dataset under three different classifiers. We see that all the three registration-based detectors are better than the Gabors under all classifiers. The only exception is that, under Naïve Bayes, Gabors achieve higher performance than the mask-based registration technique. Notice that, while Gabors benefit from Naïve Bayes, registration-based classifiers have higher performance with quadratic Gaussian classifier.

Table 6.1 compares multiple reference registration-based and Gabor-based detection on the Bosphorus DS2 dataset. We observe that under all classifiers registration-based detectors are better. The highest values are obtained by SVM classifiers with RBF kernels, and registration-based detectors achieve 96.7% average AuC, 1% higher than the Gabors. Figure 6.17 displays these results for each AU. We observe a big improvement for AU 23. Other statistically significant improvements are on AU 7 and AU 14. Also, for many AUs we obtain p-values slightly higher than 0.05, and thus though they are failed according to paired t-test for 5% significance level their differences are

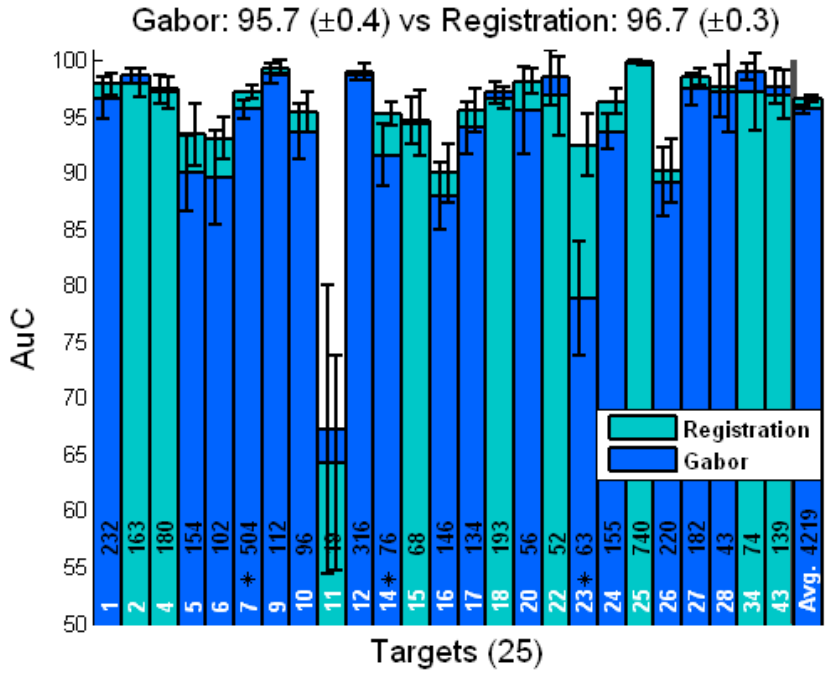


Figure 6.17. Comparisons of registration-based and Gabor-based detectors with 400 features. On each AuC bar AU code (bottommost), a star if performances are significantly different, positive samples and 95% confidence intervals are displayed.

still remarkable.

Another noteworthy observation from Table 6.1 is the very close performances of by Naïve Bayes and quadratic diagonal Gaussian classifiers used in registration-based detectors. Recall from Figure 6.12 that registration-based detectors achieve higher performance with quadratic classification on the DS1 dataset. This outcome validates our assumptions stated in Section 6.3.2 that variances are informative in discrimination between positive and negative deformation samples. However, with the DS2 dataset the two average performances are the same, only with one noticeable difference which is the higher confidence of the quadratic classification results. This contradiction can be explained by the contents of DS1 and DS2 datasets. DS2 dataset is composed of AUs that have higher degree of variability due to the many types of AU combinations in contrast to monotype and clearly occurring AUs of DS1. Therefore, we can benefit more by quadratic classification on the DS1 dataset since the employed expression references are good representatives of the AUs that we analyze and thus the variances represent

genuine within class variations. These results suggest that quadratic classification can be the appropriate choice when our problem is recognition of prototypical deformations for which we have good representative references.

On the other hand, when we compare the DS1 and DS2 datasets according to linear and quadratic classifiers from the Gabor analysis point of view, we do not attain inconsistent results as seen from Figure 6.12 and Table 6.1. Naïve Bayes always performs much better than the quadratic classification. We may think of two reasons for the failure of quadratic classifiers in case of Gabors: the amount of data samples may be insufficient for Gabor analysis to have reliable estimates of the variances from the positive samples, or the quadratic classification model may not be the right assumption at all for the underlying distribution of the samples.

We also implemented fusion of registration curvature values with luminance Gabor features by combining 400 features from both modalities, then selecting 400 out of 800 via AdaBoost. Results are given in Table 6.2 where we also compare them with the 2D luminance only case. We see that in the modality fusion, i.e., 2D luminance and 3D information, using 2D luminance with registration-based detectors is better than fusion with Gabor-based features. Under RBF classifiers we obtain 97.1% and under quadratic Gaussian 97.2%. The performance with fusion with Gabor features were resulted with 96.7%.

In Figure 6.18 we compare 2D luminance, 3D geometry and their fusion for varying number of features. The 3D geometry is represented in two ways, registration-based and Gabor wavelet-based. We see that for all the methods the performance is monotonically increasing function of the number of the features, and the rankings do not change depending on the feature amount. These results prove the superiority of 3D over 2D, registration over Gabors, and modality fusion over single modality.

Table 6.1. Average AuC values and 95% confidence interval estimates of Gabor and registration-based AU detectors. All of the classifiers use 400 features that are selected by AdaBoost.

Classifier	**Gabor**	**Registration**
AdaBoost	95.4 ± 0.4	96.3 ± 0.3
Linear-SVM	95.1 ± 0.4	96.3 ± 0.3
RBF-SVM	**95.7 ± 0.5**	**96.7 ± 0.3**
Quadratic D.	93.2 ± 0.5	**96.3 ± 0.3**
Quadratic S.	85.2 ± 0.7	92.8 ± 0.5
Naïve Bayes	**95.4 ± 0.4**	96.3 ± 0.4
Nearest Mean	92.3 ± 0.6	95.4 ± 0.5

Table 6.2. Average AuC values and 95% confidence interval estimates of 2D luminance and modality fusion-based AU detectors. Registration is applied only on surface curvature modality for Fusion-Reg. detection, otherwise Gabors are used.

Classifier	**2D Lum.**	**Fusion-Gabor**	**Fusion-Reg.**
AdaBoost	92.8 ± 0.5	95.7 ± 0.3	95.5 ± 0.3
Linear-SVM	93.0 ± 0.6	96.4 ± 0.3	96.8 ± 0.3
RBF-SVM	**94.0 ± 0.5**	**96.7 ± 0.3**	**97.1 ± 0.3**
Quadratic D.	90.1 ± 0.7	95.1 ± 0.4	96.7 ± 0.3
Quadratic S.	63.4 ± 1.1	91.4 ± 0.5	95.1 ± 0.4
Naïve Bayes	91.3 ± 0.6	**96.7 ± 0.2**	**97.2 ± 0.3**
Nearest Mean	82.8 ± 1.0	**96.7 ± 0.2**	**97.2 ± 0.3**

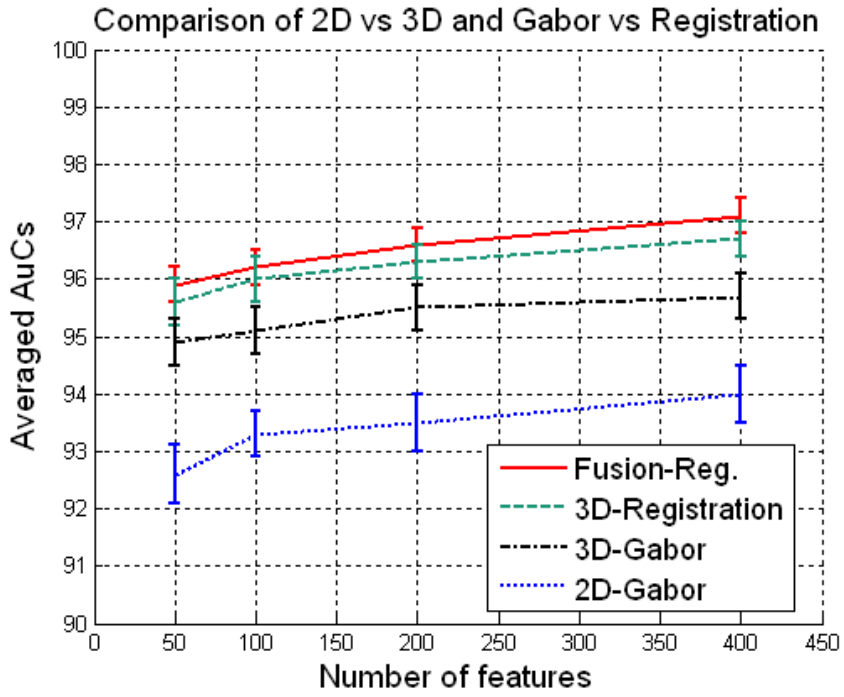


Figure 6.18. Average AuC values and 95% confidence interval estimates of detectors employing Gabors on 2D luminance and 3D surface curvature, and registration on 3D surface curvature, under RBF-SVM classifier. "Fusion-Reg.": Feature fusion of Gabors on 2D luminance and registration on 3D surface curvature using AdaBoost.

6.5.7. 3D Pose Robustness of Registration-based Detectors

In Section 4.9.6, we showed in Table 4.2 that current data-driven expression analyzers benefit from improved head pose estimation. With Gabor wavelet-based detectors, we observe a drop from 95.5% to 94.8% on average AuC values when we train and test the AU detectors after ICP-based automatic 3D pose normalization instead of landmark-based manual normalization. Here, we conjecture that non-rigid registration-based detectors must inherently be robust to 3D pose since after detailed correspondence estimation by non-rigid registration, pose related distortion is compensated.

While working on 2D orthogonal projection of 3D faces, out-of-plane pose transformations, i.e., translation in depth and rotations as yaw and pitch, will not be linear as in the 3D space. Nevertheless, thanks to non-rigid registration we can handle these

non-linear transformations adequately, which would not be possible with a linear model in 2D space. An example is demonstrated in Figure 6.19. First row shows the reference and the input mean curvature images. Notice the differences between the reference and the input face: 3D pose of the input with respect to the reference is 20 ° yaw rotation, subject identities are different and the reference mouth is more opened than the input. The deformation of the reference and the resulting registration of the input obtained with $\rho = 0$, $\rho = 100$ and $\rho = 500$ are shown below. We see from registered images how 3D pose, identity and expression differences are normalized, with higher rigidity value less motion is compensated. Notice from the deformed reference images that, even with high rigidity value, some pose and also mouth closing deformation can be estimated.

In order to test whether our conjecture is valid or not, we trained 25 non-rigid registration-based AU detectors with ICP normalized faces. After testing with several different values of ρ, we attained exactly the same performance levels of the landmark-based manual aligned faces. For instance, selecting 400 pixels from faces registered onto multiple references with $\rho = 0$, we obtained the same 96.3% average AuC and also the same performances for every AU. This experiment proves that by non-rigid registration-based detectors, we are able to compensate for any pose estimation imperfections in automatic processing as compared to manual face alignment.

6.5.8. Non-rigid Registration as a Preprocessor?

We investigated if we could benefit from non-rigid registration employed in the preprocessing role for conventional feature extraction-based detectors. For this purpose we apply non-rigid registration at the frontend of Gabor-based expression analyzers, and also compare their performances with non-rigid registration-based detectors. The motivation is that, since we are able to reduce the confounding effects of residual pose and facial physiognomy differences, we may subsequently attain higher performances also with the traditional techniques.

To examine this conjecture we performed three experiments where we do clas-

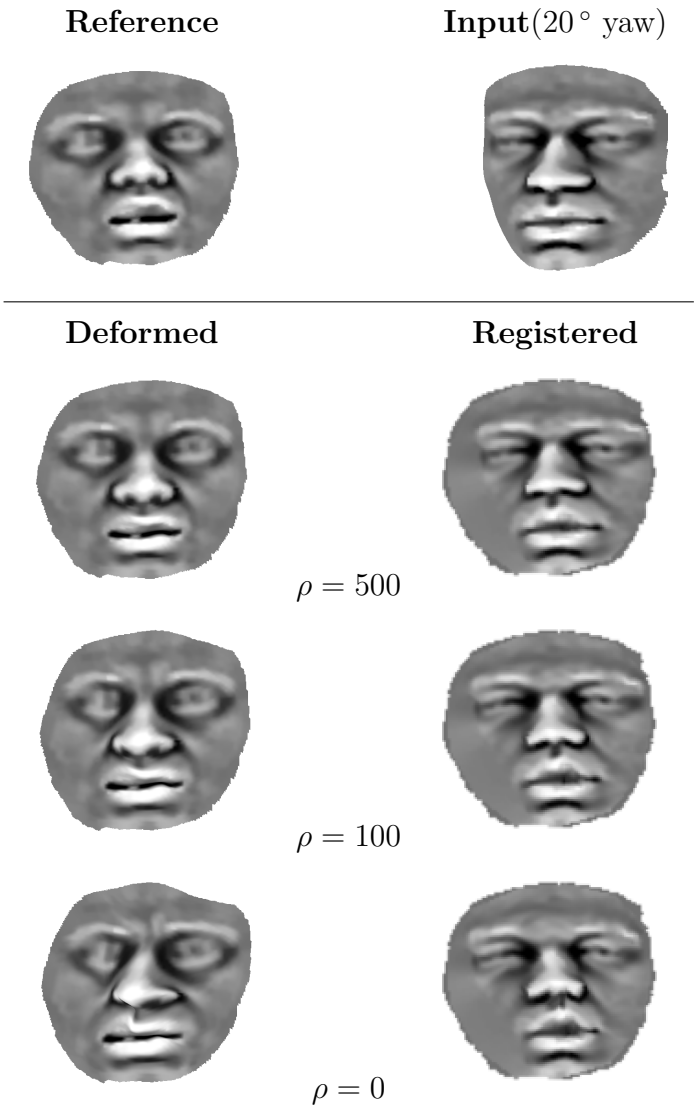


Figure 6.19. Orthogonal projection of an input 3D face with 20° yaw motion is registered onto a reference face of a different subject and of a wider mouth opening under three rigidity (ρ) values.

sification with and without non-rigid registration. In all these experiments we apply landmark-based manual pose alignment followed by registration performed on the curvature fields with $\rho = 500$. The reference is always a neutral face in order to suppress physiognomical differences while preserving the shape changes due to the expressions. In the first experiment we use surface mean curvature images; in the second experiment, we apply Gabor feature extraction on curvature images; and in the third case, we apply Gabors on luminance data.

The results obtained by using 200 features under SVM classifiers with RBF kernel are shown in Figure 6.20. The relevant observations from this figure and from experiments with other classifiers are listed below:

- With direct use of registered curvature images we obtain a considerable performance gain from 94.4% to 95.9% thanks to non-rigid registration. We have also noticed that the second best performing classifier for raw curvature data was the quadratic Gaussian classifier and it achieves 95.1% average AuC with registration, however, deteriorates seriously in the absence of non-rigid registration, to 92.9%. Other classifiers also incurred in big differences when we did not perform non-rigid registration.
- In the case of Gabors the differences due to non-rigid registration, whether on surface curvature or luminance data, seem statistically not significant since confidence intervals overlap considerably. The differences have also been found insignificant for other classifiers.
- The Gabor analysis on curvature images achieves slightly lower performance as compared to the direct use of registered curvatures but much better than curvature pixels without registration.

To recapitulate, we have discussed and AU detector which is using AdaBoost-selected Gabor features, and registers faces in two successive steps, first, with 3D alignment with manual landmarks and then via scaling of horizontal and vertical dimensions of 2D curvature maps to best fit to the square image analysis domain. The results in Figure 6.20 prove that this Gabor-based scheme is quite robust to residual motion and shape differences of facial parts. The reason behind this robustness may be that, spatial frequencies selected by AdaBoost do not change much with small deformations of face parts relative to the regions effectively supported by Gabor wavelets, while the individual pixels vary a lot. Since in this experiment we align face pose by manual landmarking which is quite accurate, we anticipate that effect of pose in an estimated deformation field can be negligible compared to identity related differences. Therefore, these findings suggest that non-rigid registration employed as a pre-processor to compensate for identity differences is not necessary if we apply Gabor-based data-driven

expression analysis. However, in light of the experimental results in Section 4.9.6 and Section 6.5.7 where 3D pose robustness of Gabor-based and registration-based detectors are investigated respectively, we deduce that Gabor wavelets are sensitive to pose alignment inaccuracies and can benefit from the better pose alignment that non-rigid registration offers.

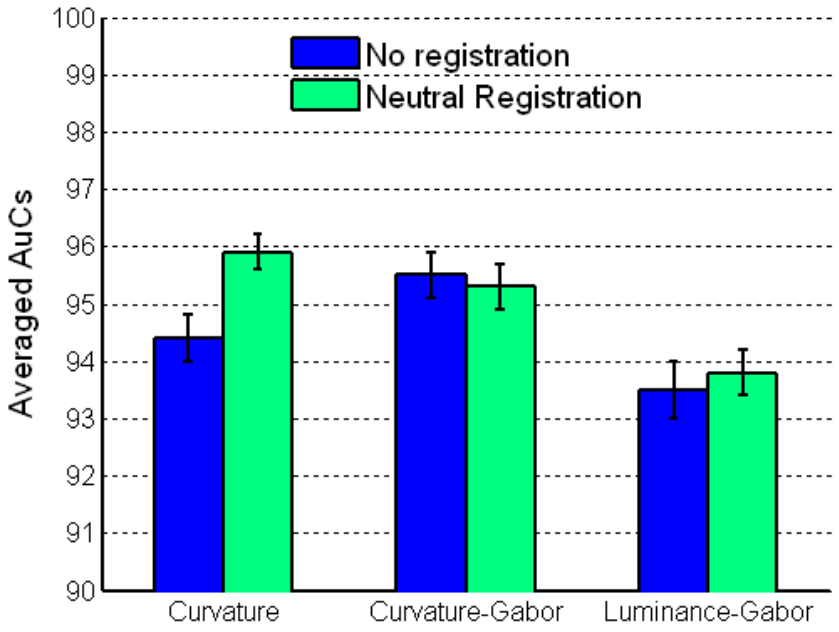


Figure 6.20. Average AuC values and 95% confidence interval estimates of three types of detectors with and without non-rigid registration as a preprocessor. Registration is done onto a neutral face. "Curvature": mean curvature data, "Curvature-Gabor": Gabor features applied on mean curvature data, and "Luminance-Gabor": Gabor features applied on luminance data. SVM classifier with RBF kernel is used to classify 200 features selected by AdaBoost.

7. CONCLUSIONS

In this thesis we have dealt with the facial action unit recognition problem via 3D modality, and have concentrated on person-independent data-driven detection using static images of expressions. The two hypotheses that we had advanced were:

(i) 3D acquisition must be a better data modality as compared to conventional 2D cameras since the former captures surface information directly;
(ii) It is possible to utilize detailed face registration in a data-driven way, i.e., without resorting to any face modeling, in order to improve the state-of-the-art in automatic expression analysis by handling facial features of expressions effectively, avoiding all the drawbacks of model-driven analysis (see Table 2.1).

In order to validate these two hypotheses, we prepared an extensive 3D face database covering a range of expressions, and made our evaluations within the framework of the FACS paradigm. This database is a contribution to the open-access repertoire of expression databases and its importance stems from the fact it contains up to 54 face scans per subject including various posed expressions (up to 35 per subject) that are annotated by a certified FACS coder who scored all the AUs with intensity codes. In addition to expressions, this database includes systematic head pose variations and different types of face occlusions.

In the following section we summarize the main contributions of this thesis. Then in Section 7.2 we discuss possible future directions of our study.

7.1. Original Contributions

7.1.1. Novelties

This thesis presents the first comprehensive study on 3D AU recognition in the literature, dealing with 25 AUs that occur singly or in various combinations. We

also address the detection of lower intensity AUs and estimation of their intensities. Previous 3D studies have focused on prototypical expression recognition.

The novel aspects of the methods developed in this thesis are summarized as follows:

(i) *Data-driven Analysis of Expressions on 2D Maps of 3D Facial Surfaces.* We have developed and compared two categories of data-driven algorithms for facial expression analysis: recognition using the conventional scheme, i.e., applying feature transform techniques on pose-normalized face images and recognition based on non-rigid registration. The property underlying the 3D expression recognition is the use of 2D maps of 3D facial surface properties. Although we are not the first one to propose 2D mapping of 3D surfaces, the difference of our approach is the direct use of 2D-resampled surface curvature values in contrast to previous studies that use 2D luminance texture maps [48, 49]. This technique provides efficient processing of 3D data. Another advantage is that, after resampling on an image grid, the analysis is not affected by differences in surface mesh resolution and topology.

(ii) *Regression-based Facial Action Intensity Estimation.* We have developed regression-based AU intensity estimators that are generic in that they are not subject-dependent. The estimators are based on ε-SVM regression on AU-specific Gabor features. To the best of our knowledge this is the first work employing regression for intensity estimation, whether in a subject-independent or subject-dependent scheme.

(iii) *Efficient Non-rigid Registration of 3D Surfaces by Deformable 2D Triangular Meshes.* Registration of faces in man-machine interfaces require computational efficiency and capability to handle large deformations. Our proposed technique based on the Green-Lagrange strain allows for larger deformations than the classical linear elastic models used in registration. The algorithm is also computationally efficient since it solves the deformation strain forces in triangular elements over which deformation gradient tensor is constant. Also, our method enables adaptive mesh generation for more detailed calculations wherever neces-

sary. Furthermore, we develop an efficient surface extrapolation technique that conveniently completes missing correspondences. The types and topologies of surfaces that can be handled with this method depend on the 3D-to-2D mapping technique. For instance, in the case of orthogonal projection one cannot work when a non-negligible portion of the surface has normal direction very different from the projection direction. For instance, for registration of the whole head spherical projection could be used. On the other hand, the limitations of the LSCM method is that the surfaces must possess disk topology (0-genus mesh with borders). However, some violations to disk topology constraints like holes can be recovered by surface interpolation as done in our work.

(iv) *Non-rigid Registration-based Data-driven Expression Analysis.* The most important novelty of this thesis is the utilization of non-rigid registration for data-driven expression analysis. Our method enjoys all the advantages of data-driven analysis, and also benefits from detailed registration that was previously only possible within a model-driven method. The attractive aspect is that, the registration advantage does not come at the price of face modeling (see Section 2.2.3), i.e., error prone model fitting, tedious model preparation, and bias by the assumed models. There are two key points of the proposed non-rigid registration-based scheme. First, we achieve detailed estimation of facial feature correspondences, and thus we are able to suppress confounding effects of mild head pose and subject-identity related spatial deformations. Second, in our approach we use expression specific references in order to effectively handle various types of transient features of expressions like furrows, as well as permanent facial features.

7.1.2. Main Findings

The main findings of 3D modality for expression analysis are as follows:

(i) *Usefulness of 3D Surface Data vis-à-vis Light Camera Images.* We have systematically evaluated the use of 3D data for subject-independent facial action unit detection and compared with conventional 2D camera images. By means of 3D-to-2D mapping, we make one-to-one comparisons of the two modalities under the

same set of algorithms. Furthermore our wholly data-driven analysis precludes any bias of model-driven techniques, a crucial factor for fair assessment since none of the modalities are favored by model design. With extensive experimentation over 25 selected AUs via ROC analyses and evaluating statistical significance we showed that 3D modality offers significant advantages in AU detection and performs overall better than the 2D. In general, lower face AU detections benefit more from 3D. A case in point is the considerable improvement in the detection rate of AU 23, which is difficult even for certified coders [21]. This AU is useful, for instance, in telling a genuine expression of pain from a faked one [17]. 3D also proves its value for low intensity expressions. In the case of the next to the lowest intensity level (level B), while many AUs degrade in 2D, 3D data can maintain overall higher performance. Nevertheless, there are some upper face AUs where 2D outperforms 3D.

(ii) *3D+2D Fusion Potential.* We have found out that neither 3D nor 2D is uniformly better over all AUs. We have showed that for upper face AUs that are related with actions at the eyes, success rate of 3D is lower than 2D. An explanation for this situation is the fact that 3D sensing noise is excessive in the eye region, and 3D acquisitions miss the eye texture information. Our experiments with feature fusion of the two modalities prove advantageous and indicate their complementary roles since detection of most of the lower and upper face AUs improves. In particular, feature fusion compensates for weakness of 3D on the upper face. However, we have also observed a crucial performance drop in AU 23 as consequence of fusion, which points out to the deficiency of 2D modality for that AU.

(iii) *Best Features and Best Classifiers.* The best 3D representation for AU detection is the surface curvature, particularly the mean curvature. The most effective feature set is Gabor wavelets that are selected by AdaBoost. Alternative local analysis tools such as ICA and NMFSC do not prove as effective. For Gabor wavelet-based 3D AU detection, RBF-SVM and Naïve Bayes classifiers achieve superior performance than AdaBoost and linear-SVM. RBF-SVM is the choice classifier since it even improves the results over the state-of-the-art in 2D AU detection. Finally, if neutral faces are available, use of difference images slightly

improves the performance for 3D, but for 2D, it does not seem to be beneficial.

(iv) *Contribution of 3D Normalization of Luminance Images.* We have shown that when luminance images are adaptively resampled according to 3D information (referred to as 3D Lum. in Table 4.2) a significant increase in the detection performance occurs compared to direct luminance images (referred to as 2D Lum. in the tables). Resampling the luminance image at the grid points indicated by the 3D surface data after normalization implicitly corrects for the effects of pose. Notice that although our data set is comprised in principle of frontal faces, small pose variations inevitably occur.

(v) *Facial Action Intensity Estimation.* We have proposed a method for AU intensity estimation based on regression of appearance features. This method proves to be superior to the one based on SVM margins, both for 2D and 3D data modalities. In fact, the only other person-independent AU intensity estimation study in the literature was that of [25], who use SVM margins and Gabor features, but address only eight AUs. Our 3D experiments show improvements on some AUs but also performance drops on some other AUs, both in the detection and intensity estimation problems. However, when 3D is fused with 2D luminance images, the overall performance increases significantly. We have observed that whenever a modality is better for detection of an AU, its intensity estimation is also superior in the same modality. However, the performance drop in intensity estimation for certain AUs with 3D data is more pronounced as compared to the performance differential for detection. This may be due to the same reasons as discussed above, i.e., noise on eye regions and importance of eye texture. However, this may also because ground-truth FACS scores that were created based on 2D appearance observations causes bias favored to 2D data.

The main findings concerning our hypothesis on non-rigid registration based data-driven expression analysis can be summarized as follows:

(i) *Superiority of Registration-based Detection.* We prove the superiority of non-rigid registration-based detectors for 3D expressions over the state-of-the-art AU detection based on a multitude of experiments. Notice that registration-based

detection does not mean its application as a preprocessing step in conventional methods. In effect, non-rigid registration is the core step to extract features for classification. To avoid any misinterpretation, Gabor or other feature extraction techniques need rigid registration for pose normalization as a preprocessing step, but the discussion here is the contrast between Gabor or the like feature extraction on pose normalized faces versus the non-rigid registration as feature extractor per se. We have developed three types of nonrigid registration-based detectors, namely RoI masking, single-reference scheme and multiple-reference scheme, which all attained better performance than those of Gabor-based detectors. Among all the competing techniques in this thesis, the best AU detection performance resulted with the fusion of features coming from non-rigidly registered surface mean curvature and Gabor features from luminance images. We want to mention as a side note that quadratic Gaussian Bayes classifier is quite effective for our registration-based approach whereas it performs very poorly for Gabor-based analysis. We showed its prominent advantage especially for monotype AUs, i.e., for AUs that involved a single deformation type.

(ii) *Compensation of Pose Differences.* Our experiments prove that registration-based detection is very robust to mild 3D pose inaccuracies. Even though use of 3D data should in principle overcome the weakness of 2D data to pose effects, we still have to confront with some residual pose variations since perfect initial pose alignment is difficult to achieve even with 3D data, especially when an automatic alignment procedure is applied. When the performances of manual landmarking-based and automatic ICP-based 3D pose alignment techniques are compared, we have found that while Gabor-based detectors deteriorate significantly with ICP-based automatic alignment performance, registration-based detectors maintain their performance for all AUs. We also showed visually how high degree out-of-plane rotations can be effectively compensated by non-rigid registration on 2D orthogonal projections. The non-linear distortions can be corrected appropriately. Thanks to very detailed correspondence estimation, registration-based methods are inherently robust to pose.

(iii) *Compensation of Subject-Identity Differences.* We tested whether non-rigid registration could be adjusted to suppress subject-identity related differences, like

variations in individual physiognomy, while preserving differences in expression. When we compared direct use of surface curvature pixels with and without neutral face registration, we recorded improvements by non-rigid registration, even on top of accurate pose alignment by manual landmarking. This outcome proves the compensation ability of non-rigid registration for subject-identity related differences. However, when we repeated the same experiment with Gabor wavelet responses of surface curvature pixels, we did not observe improvement on average performance over 25 AUs. This means Gabor-based detectors are robust to subject-identity related spatial deformations in contrast to their high sensitivity to pose differences.

(iv) *Insensitivity to Elasticity.* We showed that our registration-based detectors are practically not sensitive to elasticity level since the results produced with selection of the elasticity parameter ρ in a wide range were very similar. More importantly, we found out that enforcing elasticity is not necessary for our deformable triangular-based registration algorithm, and in fact the highest performances were obtained when we did not apply elasticity at all ($\rho = 0$). This finding is interesting because elasticity means regularization, which is required to estimate plausible shape differences. We also had a similar observation when we developed our detectors based on deformation fields, i.e., estimated shape differences, rather than curvature fields; the best results were obtained without regularization. Therefore, we conclude that for registration-based expression recognition on 3D data, realistic shape estimation hence regularization is not required.

(v) *Benefits of Multiple Reference Scheme.* We have experimentally proved the usefulness of multiple registration references. We first showed that, even for monotype AUs (in the absence of non-additive deformation combinations), incorporating references representing other deformation types improves the AU detection results significantly. Second, gradual improvements are observed with additions to reference pool. An explanation of the improvement with increasing number of references is as follows. In AU detection problem high degree of within-class variations are inevitable. First of all, negative classes comprise various types of deformations not belonging to target AU. Second, an AU can appear differently, especially when combined non-additively with other AUs, hence one has

to confront the within-class scatter also in positive classes. Since the rationale of registration onto expression specific references is to generate more suitable expression features, by utilizing more references that bear variants of the same expression we can better cope with different types of deformations existing in a class.

7.2. Future Directions

There are several interesting future research avenues following the present state of this study:

(i) *3D Video Data.* This work can be extended for automatic encoding of AUs in 3D facial video streams. Video data carry more information on facial expressions especially due to the dynamics and interactions between the AUs. Therefore, one can expect improvements with 3D video over 3D stills as well as over 2D video. As 3D expression videos become publicly available, we intend to extend non-rigid registration-based scheme to video.

(ii) *3D Spontaneous Expressions.* One of the current trends in expression analysis is to study spontaneous expressions. Spontaneous expressions differ from posed ones by involving higher degree of out-of-plane head movements, involvement of more subtle expressions, and not clearly delineated onset and offset instances. This is the next challenge in automatic expression analysis, and 3D with non-rigid registration seems to be the most promising candidate method, especially in view of its capability for 3D pose handling and subtle low intensity AU detection. 3D video expression databases have just started appearing, albeit not yet for 3D spontaneous expressions [71]. Notice that the requirement for 3D acquisition devices are not only to be accurate, to operate in real-time but also to be non-invasive (not disturbing the subjects by visible light). With recent progress in 3D sensing technology [100, 86] spontaneous databases will be more feasible.

(iii) *AU Intensity.* Detection of low intensity expressions and estimation of AU intensity is a topic which has not been sufficiently explored. As we have shown, person-independent low-intensity detection and intensity estimation performances

are quite low and there must be room for further improvements. This topic is important because intensity differences of the same AU can lead to different interpretations, e.g., FACS codes of some expressions given in EMFACS depend also on the AU intensities. To predict AU intensities we have used features optimally selected for AU detection. It is possible to redesign features specifically for intensity estimation.

(iv) *Smaller Amount of Training Samples.* Due to the variability in the sample sets, the size of training sets has great importance for data-driven methods. For instance, Whitehill09 [40] predicted based on their recent experiments on real world data that, for robust AU detectors $1,000$ to $10,000$ example images per target AU are required in order to capture the variability in personal characteristics and illumination. In this respect, use of 3D data and non-rigid registration approach may be advantageous. As we have illustrated, 3D data does not get affected much from variations like pose, illumination and facial albedo. Also, the proposed registration-based approach reduces variations originating from physiognomy traits, and better compensates for initial alignment inaccuracies and for out-of-plane motion. Therefore, we conjecture that smaller training set sizes will suffice thanks to both the 3D modality and our non-rigid registration technique.

(v) *Registration-based Recognition on 2D Luminance Images.* A non-rigid registration-based approach may also be utilized for luminance data. However, this would be a more challenging problem than the use of surface curvatures. This is because luminance image is not direct indicator of surface deformations and exhibits higher appearance variability due to illumination and confounding albedo variations as in facial hair. To work with luminance images we should first do some modification in the registration algorithm. Rather than using sum of squares error in the matching term Equation (6.2), we can employ other measures to mitigate lighting effects, such as normalized cross-correlation metric and mutual information measure. Obviously judicious preprocessing will help to mitigate illumination effects.

(vi) *Non-rigid Registration-based Lip Reading.* The proposed non-rigid registration-based approach can also be applied to 3D lip reading for improved speech recognition. Similar to phonemes which are acoustic units of speech, visemes are the

visible units of visual speech, hence lip reading requires recognition of visemes. AU recognition and viseme recognition share a common ground since both problems involve recognition of facial deformations. However, visual speech recognition differs from expression recognition in several aspects. First of all, dynamics of deformations are very crucial for speech recognition in contrast to expressions. While in expressions co-articulations of AUs are in spatial domain, viseme co-articulations occur in time domain. Also, visemes appear and disappear very rapidly. Another important difference is that, visemes happen only in lower face, described mainly by lip shapes, tooth and tongue. Moreover, deformations during speech usually are not as large as in expressions. Having found that 3D modality is very effective in coping with lower face AUs, especially with high detection rates even for low intensity instances of certain AUs that are closely related with visemes, such as AU 25 (Lips Part) and AU 22 (Lips Funneler), we anticipate that visual speech recognition can benefit considerably from 3D data. Effectiveness of 3D modality for lower face as well as robustness to head pose and illumination points out to its potential in visual speech recognition if we incorporate also the dynamics.

APPENDIX A: GRADIENTS OF THE STRAIN TENSORS ON TRIANGLES

The gradients of the norm of the strain tensors with respect to triangle vertex coordinates can be derived in a straightforward manner. Before deriving the gradients, let's first write the derivatives of the mapping function over a triangular element, which is given in Equation (5.4) and (5.5), with respect to spatial coordinates (u, v). This gives the Jacobian matrix of the mapping function. It is called as deformation gradient tensor in elastic theory. The same notation used in Section 5.1 is followed here. Area of a triangle can be obtained by

$$Area(\widehat{\mathbf{p_1}\mathbf{p_2}\mathbf{p_3}}) = [(u_2 - u_1)(v_3 - v_1) - (u_3 - u_1)(v_2 - v_1)]/2. \tag{A.1}$$

Thus we obtain the derivatives of a mapped coordinate $\mathbf{q} = (x, y)^T = \boldsymbol{\phi}(\mathbf{p})$ as

$$\begin{aligned}
\frac{\partial \mathbf{q}}{\partial u} &= \frac{1}{2A} \cdot \sum_{k=1}^{3} c_{uk}\mathbf{q_k} \\
\frac{\partial \mathbf{q}}{\partial v} &= \frac{1}{2A} \cdot \sum_{k=1}^{3} c_{vk}\mathbf{q_k} \\
& \qquad c_{u1} = v_2 - v_3\,, \quad c_{v1} = u_3 - u_2 \\
\textit{where} \qquad & \qquad c_{u2} = v_3 - v_1\,, \quad c_{v2} = u_1 - u_3 \\
& \qquad c_{u3} = v_1 - v_2\,, \quad c_{v3} = u_2 - u_1
\end{aligned} \tag{A.2}$$

Equation (A.2) shows that $\partial\mathbf{q}/\partial u$ and $\partial\mathbf{q}/\partial v$ do not depend on the spatial coordinates (u, v). This means that the Jacobian matrix of the mapping function over a triangular element is constant, i.e., is not varying with the coordinates (u, v). Therefore, compared to quadrangle or higher order elements, which do not have constant Jacobian matrix, using triangular elements greatly simplifies the calculation of deformation energy and its gradients.

A.1. Gradients for Green-Lagrange Strain

Gradients of the Froebenius norm of the Green-Lagrange strain tensor over a triangle t, $\|\boldsymbol{E}_t\|_F^2$, are as follows. The gradients with respect to triangle vertex coordinates, $\mathbf{q_k}$, are obtained by using the expressions in Equation (A.2).

$$\begin{aligned} &\frac{\partial \|\boldsymbol{E}\|_F^2}{\partial q_{ik}} = (a-1)\frac{\partial a}{\partial q_{ik}} + 2b\frac{\partial b}{\partial q_{ki}} + (c-1)\frac{\partial c}{\partial q_{ki}} \\ &where \quad \mathbf{q_k} = (q_{k1}, q_{k2})^T = (x_k, y_k)^T \end{aligned} \tag{A.3}$$

$$\begin{aligned} \frac{\partial a}{\partial x_k} &= 2 \cdot \frac{\partial \mathbf{q}^T}{\partial u} \cdot \frac{\partial}{\partial x_k}\left(\frac{\partial \mathbf{q}}{\partial v}\right) \\ &= 2 \cdot \frac{\partial \mathbf{q}^T}{\partial u} \cdot \frac{1}{2A} \cdot \begin{bmatrix} c_{uk} \\ 0 \end{bmatrix} \qquad \frac{\partial a}{\partial y_k} = \frac{1}{2A} \cdot 2c_{uk} \cdot \frac{\partial y}{\partial u} \\ &= \frac{1}{2A} \cdot 2c_{uk} \cdot \frac{\partial x}{\partial u} \end{aligned} \tag{A.4}$$

$$\frac{\partial c}{\partial x_k} = \frac{1}{2A} \cdot 2c_{vk} \cdot \frac{\partial x}{\partial v} \qquad \frac{\partial c}{\partial y_k} = \frac{1}{2A} \cdot 2c_{vk} \cdot \frac{\partial y}{\partial v} \tag{A.5}$$

$$\begin{aligned} \frac{\partial b}{\partial x_k} &= \frac{\partial \mathbf{q}^T}{\partial u} \cdot \frac{\partial}{\partial x_k}\left(\frac{\partial \mathbf{q}}{\partial v}\right) + \frac{\partial \mathbf{q}^T}{\partial v} \cdot \frac{\partial}{\partial x_k}\left(\frac{\partial \mathbf{q}}{\partial u}\right) \\ &= \frac{\partial \mathbf{q}^T}{\partial u} \cdot \frac{1}{2A} \cdot \begin{bmatrix} c_{vk} \\ 0 \end{bmatrix} + \frac{\partial \mathbf{q}^T}{\partial v} \cdot \frac{1}{2A} \cdot \begin{bmatrix} c_{uk} \\ 0 \end{bmatrix} \\ &= \frac{1}{2A}\left[c_{vk} \cdot \frac{\partial x}{\partial u} + c_{uk} \cdot \frac{\partial x}{\partial v}\right] \end{aligned} \tag{A.6}$$

$$\frac{\partial b}{\partial y_k} = \frac{1}{2A}\left[c_{vk} \cdot \frac{\partial y}{\partial u} + c_{uk} \cdot \frac{\partial y}{\partial v}\right]$$

A.2. Gradients for Cauchy's Strain

Gradients of the Froebenius norm of the Cauchy's strain tensor over a triangle t, $\|\boldsymbol{\varepsilon}_t\|_F^2$ are as follows. The gradients with respect to triangle vertex coordinates, $\mathbf{q_k}$, are obtained by using the expressions in Equation (A.2).

$$\|\boldsymbol{\varepsilon}\|_F^2 = \left(\frac{\partial x}{\partial u} - 1\right)^2 + \left(\frac{\partial y}{\partial v} - 1\right)^2 + \frac{1}{2}\left(\frac{\partial x}{\partial v} + \frac{\partial y}{\partial u}\right)^2 . \tag{A.7}$$

Then, via the chain rule

$$\begin{aligned}\frac{\partial\|\boldsymbol{\varepsilon}\|_F^2}{\partial q_{ki}} &= 2\left(\frac{\partial x}{\partial u} - 1\right)\cdot\frac{\partial}{\partial q_{ki}}\left(\frac{\partial x}{\partial u}\right) + 2\left(\frac{\partial y}{\partial v} - 1\right)\cdot\frac{\partial}{\partial q_{ki}}\left(\frac{\partial y}{\partial v}\right) \\ &\quad + \left(\frac{\partial x}{\partial v} + \frac{\partial y}{\partial u}\right)\cdot\left[\frac{\partial}{\partial q_{ki}}\left(\frac{\partial x}{\partial v}\right) + \frac{\partial}{\partial q_{ki}}\left(\frac{\partial y}{\partial u}\right)\right] \\ \textit{where} \quad \mathbf{q_k} &= (q_{k1}, q_{k2})^T = (x_k, y_k)^T .\end{aligned} \tag{A.8}$$

Thus, the derivatives with respect to each vertex coordinate becomes

$$\begin{aligned}\frac{\partial\|\boldsymbol{\varepsilon}\|_F^2}{\partial x_k} &= \frac{1}{2A}\left[2c_{uk}\left(\frac{\partial x}{\partial u} - 1\right) + c_{vk}\left(\frac{\partial x}{\partial v} + \frac{\partial y}{\partial u}\right)\right] \\ \frac{\partial\|\boldsymbol{\varepsilon}\|_F^2}{\partial y_k} &= \frac{1}{2A}\left[2c_{vk}\left(\frac{\partial y}{\partial v} - 1\right) + c_{uk}\left(\frac{\partial x}{\partial v} + \frac{\partial y}{\partial u}\right)\right]\end{aligned} \tag{A.9}$$

APPENDIX B: NON-RIGID REGISTRATION EXAMPLES

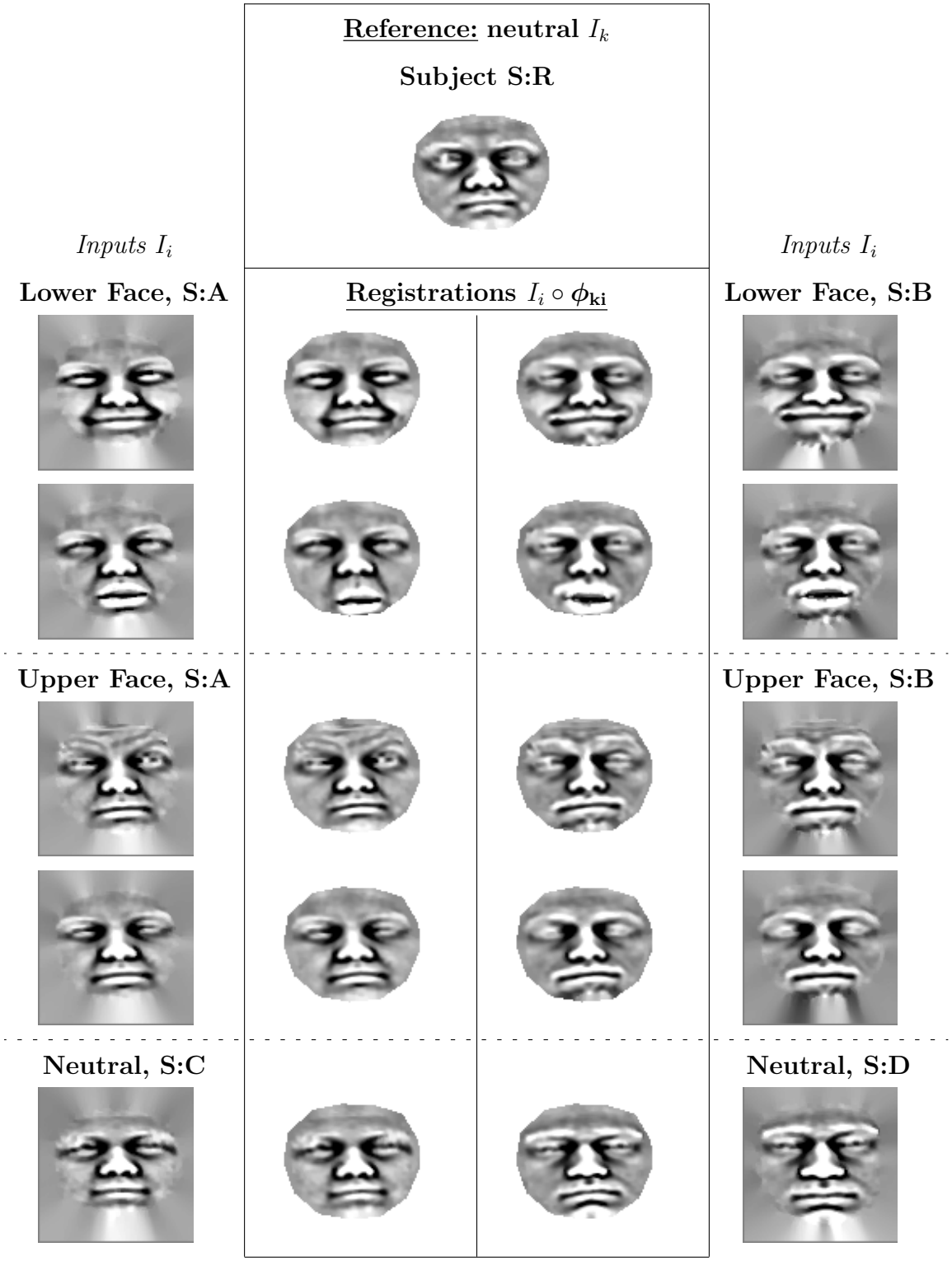


Figure B.1. Registration ($\rho = 500$) of different identity and expression faces onto neutral reference. The dominant deformations from second (top) row to fourth row are pulled lip corners, open mouth, raised eyebrows and squinted.

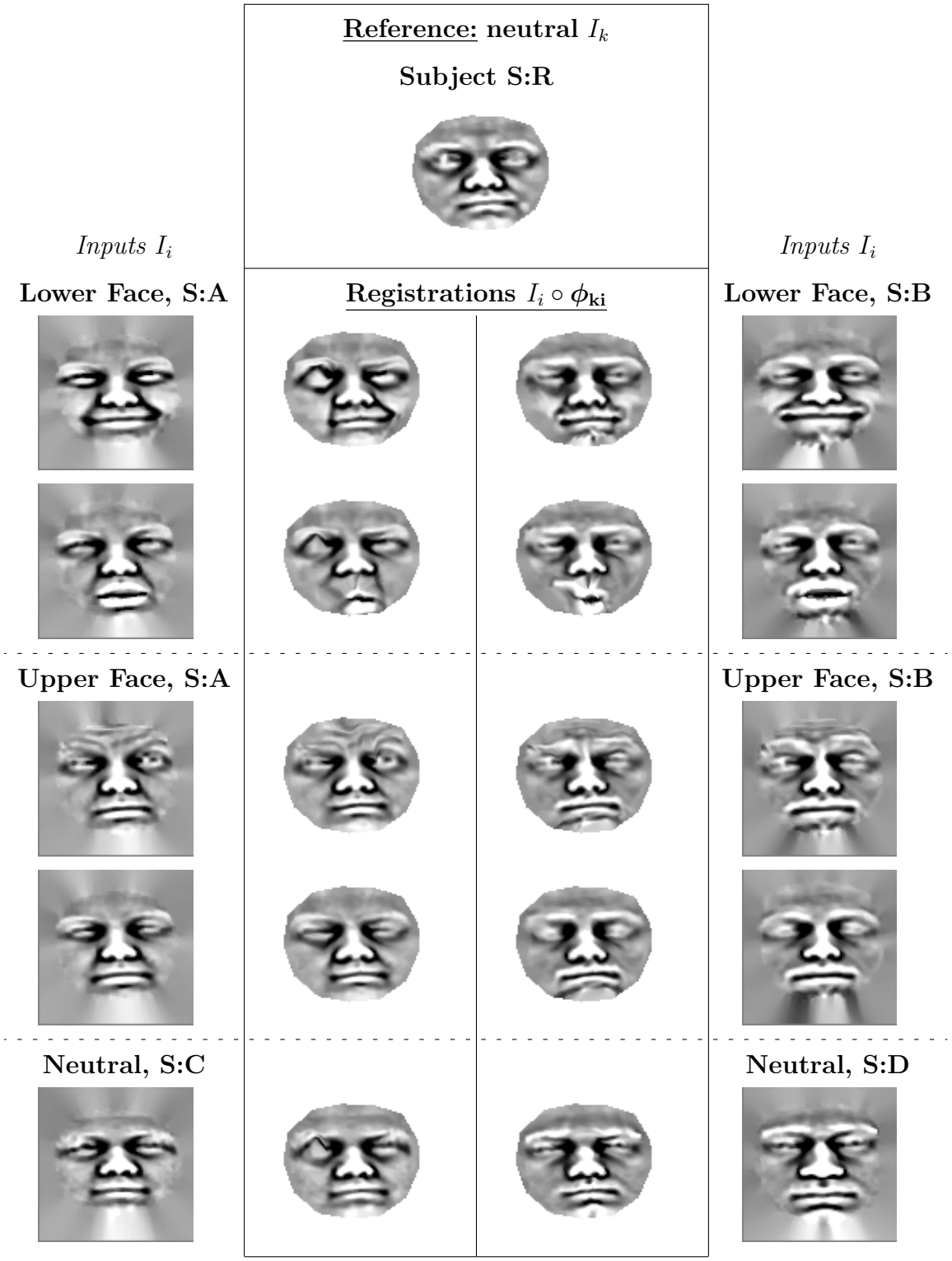


Figure B.2. Registration ($\rho = 0$) of different identity and expression faces onto neutral reference. The dominant deformations from second (top) row to fourth row are pulled lip corners, open mouth, raised eyebrows and squinted.

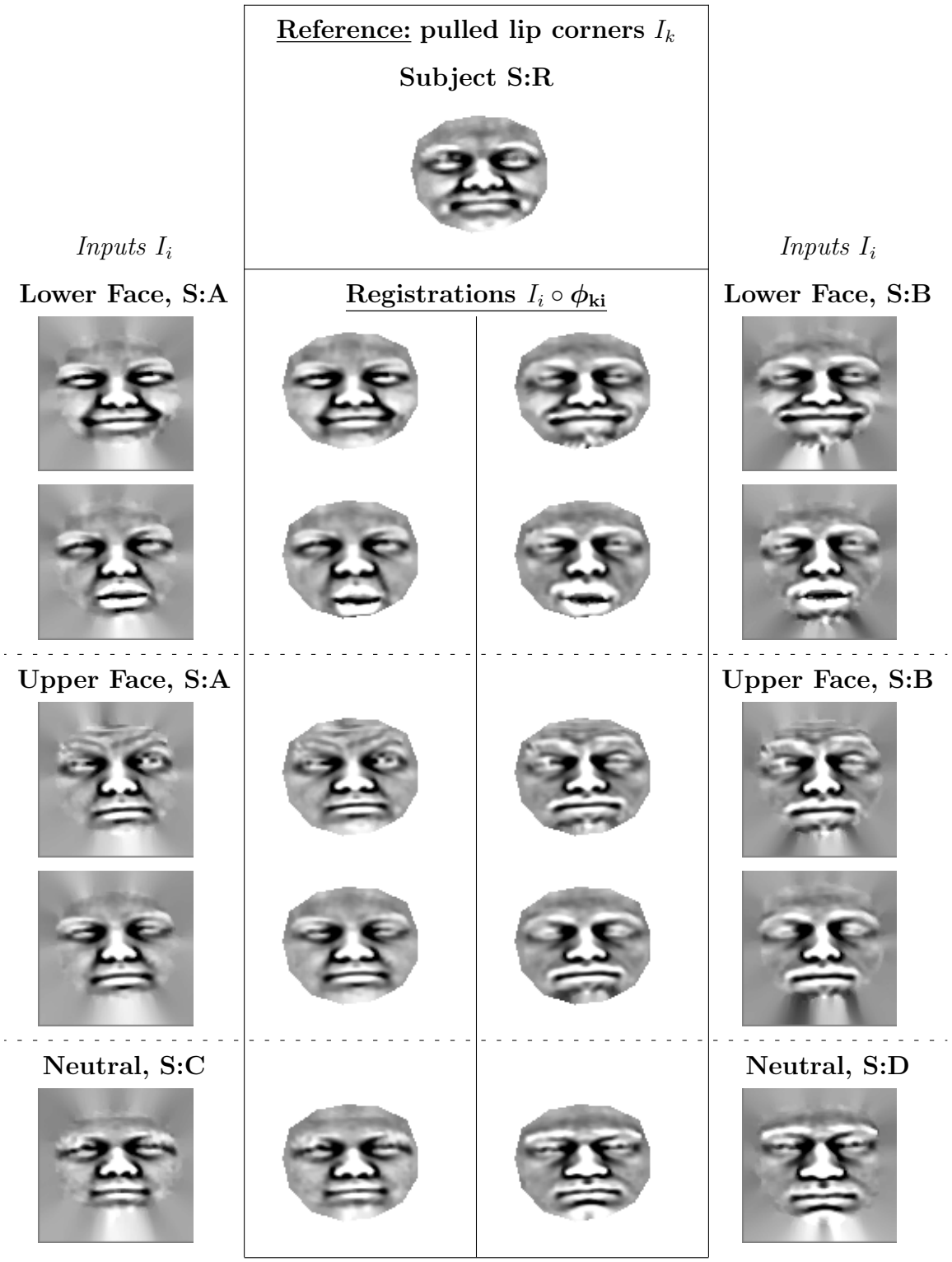


Figure B.3. Registration ($\rho = 500$) of different identity and expression faces onto pulled lip corners reference. The dominant deformations from second (top) row to fourth row are pulled lip corners, open mouth, raised eyebrows and squinted.

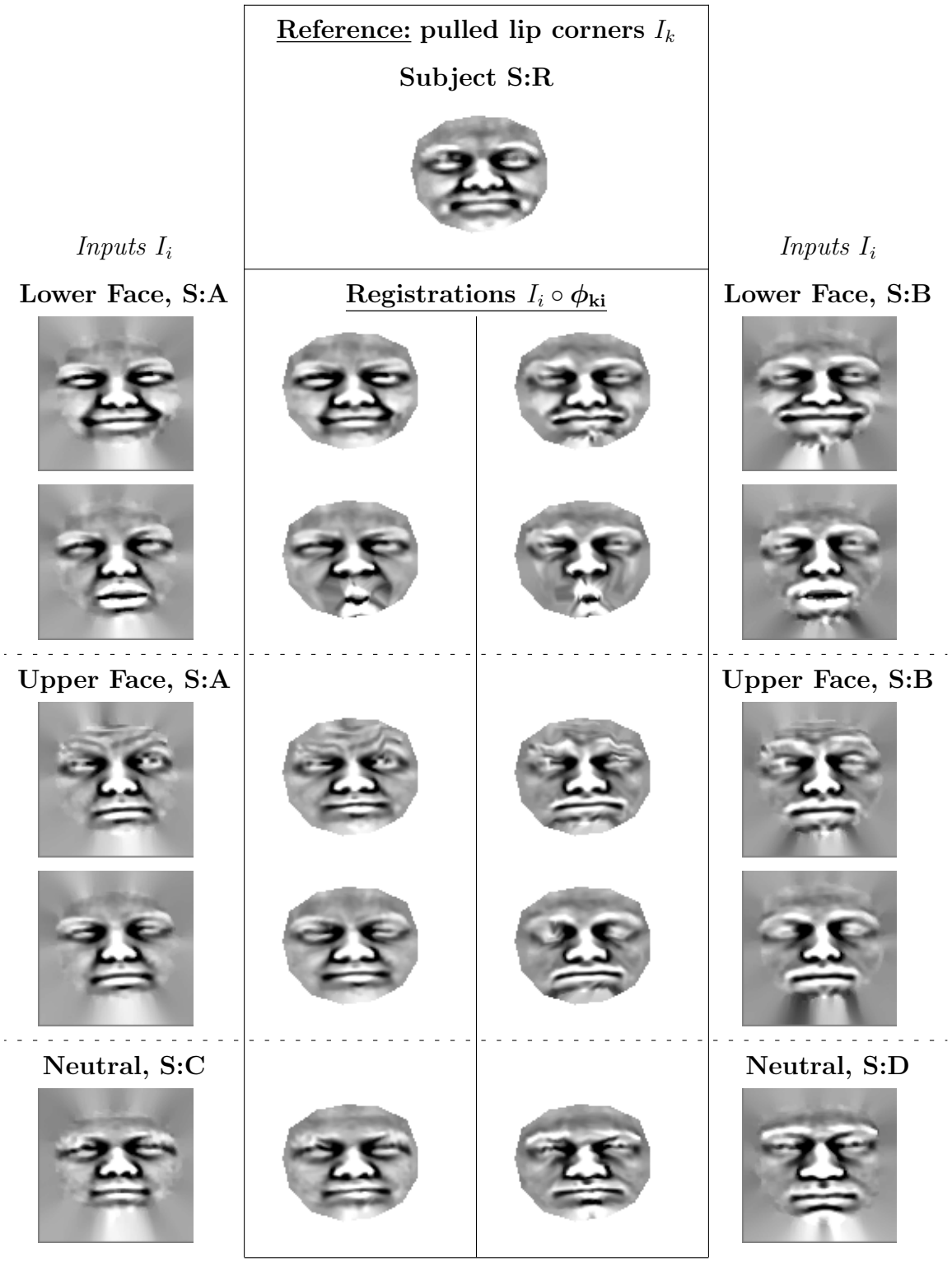


Figure B.4. Registration ($\rho = 0$) of different identity and expression faces onto pulled lip corners reference. The dominant deformations from second (top) row to fourth row are pulled lip corners, open mouth, raised eyebrows and squinted.

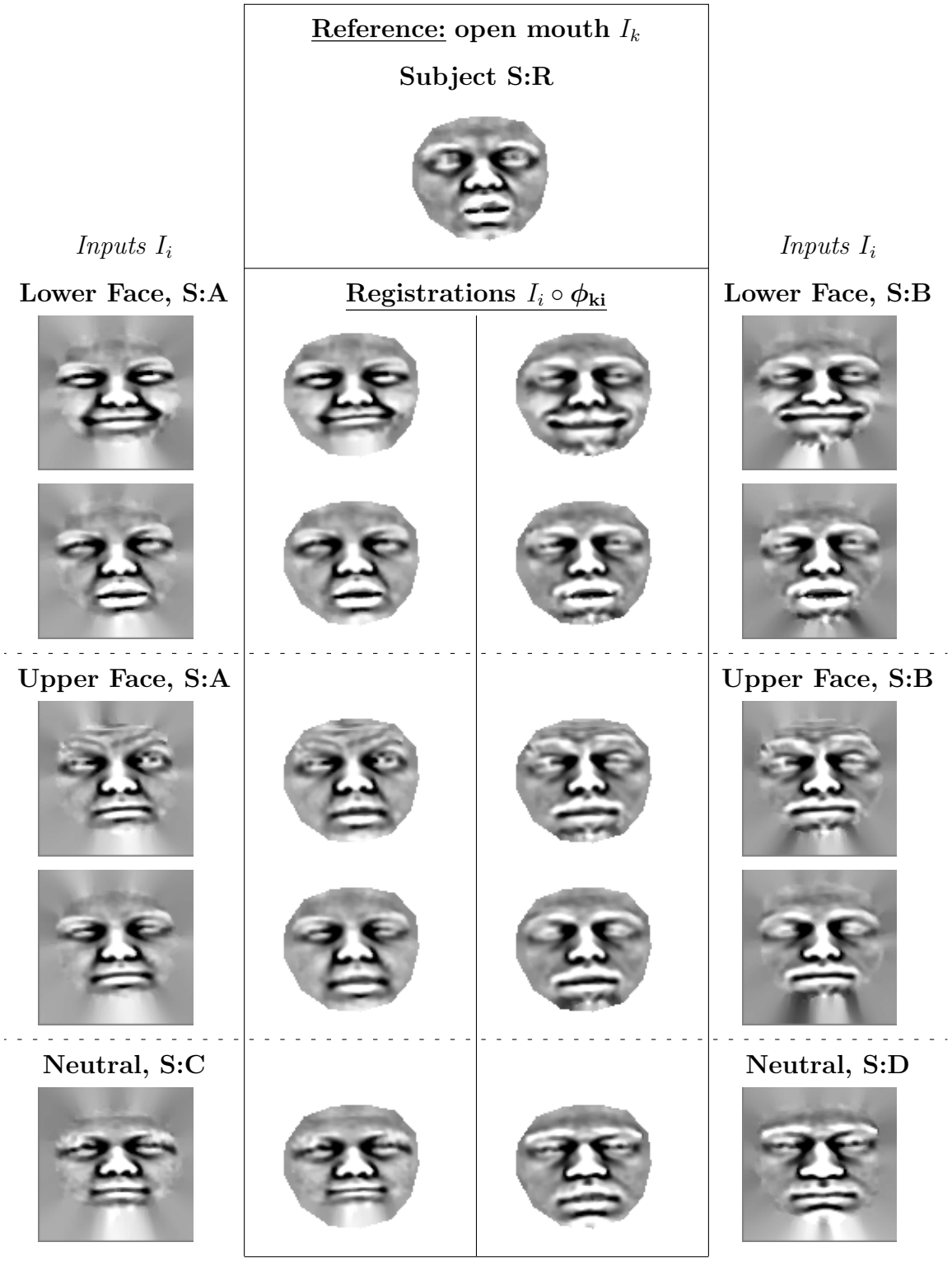


Figure B.5. Registration ($\rho = 500$) of different identity and expression faces onto open mouth reference. The dominant deformations from second (top) row to fourth row are pulled lip corners, open mouth, raised eyebrows and squinted.

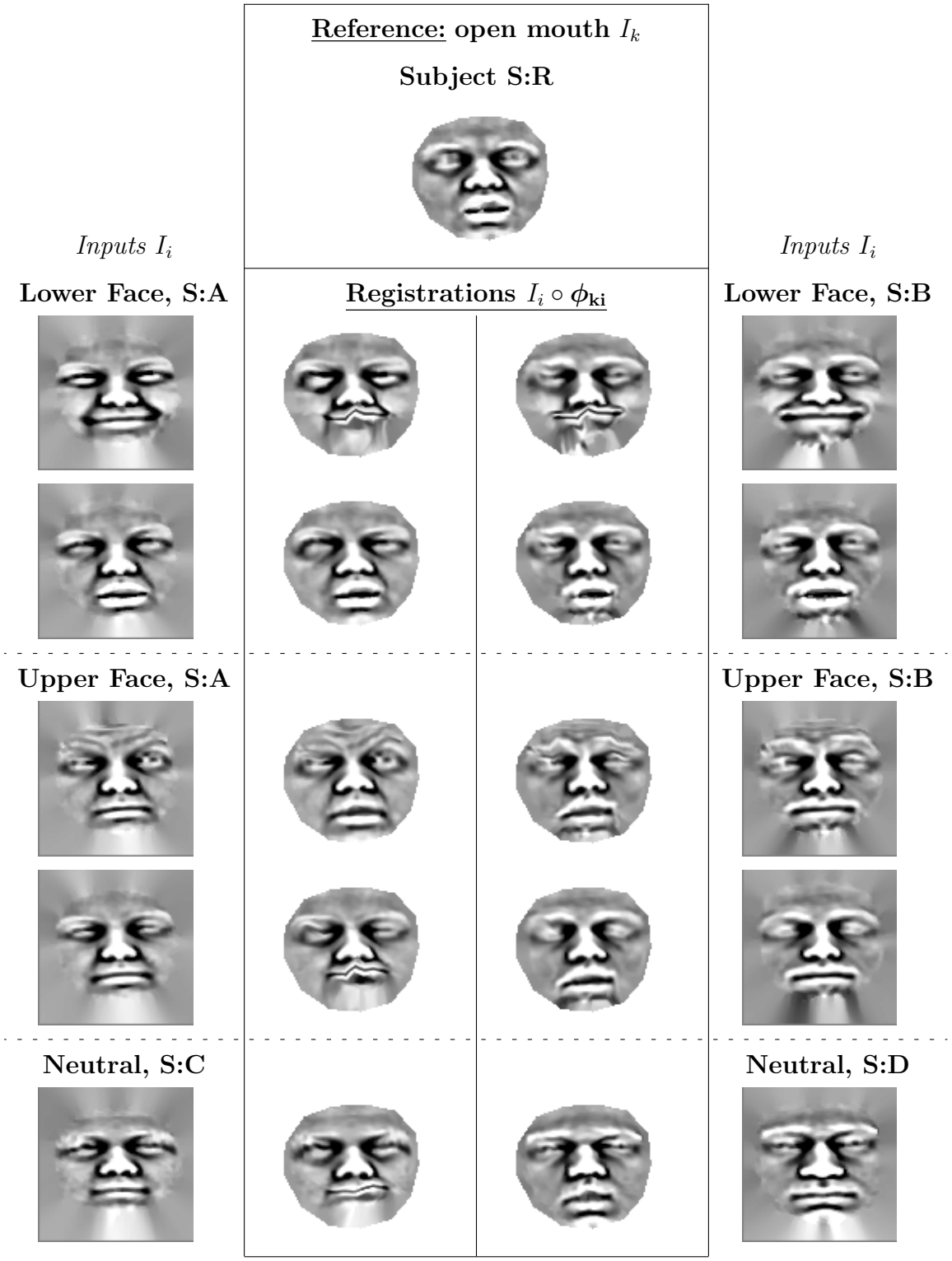


Figure B.6. Registration ($\rho = 0$) of different identity and expression faces onto open mouth reference. The dominant deformations from second (top) row to fourth row are pulled lip corners, open mouth, raised eyebrows and squinted.

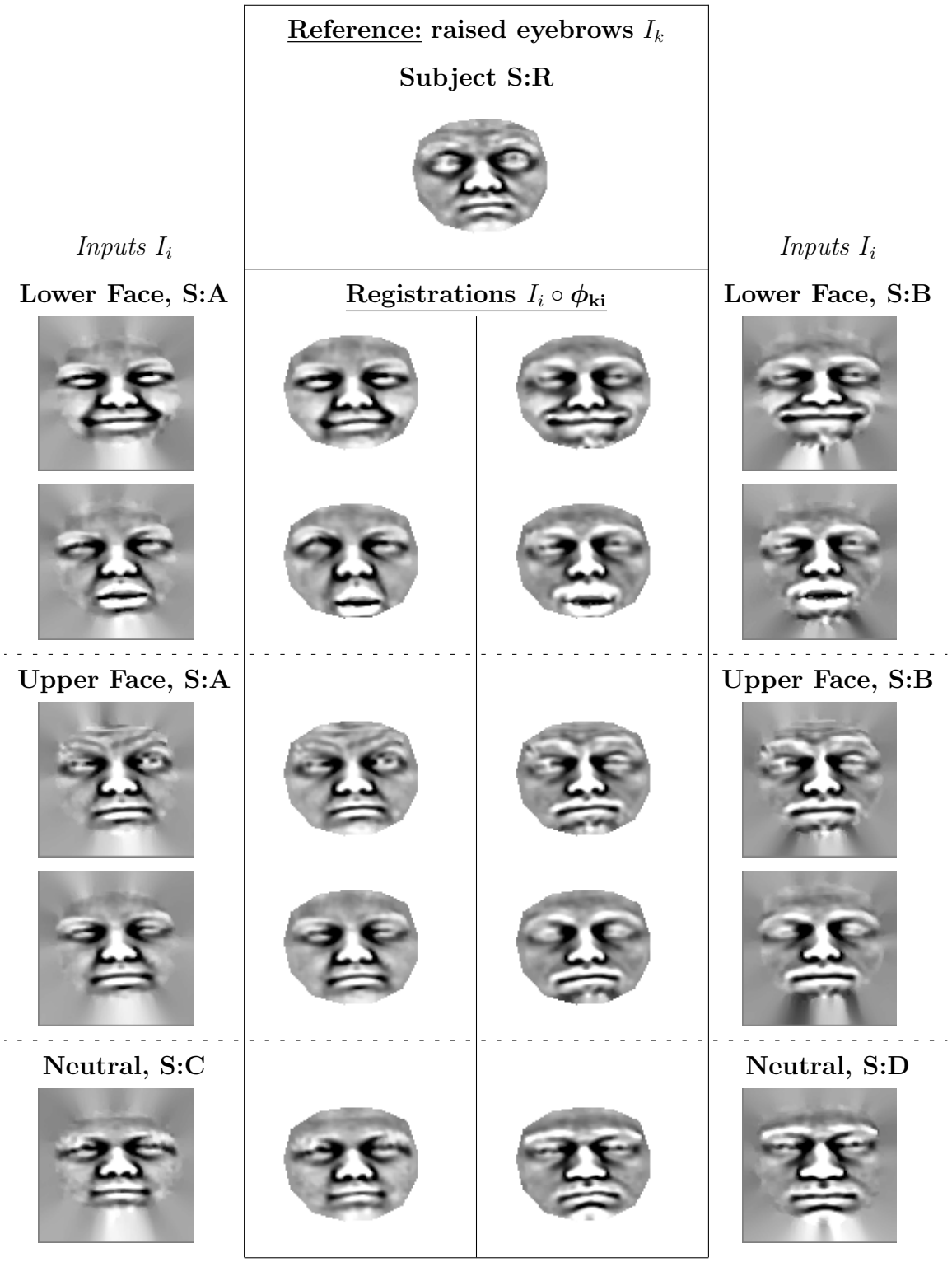


Figure B.7. Registration ($\rho = 500$) of different identity and expression faces onto raised eyebrows reference. The dominant deformations from second (top) row to fourth row are pulled lip corners, open mouth, raised eyebrows and squinted.

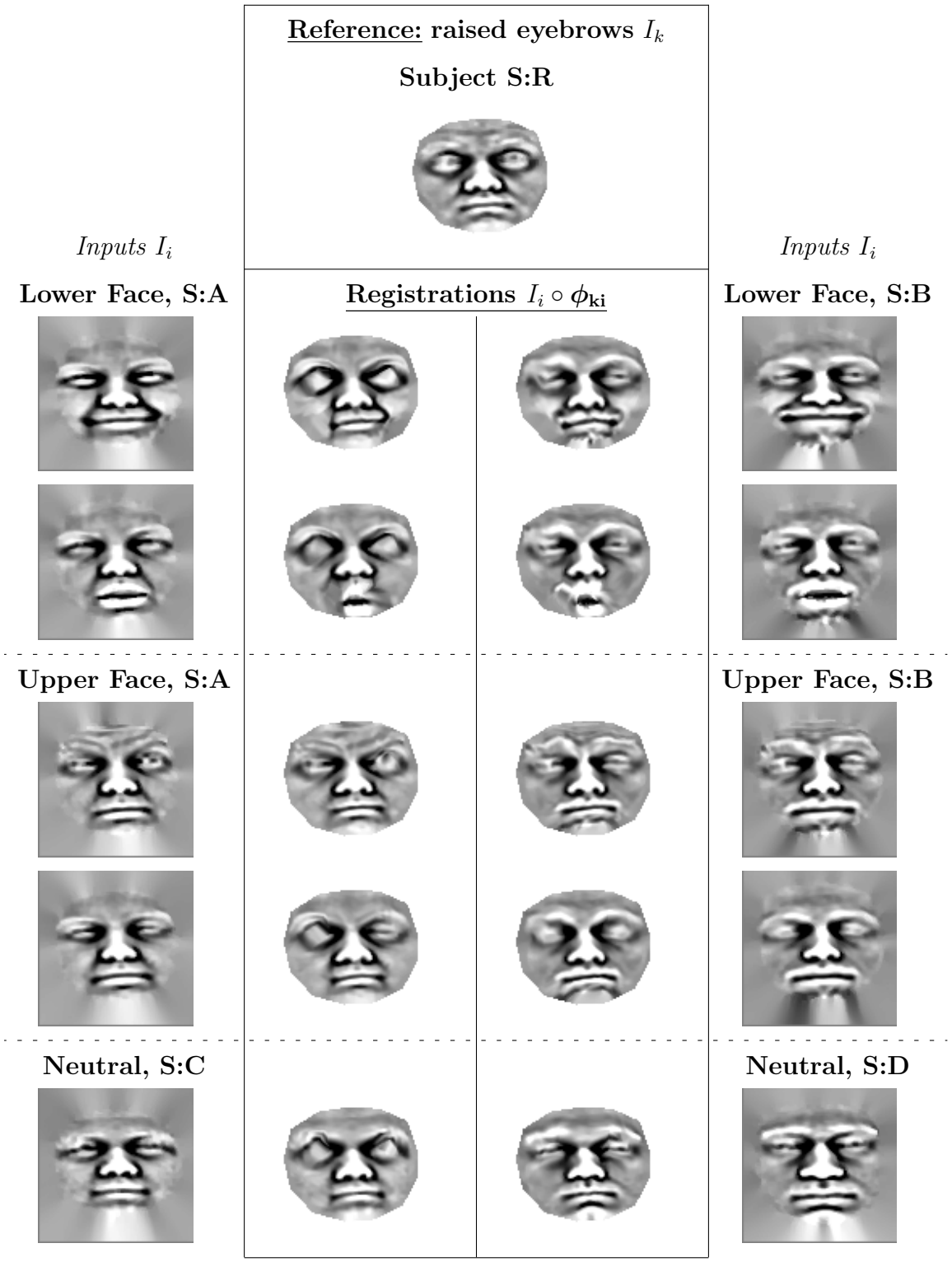


Figure B.8. Registration ($\rho = 0$) of different identity and expression faces onto raised eyebrows reference. The dominant deformations from second (top) row to fourth row are pulled lip corners, open mouth, raised eyebrows and squinted.

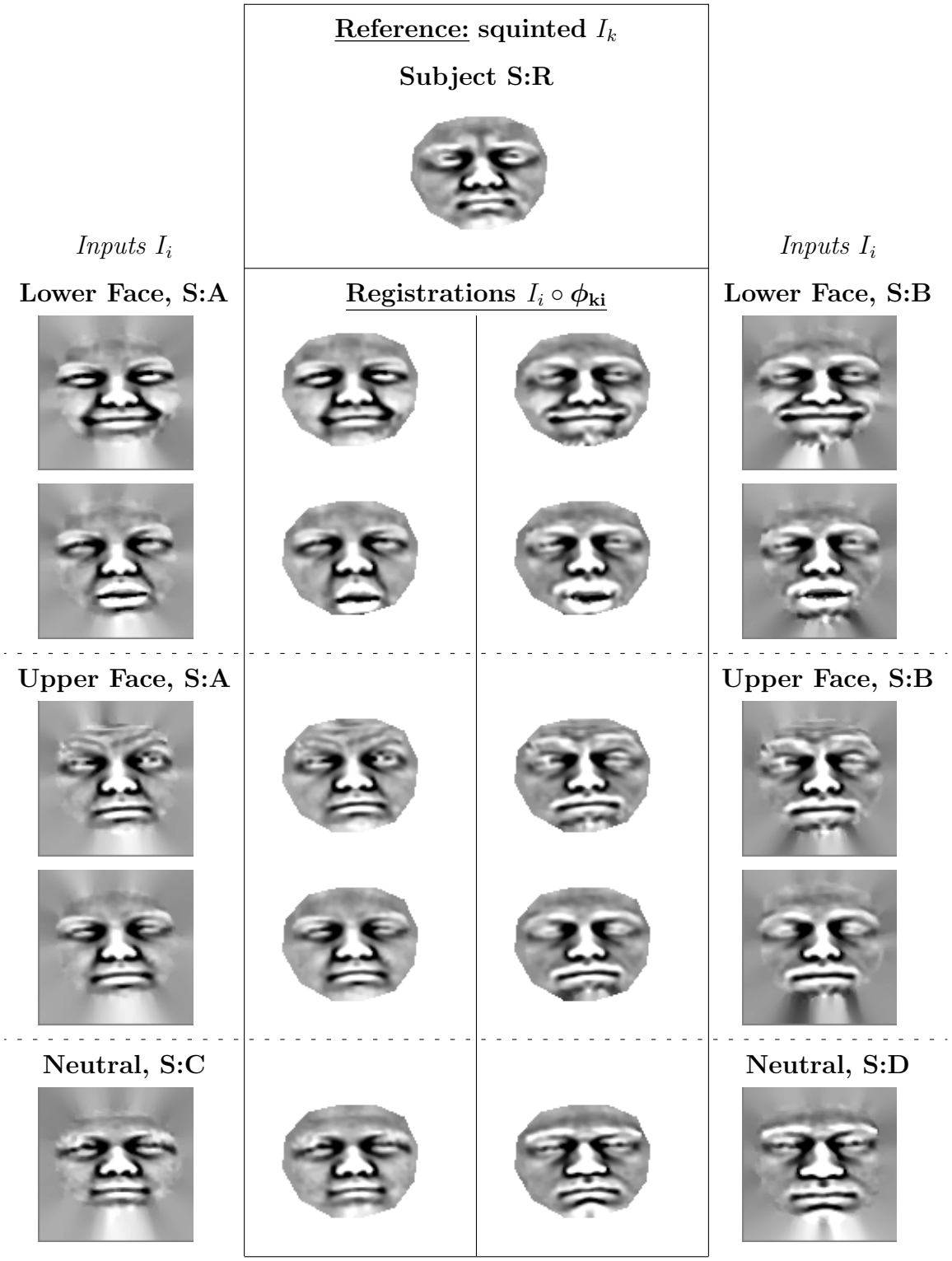


Figure B.9. Registration ($\rho = 500$) of different identity and expression faces onto squinted reference. The dominant deformations from second (top) row to fourth row are pulled lip corners, open mouth, raised eyebrows and squinted.

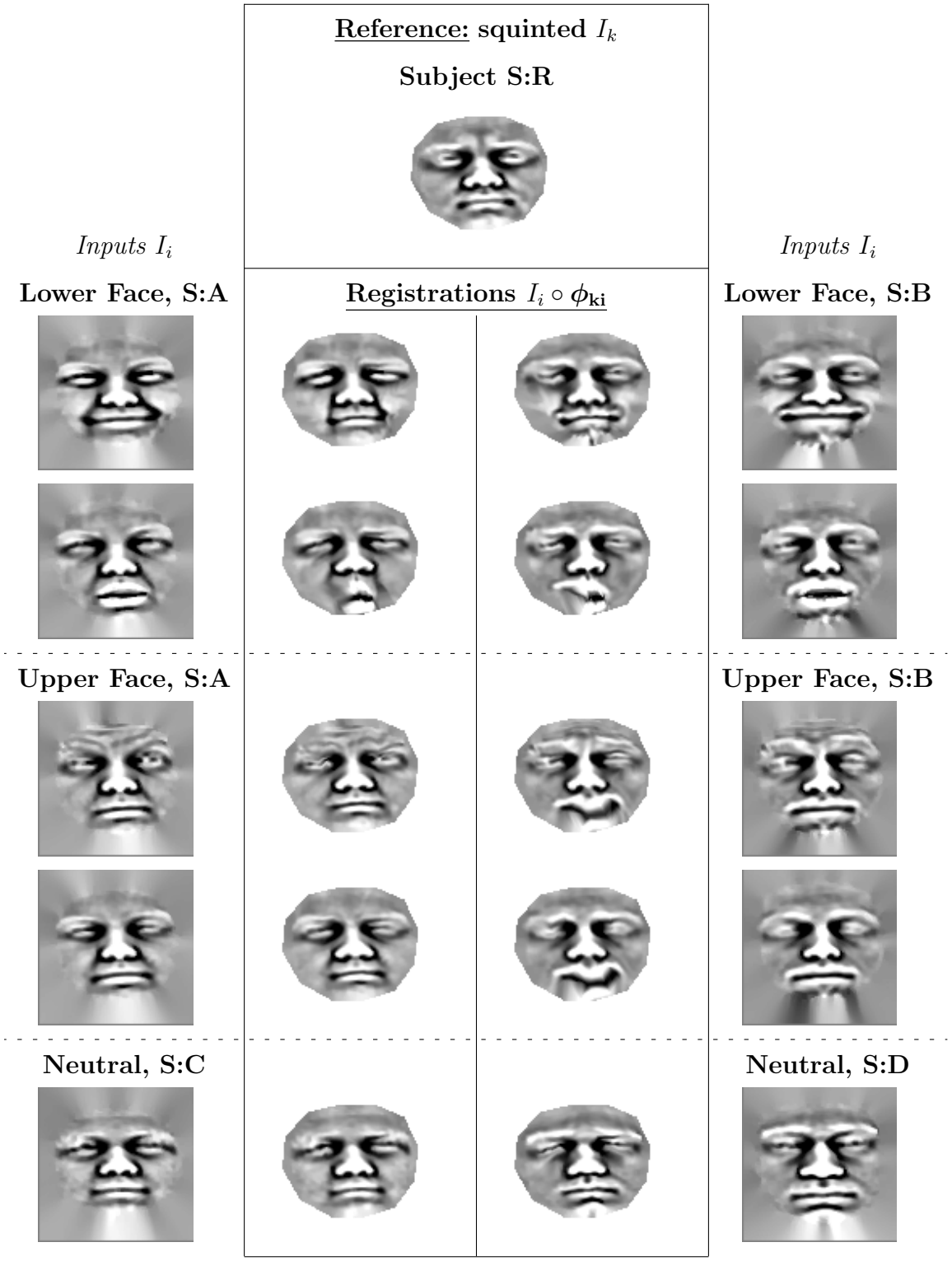


Figure B.10. Registration ($\rho = 0$) of different identity and expression faces onto squinted reference. The dominant deformations from second (top) row to fourth row are pulled lip corners, open mouth, raised eyebrows and squinted.

REFERENCES

1. Ekman, P. and W. V. Friesen, "Constants across cultures in the face and emotion", *Journal of Personality and Social Psychology*, Vol. 17, No. 2, pp. 124–129, 1971.

2. Baron-Cohen, S., A. Riviere, M. Fukushima, D. French, J. Hadwin, P. Cross, C. Bryant, and M. Sotillo, "Reading the Mind in the Face: A Cross-cultural and Developmental Study", *Visual Cognition*, Vol. 3, 1 March 1996.

3. Williams, A. C. C., "Facial expression of pain: An evolutionary account", *Behavioral and Brain Sciences*, Vol. 25, No. 04, pp. 439–455, 2002.

4. Krumhuber, E. and A. S. R. Manstead, "Are you joking? The moderating role of smiles in the perception of verbal statements", *Cognition and Emotion*, Vol. 23, No. 8, pp. 1504–1515, 2009.

5. Knapp, M. and J. Hall, *Nonverbal Communication in Human Interaction*, Wadsworth, Belmont, CA, sixth edition, 2010.

6. Pantic, M. and L. J. M. Rothkrantz, "Automatic Analysis of Facial Expressions: The State of the Art", *IEEE Transaction on Pattern Analysis and Machine Intelligence*, Vol. 22, No. 12, pp. 1424–1445, 2000.

7. Fasel, B. and J. Luettin, "Automatic Facial Expression Analysis: A Survey", *Pattern Recognition*, Vol. 36, No. 1, pp. 259–275, 2003.

8. Ekman, P. and W. Friesen, *Facial Action Coding System: A Technique for the Measurement of Facial Movement*, Consulting Psychologists Press, Palo Alto, 1978.

9. Tian, Y.-L., T. Kanade, and J. Cohn, "Facial Expression Analysis", Li, S. Z. and A. K. Jain (editors), *Handbook of Face Recognition*, Springer-Verlag, June 2004.

10. Zeng, Z., M. Pantic, G. I. Roisman, and T. S. Huang, "A Survey of Affect Recognition Methods: Audio, Visual, and Spontaneous Expressions", *IEEE Transaction on Pattern Analysis and Machine Intelligence*, Vol. 31, No. 1, pp. 39–58, January 2009.

11. Ekman, P., W. V. Friesen, and J. C. Hager, *Facial Action Coding System, The Manual on CD ROM*, 2002.

12. Hager, J. C. and P. Ekman, "The Asymmetry of Facial Actions is Inconsistent with Models of Hemispheric Specialization.", *Psychophysiology*, Vol. 22, No. 3, pp. 307–318, 1985.

13. Ekman, P., W. V. Friesen, and R. C. Simons, "Is the startle reaction an emotion?", *Journal of Personality and Social Psychology*, Vol. 49, No. 5, pp. 1416–26, 1985.

14. Frank, M. and P. Ekman, "Appearing truthful generalizes across different deception situations.", *Journal of Personality and Social Psychology*, Vol. 86, No. 3, pp. 486–95, 2004.

15. Griffin, K. M. and M. A. Sayette, "Facial Reactions to Smoking Cues Relate to Ambivalence About Smoking", *Psychology of Addictive Behaviors*, Vol. 22, No. 4, pp. 551 – 556, 2008.

16. Kamman, T., L. Muir, L. S. Koester, and D. M. Dimitrov, "Linking Maternal Perceptions to Behavior: Nurturing Attitudes and Facial Expressions of Affect.", *Parenting: Science & Practice*, Vol. 5, No. 3, pp. 237 – 258, 2005.

17. Larochette, A.-C., C. T. Chambers, and K. D. Craig, "Genuine, suppressed and faked facial expressions of pain in children", *Pain*, Vol. 126, No. 1-3, pp. 64 – 71, 2006.

18. Reed, L. I., M. A. Sayette, and J. F. Cohn, "Impact of Depression on Response to Comedy: A Dynamic Facial Coding Analysis", *Journal of Abnormal Psychology*,

Vol. 116, No. 4, pp. 804 – 809, 2007.

19. Ekman, P. and E. Rosenberg, *What the Face Reveals: Basic and Applied Studies of Spontaneous Expression using the Facial Action Coding System (FACS)*, Oxford University Press, New York, second edition, 2005.

20. Matsumoto, D. and P. Ekman, "Facial Expression Analysis", 2008, `http://www.scholarpedia.org/article/Facial_expression_analysis/`.

21. Sayette, M. A., J. F. Cohn, J. M. Wertz, M. A. Perrott, and D. J. Parrott, "A Psychometric Evaluation of the Facial Action Coding System for Assessing Spontaneous Expression", *Journal of Nonverbal Behavior*, Vol. 25, No. 3, pp. 167–185, 09 2001.

22. Donato, G., M. S. Bartlett, J. C. Hager, P. Ekman, and T. J. Sejnowski, "Classifying Facial Actions", *IEEE Transactions on Pattern Analysis and Machine Intelligence*, Vol. 21, pp. 974–989, 1999.

23. Lucey, S., I. Matthews, C. Hu, Z. Ambadar, F. de la Torre, and J. Cohn, "AAM Derived Face Representations for Robust Facial Action Recognition", *IEEE International Conference on Automatic Face and Gesture Recognition (FG)*, 2006.

24. Cootes, T. F., G. J. Edwards, and C. J. Taylor, "Active Appearance Models", *IEEE Transactions on Pattern Analysis and Machine Intelligence*, Vol. 23, pp. 681–685, 2001.

25. Bartlett, M. S., G. Littlewort, M. G. Frank, C. Lainscsek, I. Fasel, and J. R. Movellan, "Automatic recognition of facial actions in spontaneous expressions", *Journal of Multimedia*, Vol. 1, No. 6, pp. 22–35, 2006.

26. Cohn, J. F., A. J. Zlochower, J. Lien, T. Kanade, and A. F. Analysis, "Automated Face Analysis by Feature Point Tracking Has High Concurrent Validity with Manual FACS Coding", *Psychophysiology*, Vol. 36, pp. 35–43, 1999.

27. Lucey, S., A. B. Ashraf, and J. Cohn, "Investigating Spontaneous Facial Action Recognition through AAM Representations of the Face", *Face Recognition Book*, Pro Literatur Verlag, Mammendorf, Germany, April 2007.

28. Brick, T., M. Hunter, and J. Cohn, "Get the FACS fast: Automated FACS face analysis benefits from the addition of velocity", pp. 1 –7, sep. 2009.

29. Yang, P., Q. Liu, and D. N. Metaxas, "Boosting Coded Dynamic Features for Facial Action Units and Facial Expression Recognition", *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 1–6, 2007.

30. Koelstra, S., M. Pantic, and I. Patras, "A Dynamic Texture-Based Approach to Recognition of Facial Actions and Their Temporal Models", *IEEE Transactions on Pattern Analysis and Machine Intelligence*, Vol. 32, No. 11, pp. 1940 –1954, 2010.

31. Tong, Y., W. Liao, and Q. Ji, "Facial Action Unit Recognition by Exploiting Their Dynamic and Semantic Relationships", *IEEE Transactions on Pattern Analysis and Machine Intelligence*, Vol. 29, No. 10, pp. 1683–1699, 2007.

32. Tong, Y., J. Chen, and Q. Ji, "A Unified Probabilistic Framework for Spontaneous Facial Action Modeling and Understanding", *IEEE Transactions on Pattern Analysis and Machine Intelligence*, Vol. 32, No. 2, pp. 258–273, February 2010.

33. Tian, Y.-L., T. Kanade, and J. Cohn, "Recognizing action units for facial expression analysis", *IEEE Transaction on Pattern Analysis and Machine Intelligence*, Vol. 23, No. 2, pp. 97–115, Feb 2001.

34. Pantic, M. and I. Patras, "Dynamics of facial expression: recognition of facial actions and their temporal segments from face profile image sequences", *IEEE Transactions on Systems, Man, and Cybernetics, Part B: Cybernetics*, Vol. 36, No. 2, pp. 433–449, 2006.

35. Wen, Z. and T. S. Huang, "Capturing Subtle Facial Motions in 3D Face Tracking", *IEEE International Conference on Computer Vision (ICCV)*, Vol. 2, p. 1343, 2003.

36. Cohen, I., N. Sebe, Y. Sun, M. S. Lew, and T. S. Huang, "Towards authentic emotion recognition", *IEEE International Conference on Systems, Man and Cybernetics*, 2004.

37. Cootes, T. F., C. J. Taylor, D. H. Cooper, and J. Graham, "Active Shape Models - Their Training and Application", *Computer Vision and Image Understanding*, Vol. 61, No. 1, pp. 38–59, January 1995.

38. Cristinacce, D. and T. Cootes, "Feature Detection and Tracking with Constrained Local Models", 17^{th} *British Machine Vision Conference, Edinburgh, UK*, pp. 929–938, 2006.

39. Saragih, J. M., S. Lucey, and J. Cohn, "Face Alignment through Subspace Constrained Mean-Shifts", *IEEE International Conference of Computer Vision (ICCV)*, September 2009.

40. Whitehill, J., G. Littlewort, I. Fasel, M. Bartlett, and J. Movellan, "Toward Practical Smile Detection", *IEEE Transactions on Pattern Analysis and Machine Intelligence*, Vol. 31, pp. 2106–2111, 2009.

41. Pantic, M. and L. Rothkrantz, "Facial action recognition for facial expression analysis from static face images", *IEEE Transactions on Systems, Man, and Cybernetics, Part B: Cybernetics*, Vol. 34, No. 3, pp. 1449–1461, June 2004.

42. Mpiperis, I., S. Malassiotis, and M. Strintzis, "Bilinear elastically deformable models with application to 3D face and facial expression recognition", *IEEE International Conference on Automatic Face and Gesture Recognition (FG)*, Amsterdam, Netherlands, 2008.

43. Gokberk, B., H. Dutagaci, A. Ulas, L. Akarun, and B. Sankur, "Representation plurality and fusion for 3-D face recognition", *IEEE Transactions on Systems Man and Cybernetics Part B*, Vol. 38, No. 1, pp. 155–173, February 2008.

44. Wang, J., L. Yin, X. Wei, and Y. Sun, "3D Facial Expression Recognition Based on Primitive Surface Feature Distribution", *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Washington, DC, USA, 2006.

45. Soyel, H. and H. Demirel, "Facial Expression Recognition Using 3D Facial Feature Distances", *International Conference on Image Analysis and Recognition, (ICIAR)*, Montreal, Canada, 2007.

46. Mpiperis, I., S. Malasiotis, V. Petridis, and M. G. Strintzis, "3D Facial Expression Recognition Using Swarm Intelligence", *IEEE International Conference on Accoustics, Speech and Signal Processing (ICASSP)*, Las Vegas, Nevada, USA, 2008.

47. Tang, H. and T. Huang, "3D facial expression recognition based on automatically selected features", *IEEE CVPR Workshop on 3D Face Processing*, Anchorage, Alaska, USA, 2008.

48. Sun, Y., M. Reale, and L. Yin, "Recognizing partial facial action units based on 3D dynamic range data for facial expression recognition", *IEEE International Conference on Automatic Face and Gesture Recognition (FG)*, Amsterdam, Netherlands, 2008.

49. Tsalakanidou, F. and S. Malassiotis, "Real-time 2D+3D facial action and expression recognition", *Pattern Recognition*, Vol. 43, No. 5, pp. 1763 – 1775, 2010.

50. Pantic, M. and L. J. M. Rothkrantz, "An Expert System for Recognition of Facial Actions and their Intensity", *AAAI/IAAI*, pp. 1026–1033, 2000.

51. Yang, P., Q. Liu, and D. N. Metaxas, "IEEE RankBoost with l_1 regularization

for Facial Expression Recognition and Intensity Estimation", *International Conference of Computer Vision (ICCV)*, September 2009.

52. Mahoor, M., S. Cadavid, D. Messinger, and J. Cohn, "A framework for automated measurement of the intensity of non-posed Facial Action Units", *IEEE CVPR Workshop on Human Communicative Behaviour Analysis*, 2009.

53. Koelstra, S. and M. Pantic, "Non-rigid registration using free-form deformations for recognition of facial actions and their temporal dynamics", *IEEE International Conference on Automatic Face and Gesture Recognition (FG)*, IEEE Computer Society Press, Washington, DC, USA, September 2008.

54. Blanz, V. and T. Vetter, "A Morphable Model for the Synthesis of 3D Faces", *ACM Transactions on Graphics: Proc. of SIGGRAPH*, pp. 187–194, 1999.

55. Ramanathan, S., A. A. Kassim, Y. V. Venkatesh, and W. S. Wah, "Human Facial Expression Recognition using a 3D Morphable Model", *IEEE International Conference on Image Processing (ICIP)*, pp. 661–664, 2006.

56. Tenenbaum, J. B. and W. T. Freeman, "Separating style and content with bilinear models", *Neural Computation*, 2000.

57. Bookstein, F., "Principal Warps: Thin-Plate Splines and the Decomposition of Deformations", *IEEE Transactions on Pattern Analysis and Machine Intelligence*, Vol. 11, pp. 567–585, 1989.

58. Nielson, G. M., "Scattered Data Modeling", *IEEE Computer Graphics and Applications*, Vol. 13, pp. 60–70, 1993.

59. Praun, E., W. Sweldens, and P. Schrder, "Consistent Mesh Parameterizations", *ACM Transactions on Graphics: Proc. of SIGGRAPH*, pp. 179–184, August 2001.

60. Allen, B., B. Curless, and Z. Popovic, "The Space of Human Body Shapes", *ACM Trans. Graph*, Vol. 22, pp. 587–594, 2003.

61. Shelton, C. R., "Morphable Surface Models", *International Journal of Computer Vision*, Vol. 38, No. 1, pp. 75–91, 2000.

62. Yin, L., X. Wei, P. Longo, and A. Bhuvanesh, "Analyzing Facial Expressions Using Intensity-Variant 3D Data For Human Computer Interaction", *ICPR (1)*, pp. 1248–1251, 2006.

63. Bronstein, A., M. Bronstein, and R. Kimmel, "Calculus of Nonrigid Surfaces for Geometry and Texture Manipulation", *IEEE Transactions on Visualization and Computer Graphics*, Vol. 13, pp. 902–913, 2007.

64. Wang, Y., M. Gupta, S. Zhang, S. Wang, X. Gu, D. Samaras, and P. Huang, "High Resolution Tracking of Non-Rigid 3D Motion of Densely Sampled Data Using Harmonic Maps", *IEEE International Conference on Computer Vision (ICCV)*, Beijing, China, 2005.

65. Litke, N., M. Droske, M. Rumpf, and P. Schröder, "An Image Processing Approach to Surface Matching", Desbrun, M. and H. Pottmann (editors), *Third Eurographics Symposium on Geometry Processing*, pp. 207–216, 2005.

66. Hormann, K., B. Lévy, and A. Sheffer, *Mesh Parameterization: Theory and Practice*, HAL - CCSD, 2007, `http://hal.inria.fr/inria-00186795/en/`.

67. Floater, M. S. and K. Hormann, "Surface Parameterization: a Tutorial and Survey", pp. 157–186, 2005.

68. Zitov, B. and J. Flusser, "Image registration methods: a survey", *Image and Vision Computing*, Vol. 21, pp. 977–1000, 2003.

69. Crum, W. R., T. Hartkens, and D. L. G. Hill, "Non-rigid image registration: theory and practice.", *Br J Radiol*, Vol. 77 Spec No 2, pp. S140–53, 2004.

70. Yin, L., X. Wei, Y. Sun, J. Wang, and M. J. Rosato, "A 3D Facial Expression Database For Facial Behavior Research", *IEEE International Conference on Au-*

tomatic Face and Gesture Recognition (FG), Southampton, UK, 2006.

71. Yin, L., X. Chen, Y. Sun, T. Worm, and M. Reale, "A high-resolution 3D dynamic facial expression database", *FG*, pp. 1–6, 2008.

72. "Inspeck Mega Capturor 2", 2007, `http://www.inspeck.com/`.

73. Kanade, T., J. Cohn, and Y.-L. Tian, "Comprehensive Database for Facial Expression Analysis", *IEEE International Conference on Automatic Face and Gesture Recognition (FG)*, pp. 46 – 53, March 2000.

74. Besl, P. and N. McKay, "A Method for Registration of 3-D Shapes", *IEEE Transactions on Pattern Analysis and Machine Intelligence*, Vol. 14, pp. 239–256, 1992.

75. Wang, S., Y. Wang, M. Jin, X. D. Gu, and D. Samaras, "Conformal Geometry and Its Applications on 3D Shape Matching, Recognition, and Stitching", *IEEE Transactions on Pattern Analysis and Machine Intelligence*, Vol. 29, No. 7, pp. 1209–1220, 2007.

76. Levy, B., S. Petitjean, N. Ray, and J. Maillot, "Least squares conformal maps for automatic texture atlas generation", *ACM Transactions on Graphics: Proc. of SIGGRAPH*, Vol. 21, No. 3, pp. 362–371, 2002.

77. Telea, A., "An image inpainting technique based on the fast marching method", *Graphics Tools*, Vol. 9, No. 1, pp. 25–36, 2004.

78. Koenderink, J. J. and A. J. van Doorn, "Surface shape and curvature scales", *Image and Vision Computing*, Vol. 10, No. 8, pp. 557–564, 1992.

79. Daugman, J., "Complete Discrete 2D Gabor Transforms by Neural Networks for Image Analysis and Compression", *IEEE Transactions on Acoustics, Speech, and Signal Processing*, Vol. 36, No. 7, pp. 1169–1179, July 1988.

80. Bartlett, M., J. Movellan, and T. Sejnowski, "Face recognition by independent

component analysis", *IEEE Transactions on Neural Networks*, Vol. 13, No. 6, pp. 1450–1464, November 2002.

81. Li, S. Z., X. Hou, H. Zhang, and Q. Cheng, "Learning Spatially Localized, Parts-Based Representation", *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 207–212, 2001.

82. Buciu, I. and I. Pitas, "Application of non-Negative and Local non Negative Matrix Factorization to Facial Expression Recognition", *International Conference on Pattern Recognition (ICPR)*, 2004.

83. Hoyer, P. O., "Non-negative Matrix Factorization with Sparseness Constraints", *Journal of Machine Learning Research*, Vol. 5, pp. 1457–1469, 2004.

84. Vapnik, V., *The Nature of Statistical Learning Theory*, Springer, New York, 1995.

85. Zhang, L., N. Snavely, B. Curless, and S. M. Seitz, "Spacetime Faces: High-Resolution Capture for Modeling and Animation", *ACM Annual Conference on Computer Graphics*, pp. 548–558, August 2004.

86. Karpinsky, N. and S. Zhang, "High-resolution, real-time 3D imaging with fringe analysis", *Journal of Real-Time Image Processing*, pp. 1–12, 2010.

87. Bajcsy, R. and S. Kovačič, "Multiresolution elastic matching", *Computer Vision, Graphisc and Image Processing*, Vol. 46, No. 1, pp. 1–21, 1989.

88. Bro-Nielsen, M., *Medical Image Registration and Surgery Simulation*, Ph.D. thesis, Informatics and Mathematical Modelling, Technical University of Denmark, DTU, Richard Petersens Plads, Building 321, DK-2800 Kgs. Lyngby, 1996, iMM-PHD-1996-25.

89. Thirion, J.-P., "Image matching as a diffusion process: an analogy with Maxwell's demons", *Medical Image Analysis*, Vol. 2, No. 3, pp. 243 – 260, 1998.

90. Yanovsky, I., C. L. Guyader, A. Leow, P. Thompson, and L. Vese, "Nonlinear Elastic Registration with Unbiased Regularization in Three Dimensions", *The MIDAS Journal*, , No. 549, pp. 56 – 67, 2008.

91. Weickert, J. and H. Scharr, "A Scheme for Coherence-Enhancing Diffusion Filtering with Optimized Rotation Invariance", *Journal of Visual Communication and Image Representation*, Vol. 13, No. 1/2, pp. 103–118, March 2002.

92. Alliez, P., M. Meyer, and M. Desbrun, "Interactive Geometry Remeshing", *ACM Transactions on Graphics: Proc. of SIGGRAPH*, Vol. 21, pp. 347–354, 2002.

93. Boissonnat, J. D., O. Devillers, S. Pion, M. Teillaud, and M. Yvinec, "Triangulations in CGAL", *Computational Geometry*, Vol. 22, No. 1-3, pp. 5–19, 2002.

94. Shewchuk, J. R., "Mesh generation for domains with small angles", *Annual Symposium on Computational Geometry*, 2000.

95. Kanade, T. and B. Lucas, "An Iterative Image Registration Technique with an Application to Stereo Vision", *International Joint Conference on Artificial Intelligence*, pp. 674–679, Vancouver, 1981.

96. Yin, L., X. Wei, Y. Sun, J. Wang, and M. J. Rosato, "A 3D Facial Expression Database For Facial Behavior Research", *IEEE International Conference on Automatic Face and Gesture Recognition (FG)*, Washington, DC, USA, 2006.

97. Cootes, T., C. Twining, K. Babalola, and C. Taylor, "Diffeomorphic statistical shape models", *Image and Vision Computing*, Vol. 26, No. 3, pp. 326 – 332, 2008, 15th Annual British Machine Vision Conference.

98. Cootes, T. F., C. J. Twining, V. S. Petrovic, K. O. Babalola, and C. J. Taylor, "Computing Accurate Correspondences across Groups of Images", *IEEE Transactions on Pattern Analysis and Machine Intelligence*, Vol. 32, pp. 1994–2005, 2010.

99. Charpiat, G., R. Keriven, and O. Faugeras, "Image Statistics based on Diffeomorphic Matching", *IEEE International Conference on Computer Vision (ICCV)*, Vol. 1, pp. 852–857, 2005.

100. Modrow, D., C. Laloni, G. Doemens, and G. Rigoll, "A novel sensor system for 3D face scanning based on infrared coded light", SPIE, 2008.