Robust Facial Landmark Localization Under Simultaneous Real-World Degradations

Submitted in partial fulfillment of the requirements for
the degree of
Doctor of Philosophy
in the
Department of Electrical and Computer Engineering

Keshav Thirumalai Seshadri
kseshadr@andrew.cmu.edu

B. Tech., Electronics & Communication Engineering, VNIT Nagpur
M.S., Electrical and Computer Engineering, Carnegie Mellon University

Carnegie Mellon University
Pittsburgh, PA

December 2015

Carnegie Mellon University

CARNEGIE INSTITUTE OF TECHNOLOGY

THESIS

SUBMITTED IN PARTIAL FULFILLMENT OF THE REQUIREMENTS

FOR THE DEGREE OF Doctor of Philosophy

TITLE Robust Facial Landmark Localization Under Simultaneous Real-World Degradations

PRESENTED BY Keshav Thirumalai Seshadrikseshadr@andrew.cmu.edu

ACCEPTED BY THE DEPARTMENT OF

Electrical and Computer Engineering

______________________________ ____________

ADVISOR, MAJOR PROFESSOR DATE

______________________________ ____________

DEPARTMENT HEAD DATE

APPROVED BY THE COLLEGE COUNCIL

______________________________ ____________

DEAN DATE

THESIS COMMITTEE

Professor Marios Savvides, Advisor
Department of Electrical and Computer Engineering
Carnegie Mellon University
marioss@andrew.cmu.edu

Professor Vijayakumar Bhagavatula
Department of Electrical and Computer Engineering
Carnegie Mellon University
kumar@ece.cmu.edu

Professor Arun Ross
Department of Computer Science and Engineering
Michigan State University
rossarun@cse.msu.edu

Dr. Saad Bedros
Industrial Relations Director for Robotics, Sensing and Advanced Manufacturing
College of Science and Engineering
University of Minnesota
sbedros@umn.edu

Dedicated to my parents, Prof. T. P. Seshadri and Kanak Seshadri.

Without their support and encouragement, none of this would have been possible.

Abstract

The automatic localization of facial landmarks, also referred to as facial landmarking or facial alignment, is a key pre-processing step that is of vital importance to the carrying out of tasks such as facial recognition, the generation of 3D facial models, expression analysis, gender and ethnicity classification, age estimation, segmentation of facial features, accurate head pose estimation, and a variety of other facial analytic tasks. Progress in all these areas of research has heightened the need for developing accurate facial alignment algorithms that can generalize well to handle simultaneous variations in pose, illumination, expression, and high levels of facial occlusion in real-world images.

This thesis proposes a facial alignment algorithm that is not only tolerant to the joint presence of facial occlusions, pose variation, and varying expressions, but also provides feedback (misalignment/occlusion labels for the detected landmarks) that could be of use to subsequent stages in a facial analysis pipeline. Our approach proceeds from sparse to dense landmarking steps using a set of pose and expression specific models trained to best account for the variations in facial shape and texture manifested in real-world images. We also propose the use of a novel shape regularization approach that sets up this task as an ℓ_1-regularized least squares problem. This avoids the generation of implausible facial shapes and results in higher landmark localization accuracies than those obtained using prior shape models. Our approach is thoroughly evaluated on many challenging real-world datasets and demonstrates higher landmark localization accuracies and more graceful degradation than several state-of-the-art methods. We proceed to put the task of facial alignment into better context by examining its role in two applications that require alignment results as input: (1) a large-scale facial recognition scenario and (2) a project aimed at improving driver safety by assessing facial cues. Finally, we also carry out a rigorous set of experiments to analyze the performance of our approach when dealing with low-resolution images and provide some insights gained from this study.

Acknowledgements

The research in this thesis was supported in part by CyLab at Carnegie Mellon University (CMU) under grants DAAD19-02-1-0389 and W911NF-09-1-0273 from the Army Research Office and by cooperative agreement W911NF-10-2-0028 from the Army Research Laboratory and also by the Federal Highway Administration's Exploratory Advanced Research Program under contract DTFH61-14-C-00006. We would also like to thank Pittsburgh Pattern Recognition for providing their software library to Carnegie Mellon University for research purposes.

No one climbs over walls without someone providing them a leg up. A lot of people helped me over my metaphorical walls and I'm making an effort here to thank all of them. For those of you who hate my constant habit of turning everything I see around me into some kind of literary or pop culture reference, I apologize in advance for the numerous ones scattered through these acknowledgements but you know me well enough to know that, as Adrian Monk would say, "It's a gift and a curse."

I'd like to start by thanking the two people to whom I owe infinitely more than a Ph.D., my parents, Prof. T. P. Seshadri and Kanak Seshadri. My parents always put my needs ahead of their own and I owe them more than I can ever express using a few sentences, though that's what I'm attempting to do here They made everything possible and backed me every step of the way. When I felt things weren't going well, which was almost always, they were always there to provide their own unique brand of advice to make sure I got back on the horse and kept going. They were always there like a *Bridge over Troubled Water*, always trying to do more, when they had already done more than they needed to, and of course, I'm not just referring to my Ph.D. stint (stint, being a euphemism in my case) here. Their inordinate patience in the face of my neurotic negativity is unfathomable, but is something I will always be grateful for. For a lot of the same reasons, Vanaja patti, thank you for all the support and the advice and for always believing in me.

My advisor, Prof. Marios Savvides, has my heartfelt gratitude for taking a chance on me around the time I was winding up my master's degree. I was always sure I wanted to pursue a

Ph.D., but that was as clear as my plan got and and it was only due to some chance conversations with Shreyas Venugopalan, circa 2008, followed by some meetings with Marios, that the plan took on a more solid form. The assurance that he provided me at that stage and the confirmation of my admission to the Ph.D. program that followed were huge moments for me. There were numerous other high points along the course of my Ph.D., a completely unexpected trip (thanks Prof., that was a cracker of a trip and a great learning experience) to Oahu, Hawaii in November, 2008 for the Robust-2008 conference that he organized just a few months after I started work at what is now the CyLab Biometrics Center, the first of many road trips when I presented my first paper at the Biometrics: Theory, Applications and Systems 2009 (BTAS-2009) conference in Washington D.C., the text message that confirmed that I had passed the dreaded qualifier exam in November, 2010, subsequent conference trips, among which a second trip to Hawaii in 2011 and one to Orlando in 2012 stand out, and the fortuitous trip to Tampa for the Biometrics Consortium Conference (BCC) in September, 2012 that helped me avoid finding out how well I would manage when posed with the problem of being stuck inside a house on fire, are just some of the top moments on the list. It's an extremely difficult job to be a professor at a university and it's something I have a new found appreciation for after seeing the number of things that Marios has to juggle on a daily basis and the many hats he wears. His passion for the job and boundless optimism are qualities that have no doubt helped in creating the great environment that is the CyLab Biometrics Center.

I've never been much of a multitasker, something Marios no doubt realized, and so it suited me perfectly to keep chipping away, under his guidance, at a problem he had identified early on in my Ph.D. stint (let's stick with this euphemism, for the sake of consistency in notation), until the referees (Prof. Savvides and my doctoral committee) felt that it would be all right to call time on the fight. I still wish I could have won the fight with a clean upper cut to the jaw and a clear knockout, but at this point, I'm quite happy to limp away from the ring with a small edge on points with the consolation that if every Ph.D. student aimed for that perfect knockout, very few would ever graduate. While my time at this ring is at an end, I know that all the current and future students of Prof. Savvides' will continue along with him, in the words of Tennyson, "To strive, to seek, to find, and not to yield."

It's hard to examine your own work in a critical and objective fashion and I really struggled with this aspect just around the time I was due to submit my Ph.D. prospectus. I am very grateful to Prof. Vijayakumar Bhagavatula, Prof. Arun Ross, and Dr. Saad Bedros, members of my doctoral committee, for taking the time to do just that. Their feedback, along with that of Prof. Savvides', was of great value and helped me understand what I needed to focus on. Looking back at my Ph.D. prospectus presentation, I realize how important that step was in planning out the home stretch that culminates in a thesis defense. Dr. Craig Thor at the Office of Safety Research and Development, Federal Highway Administration (FHWA) and Dr. Paul Karnowski at Oak Ridge National Laboratory (ORNL), also have my thanks for their help and support at the last stage of my Ph.D. when I was involved with a project sponsored by the Federal Highway Administration. I'd also like to thank Prof. Richard Stern, Prof. Kumar, and Dr. John Dolan for giving me the benefit of the doubt on my less than impressive qualifier exam performance. Their spot on assessment and feedback was greatly appreciated.

One of the first things that struck me when I started at Carnegie Mellon was how friendly and helpful all the administrative staff were, it's been a pleasure interacting with all the folks associated with CyLab, the Electrical and Computer Engineering (ECE) Department, and the Office of International Education (OIE). Thank you Michael Balderson, Kelley Conley, Megan Kearns, Samantha Stevick, Rachel Swetnam, and Tina Yankovich; Elaine Lawrence, Tara Moe, Samantha Goldstein, and Nathan Snizaski; and Jennifer McNabb and Neslihan Ozdoganlar, for all your help over the years with various matters such as making travel arrangements and processing of business reimbursements, keeping track of my progress, or lack of it, towards satisfying all degree requirements, and the issuing of all those I-20s, respectively.

At this point, I'd also like to thank Dr. Leonardo Baloa and his team at Respironics Inc. (now Philips Respironics) for giving me my first real job and guiding me through it. That internship back in the summer of 2008 was, in hindsight, crucial and helped me in developing some up good coding practices that could have proved useful in my Ph.D. but for the fact that I very quickly unlearnt them as soon as I started.

Research is a collaborative process and I've been very fortunate to be able to brainstorm and collaborate with several not only dedicated and skilled researchers, but extremely helpful people who were always there within a nerf gun's range to help out. Ramzi Abi Antoun, Vishnu Naresh Boddeti, Sasikanth Bendapudi, Chandrasekhar Bhagavatula, Khalid Harun, Jingu Heo, Aaron Jaeach, Abhinandan Krishnan, Raied Jadaany, Thi Hoang Ngan Le (Nancy), Yung-hui Li, Khoa Luu, Tahei Munemoto, Dipan Pal, Sung Won Park, Kavya Patil, Utsav Prabhu, Karanhaar Singh, Nick Vandal, Shreyas Venugopalan, and Juefei Xu (Felix) all took the time to help me out along the way (often with some pretty inane stuff that they really needn't have bothered with), and I am extremely grateful to them for all the ideas, advice, and support, over the years. Ramzi, Jingu, Tahei, Sung Won, and Yung-hui all helped me get my bearings when I first started as a Ph.D. student and as the years started rolling by, and boy did they roll by, Ramzi, Jingu, Khalid, Utsav, Shreyas, Nancy, Khoa, Sekhar, and Karanhaar were all there for me, often going above and beyond the call of duty to help out. Khalid and Karanhaar have my thanks for converting research code versions of facial alignment algorithms into deliverable code and without their work it may have taken me an extra year, maybe more, to graduate. Khoa and Nancy have been great collaborators over the years, and Sophia, you have it made with wonderful parents. I'd like to also thank Vishnu for some extremely useful research related advice that he provided to me at key points just before my qualifier exam and proposal. Divya Sharma (Subtle, that's pronounced with a non-silent b), who though not a part of the Biometrics Center, was a frequent visitor to it, has my thanks for taking the time to share her unique views on *Life, the Universe and Everything* with me (and with anyone within earshot).

Shreyas and Utsav, who started their Ph.D. journey in the same semester as I did, have helped me out in more ways than I can enumerate. In particular, Utsav has my thanks for providing code to successfully run some of the experiments in this thesis, and both Shreyas and Utsav have my thanks for their advice (research related and otherwise) and help during the final stages of my Ph.D., a period when my powers of procrastination were at their peak, and for their extreme patience during a phase (euphemism, again) when I was trying to bag my white whale, a driver's license, a big budget and epic production that also involved the collective efforts of Arvind Kandhalu, who has my eternal gratitude for spending a lot of his time with me on that quest and for trusting me with his car for several weeks, Ronald Meeder, Apurva Samudra (Dada), Sudarshan Narayanan, and Harini Venugopal, whose car was the one in which my quest finally ended.

That brings me to thanking a whole bunch of old and new friends. I guess everyone has different sets of friends from various stages of their lives and while they may not all ever meet, an acknowledgement is one place where they all belong. My old National Public School (NPS) friends, Ketan, Rahul, and Soumyajit, have always stayed in touch (may we never forget to wish each other on our birthdays) and I am the better for it. And so moving on to my undergraduate years and to set of pals, who as a close and tightly knit team, managed to do something that in a small way left some "footprints in the sands of time." Ronojoy, Vinay, Amar, and I set up Quiz Pro. Co. (QPC), the still existent (I think) quiz club at Visvesvaraya National Institute of Technology (VNIT) Nagpur. My quality of life has always been proportional to the number of good films and TV shows I'm watching, the number of matches that the Indian cricket team is winning, and the number of quizzes I'm participating in or organizing. Here's hoping for more reunions with impromptu quizzes, old chaps. That brings me to another group of close friends from my undergraduate years. Anand (Andy), Anirudh, Gauri, Kandarp (Kandu), Kritika, Soumya (Jain Saab), Vinay G., and Vikram (Pops), thank you for all the great stories and memories over the long time that we have known each other, for the numerous discussions that sometimes led to something productive, for the memorable trips across the continental United States and Canada (Niagara Falls from the Canadian side counts), for dragging me out of my comfort zone, and for showing me how to live in the present (something I still need some help with).

My years at Carnegie Mellon have had their ups and downs, but through it all I've been extremely lucky to be surrounded by the smartest and most helpful people I've had the pleasure of knowing, whom I'm also extremely fortunate to call my friends. There's again some overlapping circles here, but broadly there's the very understanding roommates who put up with my stubborn refusal to make even the most basic dishes - Ankit (thanks for the tutorials on playing pool, among other things), Apurva (that Diwali quiz idea of yours was pure gold and I'm glad you roped me in to help with what turned out to be an event of great importance), Vikram, Suyash, and Supreeth (whom I also owe big time for fishing my passport and other crucial documents out of a fire damaged and water soaked apartment); the friends who had me over for dinner and potlucks several times in spite of my stubborn refusal to make even the most basic dishes - Ashwin, Harsh, Kavya, Lavanya, Khoa and Nancy, Samrat, Sarah and Shreyas, Sneha and Salil, and Veda; the aforementioned folks who are/were a part of the Biometrics Center; and then there's the friends I got to know so well, directly or indirectly, through that most glorious of entities - the quiz club. Words from the theme song to *Cheers* are particularly apt here:

"Making your way in the world today takes everything you've got.
Taking a break from all your worries, sure would help a lot.
Wouldn't you like to get away?
Sometimes you want to go
Where everybody knows your name,
and they're always glad you came.
You wanna be where you can see,
our troubles are all the same
You wanna be where everybody knows your name."

The quiz club is my (and I know I'm not the only one) *Cheers*. Abhijeet, Ajit, Apurva, Aranya

(Bloops), Aravindh (FNU), Arvind Muralimohan (James), Arvind Suresh, Ashwati, Avinash, Ayan, Bhargav, Chitra, Erle (The Transporter 2), Gaurav, Harini, Harsha (Chuck), Janani, Lavanya, Mike, Murali, Namratha, Nishtha, Pranav, Pranjal (PJ), Raghav (Ji), Ranjini, Rohan, Ravi, Riteja, Samrat (Sammy, The Transporter), Sandeep, Sandhya, Satyajith, Saurabh, Sekhar, Shishir, Siddharth (Garg), Siddharth (Maddali), Srinivas, Sudarshan (my dessert buddy at restaurants), Supreeth (BT), Suyash, Swapnil (Tommy), Uday (Pagan), Utsav, Varun (Navman), Varun (2.0, Toupie), Vas, Venkat, Vinod, Vivek (Sesha), Vyas, and everyone who ever participated in or made a quiz, thank you for the questions, the great psych outs, and the guaranteed weekly excitement. The quiz club has made all the difference to me and while I'll greatly miss being physically present for the quizzes, it'll be great to still participate remotely (huzzah for still being on the east coast), because it's going to be a long time before that dedicated timer in my head stops going off around 6:30 PM on Fridays. It won't be the same though and I'll miss the post-quiz and birthday celebration dinners, the raucous youth hostel "gatherings", the round table at Silky's, the board games, the more than spirited Mafia games (I'm never playing it again with you guys), the epic road and cabin trips, the viewing of several terrible movies on opening night (thanks Swapnil), and so much more that all came to be because a bunch of people who all shared a love for making and answering convoluted questions came to know each other so well. Thanks for everything comrades, the ride's been much less bumpy thanks to all of you. I'm looking forward to more "Dino trips" and to meeting up regularly in the future.

As I wind up here, I'd like to mention that all the goodwill and cheer provided by many close relatives and family friends. Cheena periyappa and Kausalya periyamma, thank you for all the help over the years and for the perspective that only people with a lifelong association with universities and schools can acquire. Sharada aunty, thank you for all the advice around the time I applied to various universities. Madhav mama, Anita mami, Govind, Rohit, and Sriram, Anand mama and Suhasini mami, Raghav mama and Suman mami, Gopal chechi and Sheela chithi, Bharadwaj and Trishi, and Sharada aunty, thank you for showing me around various parts of the U.S. and putting me up when I was away from Pittsburgh.

And finally, here's my thanks to bunch of people I've never met (though I hope to in the future), but whose work has given me great joy over the years and made a huge difference, although it may seem strange to admit it. First, there's the men's Indian national cricket team. Thanks for winning the 2011 world cup guys, and for putting up a pretty good show in the 2015 one as well. Those performances, along with several great test wins, are highs I'll always remember from my Ph.D. years. Special knocks by Sehwag, Laxman, Dhoni, Kohli, and, of course, Dravid and Tendulkar, have made a lot of difference and have often turned otherwise ordinary days into special ones. I'd like to thank Christopher Nolan, Quentin Tarantino, Wes Anderson, and David Fincher for generally doing their thing and ensuring that there was a much hyped movie that would release every year or so. I'm quite confident that if I live to an age when I can't remember what this thesis is about, I'll still remember key lines from *The Dark Knight Trilogy* to go along with ones from *The Shawshank Redemption* and other random movies and books and seemingly useless trivia. In similar fashion, I'd like to thank everyone associated with the following TV shows (some of which started airing only after my Ph.D. commenced, while the others were older and made available for streaming through Netflix - what a glorious thing Netflix is): *Cheers*, *Agatha Christie's Poirot*, *Seinfeld*, *Frasier*, *Friends*, *The West Wing*, *Scrubs*, *24*, *Monk*, *House*, *The Office*, *Prison Break*, *Burn Notice*, *The Big Bang Theory*, *Breaking Bad*, *Modern Family*, *Sherlock*, *Game of Thrones*, *Veep*, *House of Cards*, and *Brooklyn Nine-Nine*. They helped with getting through my Ph.D.; they helped a lot.

At the 76^{th} Annual Academy Awards ceremony in 2004, which was swept by *The Lord of the Rings: The Return of the King*, Billy Crystal, the host, mentioned that almost everyone in New Zealand had been thanked. I guess I was going for something similar, just like Ramzi did in his dissertation. So, I hope I've not forgotten anyone in these acknowledgments, and if I have, you have my apologies. I think that about covers it, and in the words of Forrest Gump, "That's all I have to say about that."

Contents

List of Figures

List of Tables

Chapter 1

Introduction

The White Rabbit put on his spectacles. "Where shall I begin, please your Majesty?" he asked. "Begin at the beginning," the King said gravely, "and go on till you come to the end: then stop."
Alice's Adventures in Wonderland by Lewis Carroll

The automatic localization of facial landmarks, also referred to as facial landmarking or facial alignment, is an active area of research whose importance has grown dramatically over the last few years. Facial alignment is a key pre-processing step required in order to carry out face recognition [18], [21], [22], [23], [24], [25], [26], [27], [28], [29], [30], the construction of 3D facial models [23], [31], [32], [33], [34], [35], expression analysis [36], [37], [38], [39], gender and ethnicity classification [40], [41], age estimation [42], [43], segmentation of facial features [44], [45], [46], head pose estimation [47], [48], [49], [50], and a variety of other facial analytic tasks. All of these tasks require precise knowledge of the locations of facial landmarks to extract regions of interest for classification or regression or for initialization purposes. In an ideal scenario, the role of facial alignment should be almost invisible and taken for granted. However, this is not the case at the present time and even many state-of-the-art algorithms struggle to provide an acceptable level of performance on real-world images that are now the focus of attention of all the previously men-

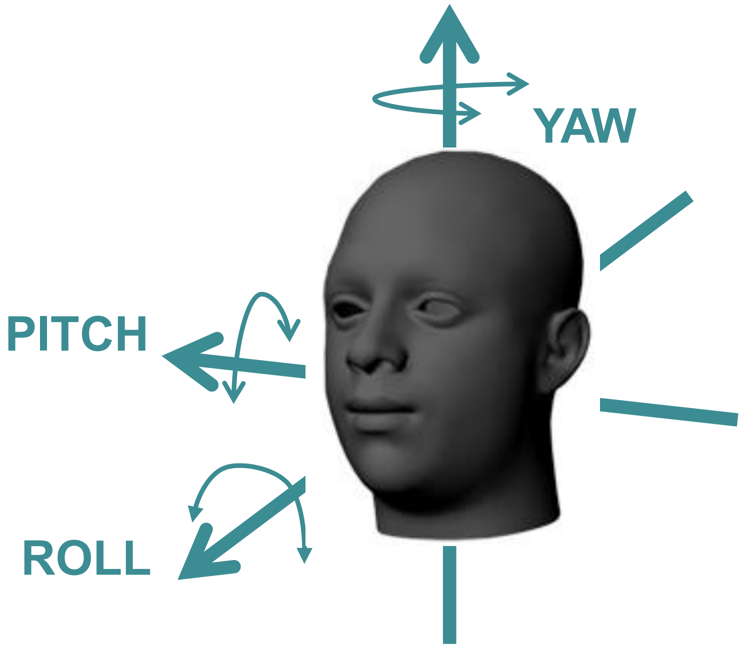


Figure 1.1: Head pose coordinate system showing what yaw, pitch, and roll refer to.

tioned research problems. With the strides made in all of these areas over the past few years, there has been a dramatic increase in research efforts aimed at developing facial alignment algorithms that can generalize well to handle variations in pose, illumination, expression, and large levels of facial occlusion in unseen test images. Many existing algorithms do not handle all of these variations when they are simultaneously present and it is this challenge that our work aims at addressing. In addition, we also study the problem of dealing with low-resolution images that also exhibit these variations.

1.1 Desirable Attributes in a Facial Alignment Algorithm

Facial landmark localization is never carried out in isolation. The results produced are usually utilized for an additional purpose, such as the construction of 3D facial models, facial recognition, or expression analysis. Thus, a question that arises is what list of desirable attributes or features must it possess in order to be useful in each of these scenarios. We believe the following list of attributes is a superset of the most important ones that a facial alignment algorithm ought to have in order for its output to be effectively used in a variety of applications.

(i) Ability to localize a dense set of landmarks: While some applications require only a sparse set of landmarks to simply normalize out the effects of scale, translation, and rotation in order to better align a face, many applications benefit from an alignment algorithm that is able to localize a dense set of landmarks. Additionally, another desirable trait would be flexibility during the training and testing stages to allow an algorithm to adapt to any landmarking scheme and output a different number of landmarks, based on the manual annotations available for training.

(ii) Ability to handle a wide range of pose variation: Figure 1.1 shows the head pose coordinate system and illustrates what the terms pitch, yaw, and roll refer to. Real-world images exhibit a wide range of yaw (turning of the head with negative yaw for cases when the subject looks to his/her right and positive yaw for cases when the subject looks to his/her left) and roll (in-plane rotation of the head) variation. Thus, it is important that a facial alignment algorithm be adaptable enough to automatically determine a suitable set of landmarks to output based on the yaw of the face, with $-90°$ to $+90°$ being the general range of interest. Many algorithms do not yet handle an absolute yaw in excess of $60°$ and are thus presently incapable of automatically landmarking profile faces. It is to be noted that while pitch variation may also be present in real-world images, its range is more constrained than those of yaw and roll and excessive pitch variation would result in cases where the face is not visible, thus making the facial alignment results produced unsuitable for further processing. Thus, in our work we focus on combating roll, and more particularly, yaw variation. Thus, when describing our facial alignment algorithm in chapter 3 the word pose refers specifically to the yaw (turning of the head), since it is this yaw variation that we explicitly account for using different models. While it is possible for our approach to also deal with small variations in pitch (nodding of the head), we do not explicitly train pitch models, and hence, in this context, the use of the word “pose” and the phrase “pose-specific” are meant to be synonymous with ”yaw” and “yaw-specific”, respectively.

(iii) Ability to handle the presence of simultaneous variations or degradations (suitable for dealing with unconstrained real-world images): The challenges that an effective facial alignment algorithm must overcome are the same ones that affect the performance of any facial analysis based algorithm, such as recognition engines, 3D modeling techniques, *etc.*. These challenges include variations in pose, expression, and illumination, the presence of facial occlusions, such as sunglasses, scarves, hair, hands, food, *etc.*, and the presence of low-resolution artifacts. Real-world images are unlikely to exhibit only one of these variations or degradations and they generally occur simultaneously. Thus, it is a key requirement for a facial alignment algorithm to be equipped to handle these variation and degradations irrespective of whether they occur individually or jointly and without making any assumptions or requiring prior information regarding their presence.

(iv) Performance feedback capability: This is a very useful attribute for an alignment algorithm to possess, again bearing in mind that that alignment is generally followed by a stage that uses the outputs produced by it. If an alignment algorithm can not only localize landmarks but also provide confidence scores that correlate to their goodness of fit, this could prove very useful for subsequent stages in an application pipeline. For example, in a face recognition scenario, prior knowledge of occluded regions on the face could prove very useful as such regions could now be treated as untrustworthy and either reconstructed using a massive training dictionary before carrying out recognition [18], [28], [29], or omitted from the matching process with smaller regions, such as the periocular (eyes and eyebrows) region, of the face used instead, as demonstrated in [51], [52], [53], [54]. Similarly, when tracking facial landmarks across the frames of a video sequence, knowledge of the goodness of fit of landmarks on previous frames could allow for higher accuracies and appropriate initialization on future frames and the easy determination of whether a subject of interest has been lost entirely. Confidence scores for individual landmarks or a general goodness of fit index for all of them considered together is also be a must for a situation in which face detection and facial landmark localization are carried out together (rather than the latter following the former) in a single step, as is the case in [6] and [16].

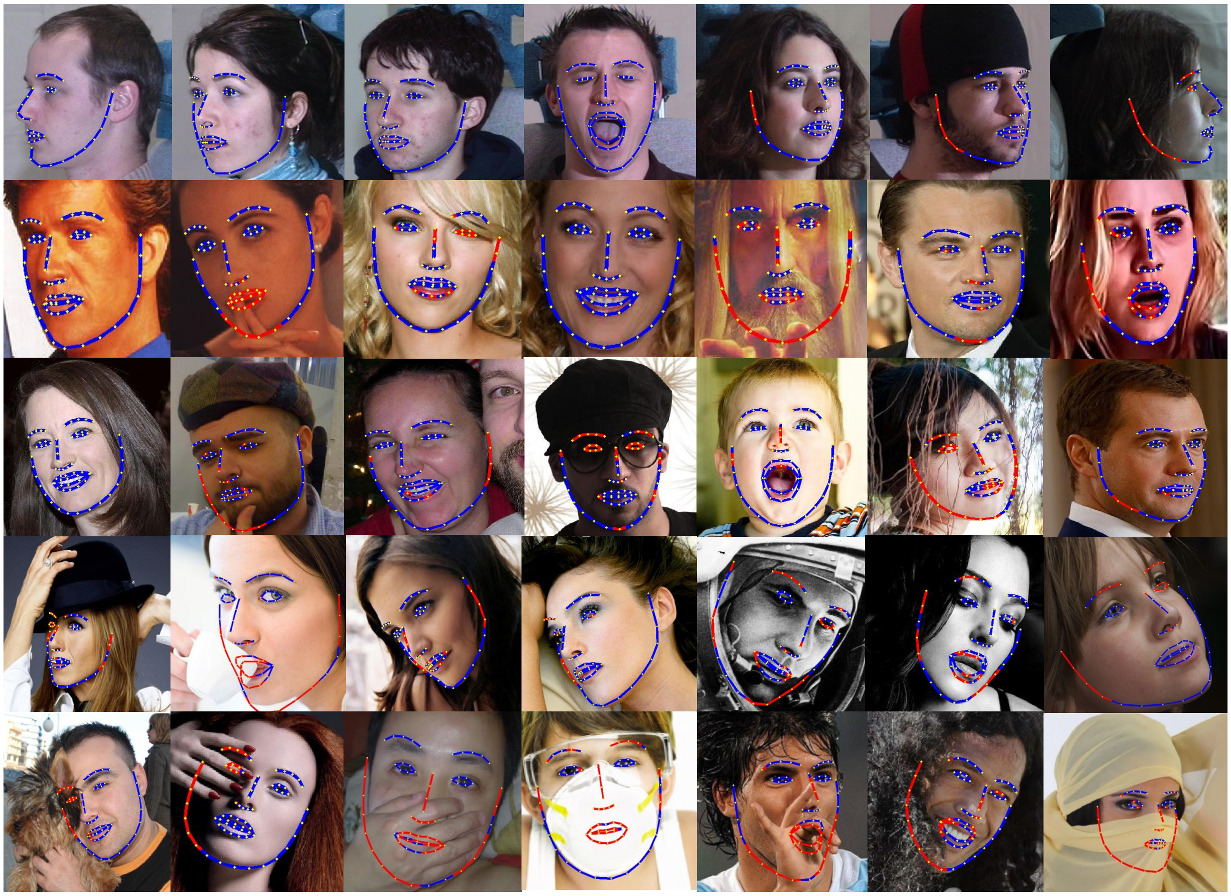


Figure 1.2: Qualitative landmark localization results produced by our approach on some images from the CMU Multi-PIE (MPIE) [1], [2], [3], Labeled Face Parts in the Wild (LFPW) [4], [5], Annotated Faces in-the-Wild (AFW) [6], [7], ibug [8], [9], [10], and Caltech Occluded Faces in the Wild (COFW) [11], [12] datasets from the top to bottom rows, respectively. In all facial images with landmarks overlaid on them, yellow dots are used to indicate the locations of facial landmarks, blue line segments indicate that the landmark at their center is accurately localized, and red line segments indicate that the landmark at their center is misaligned or potentially occluded (goodness of fit feedback). The results serve to demonstrate the pose, occlusion, and expression tolerance of our approach on challenging real-world images that are quite dissimilar to the images it was trained on (from the MPIE dataset) as well as its ability to provide performance feedback. This figure appears in [13].

(v) Flexibility and scope for modification: This is an optional attribute, but one that could prove useful as well. If a facial alignment algorithm could be easily modified to use different feature extraction techniques, classifiers or regressors, and a varying number/configuration of landmarks, this could allow for greater ease of use in a variety of circumstances with appropriate design choices made to better suit the application. This kind of flexibility could also allow the algorithm to not only localize landmarks on faces, but on other rigid objects as well, such as cars, bicycles, tables, chairs, *etc.*, using minimal changes. We demonstrate this adaptability feature of our algorithm in section 7.1.8, in which we apply it to localize a set of landmarks that lie along the contours of images of cars at various viewpoints.

The facial alignment approach we describe in this thesis was designed by keeping in mind all of the previously described attributes. Figure 1.2, that shows some landmark localization results produced using our approach on a variety of images ranging from constrained database images to more unconstrained real-world ones, serves to illustrate this. As we will demonstrate in future chapters, our approach also lends itself to easy modification and use in various applications, such as face recognition, video based landmark localization, and head pose estimation.

1.2 Contributions of this Thesis

We present an approach to facial landmark localization that is not only robust to all the previously mentioned real-world challenges (pose, illumination, expression, and occlusions), but more importantly, can handle all these **real-world variations/degradations**, even when they **occur simultaneously**. Since facial shape and the local texture around the landmarks that constitute them vary dramatically with facial pose and expressions, it is beneficial to build not one, but multiple models that can best account for these variations. In order to combat the problem of facial occlusions, we account for them using landmark and pose-specific local texture based classifiers that are trained to discriminate between the texture around a correctly localized landmark and an incorrectly localized one. Thus, we factor in the presence of occlusions without actually requiring training images with manual annotations denoting the presence of occlusions, as required by some recent approaches in the area [11]. Many existing facial alignment algorithms also rely heavily on consistent facial detection results, something that is seldom guaranteed when dealing with real-world images data as facial bounding box results produced by the same detector vary in size and location even for a similar set of images and do not always account for in-plane rotation (roll) of the face, as we learned from our initial work on using a Modified Active Shape Model (MASM) for landmark localization in frontal faces, detailed in the following papers:

[55] Keshav Seshadri and Marios Savvides, "Robust Modified Active Shape Model for Automatic Facial Landmark Annotation of Frontal Faces," *IEEE* 3rd *International Conference on Biometrics: Theory, Applications and Systems (BTAS)*, Sep. 2009.

[56] Keshav Seshadri and Marios Savvides, "An Analysis of the Sensitivity of Active Shape Models to Initialization When Applied to Automatic Facial Landmarking," *IEEE Transactions on Information Forensics and Security (TIFS)*, Vol. 7, Issue 4, Aug. 2012.

In order to enable accurate initialization of our shape models, our approach proceeds in a stage wise fashion. In our first step, we use a sliding window approach to only localize a few key landmarks, such as the centers of the eyes, the corners of the mouth, and the tip of the nose, *etc.*, that we refer to as seed landmarks. It is important to note that we do not require all of these seed landmarks to be visible and only require that any combination of two of these landmarks be reliably localized. Our next step involves aligning a denser set of landmarks (a canonical mean facial shape specific to a particular yaw range that is obtained during our training stage) using just two of the seed landmark candidates at a time and evaluating the goodness of fit of this dense set of points. Thus, we are now able to generate the most accurately aligned initial shape for each of our pose-specific models and also account for the in-plane rotation of the face (roll) and the possible presence of occlusions. *Well begun is half done* in the field of facial alignment goes and this step goes a long way towards ensuring this. The final stage involves the refinement of the top ranked shapes and the selection of a single set of final landmarks that best model the shape and texture of a given face. This stage involves the use of both a local texture guided search coupled with a shape regularization stage in order to guide the search for the optimal locations for all the landmarks. Our approach uses a shape dictionary built from the manually annotated ground truth training shapes coupled with an ℓ_1-regularized least squares approach in order to perform shape regularization. The assumption here is that the varying shapes present in the training set contain sufficient information to represent and regularize the shape produced by landmarks that lie along the contours an unseen face, since human faces can be treated as rigid objects that deform in similar fashion, for a specific pose. However, this assumption does not hold when expressions, especially those that result in excessive movement of the lower jaw, such as surprises or screams, are manifested. For this reason, we construct separate shape dictionaries (and local texture models) to better model such open mouth expressions. Our shape regularization is carried out by using these shape dictionaries and by constraining the shape coefficients (that are used to represent a new shape as a linear combination of the shapes in the dictionary) using ℓ_1-regularization. This ensures that an appropriate weight is placed on each shape in the dictionary when representing a new shape and also results in higher landmark localization accuracies than those obtained using a widely used linear subspace based shape model [57], [58], as we demonstrate in chapter 3. In addition, by only using accurately localized landmarks (inliers) during the shape regularization stage, the effect of occlusions is negated as poorly localized landmarks (outliers) do not play a role in the shape regularization process and thus do not sway the results. The contributions described in this chapter are detailed in the following paper:

[13] Keshav Seshadri and Marios Savvides, "Towards a Unified Framework for Pose, Expression, and Occlusion Tolerant Automatic Facial Alignment," *To appear in IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, 2016.

Chapter 4 provides some context for our work by demonstrating the applicability of our facial alignment algorithm in a real-world face recognition experiment on the Labeled Faces in the Wild (LFW) [59], [60] database. While it is an acknowledged fact that poor alignment results can adversely affect the performance of many existing face recognition techniques [18], [28], [29], this chapter serves to demonstrate this point by using landmark localization results produced by different facial alignment algorithms as input to the same face recognition algorithm and assessing how the recognition performance is impacted. In chapter 5, we detail experiments and results that were obtained using our algorithm in a few other allied applications. Our approach is modified to enable the localization of facial landmarks across the frames of video sequences that were acquired as part of naturalistic driving study and then utilized for head pose estimation and the determination of whether the subjects in the videos were using a cell phone or not as part of our efforts to assist with a Federal Highway Administration (FHWA) [61] project aimed at understanding driver behavior in order to improve driver safety [62]. Initial findings from this work have been published in the following paper:

[20] Keshav Seshadri, Felix Juefei-Xu, Dipan K. Pal, Marios Savvides and Craig P. Thor, "Driver Cell Phone Usage Detection on Strategic Highway Research Program (SHRP2) Face View Videos," 6$^{\text{th}}$ International Workshop on Computer Vision in Vehicle Technology (CVVT) in conjunction with the 28th IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Jun. 2015.

Finally, in chapter 6, we add an extra dimension to our work by using a rigorous set of experiments to understand the challenge that low-resolution images pose to the alignment process. We test our approach using resolution-specific texture models under various conditions in order to understand how much simultaneous degradation (low-resolution images that also exhibit varying pose, expressions, and levels of facial occlusion) it can handle.

Our key contributions in this thesis can thus be summarized as follows:

- The development of a flexible framework (pipeline) for dense facial landmark localization algorithm that incorporates all of the previously mentioned desirable attributes to **jointly deal** with the problems posed by facial pose variation (range of yaw variation from $-90°$ to $+90°$), varying facial expressions, and partial occlusion of the face (chapter 3).
- The use of a novel method to constrain shape coefficients by setting up this task as an ℓ_1-regularized least squares problem. This avoids the generation of implausible facial shapes and results in higher landmark localization accuracies than those obtained using prior shape

models (chapter 3).

- A thorough evaluation and benchmarking of our approach against many state-of-the-art approaches on several challenging real-world datasets (chapter 3).
- The carrying out of a real-world face recognition experiment using a recognition algorithm as a black box and using varying inputs provided by different alignment algorithms to provide some context for the role played by facial alignment in this key task (chapter 4).
- The application of our alignment algorithm to a large-scale experiment on facial landmark localization in challenging naturalistic driving videos and the carrying out of allied tasks that build on this in order to assess driver behavior using facial cues, such as head pose estimation (chapter 5).
- The carrying out of a thorough set of experiments to understand the challenges posed by low-resolution images to the alignment process and to determine how much simultaneous degradation (low-resolution images that also exhibit varying pose, expressions, and levels of facial occlusion) our approach can handle (chapter 6).

Chapter 2

Background

"If I have seen further, it is by standing on the shoulders of giants."

Sir Isaac Newton (translated into modern English)

Facial landmark localization has been well studied over the past few years and a variety of different techniques have been proposed in order to deal with various aspects of the problem. A detailed survey of all these approaches was carried out by Celiktutan *et al.* in [63]. We provide a broad overview of some of these techniques and focus on a few state-of-the-art techniques against which we benchmark our approach in future chapters of this thesis.

2.1 Active Appearance Models (AAMs), Active Shape Models (ASMs), and Constrained Local Models (CLMs)

Traditionally facial landmarking has been carried using deformable template (parametric) based models, such as Active Appearance Models (AAMs) [57], [64] and Active Shape Models (ASMs) [58], [65], [66]. Both build shape models, also referred to as Point Distribution Models (PDMs), that model the shape of a typical face (represented by a set of constituent landmarks), and texture

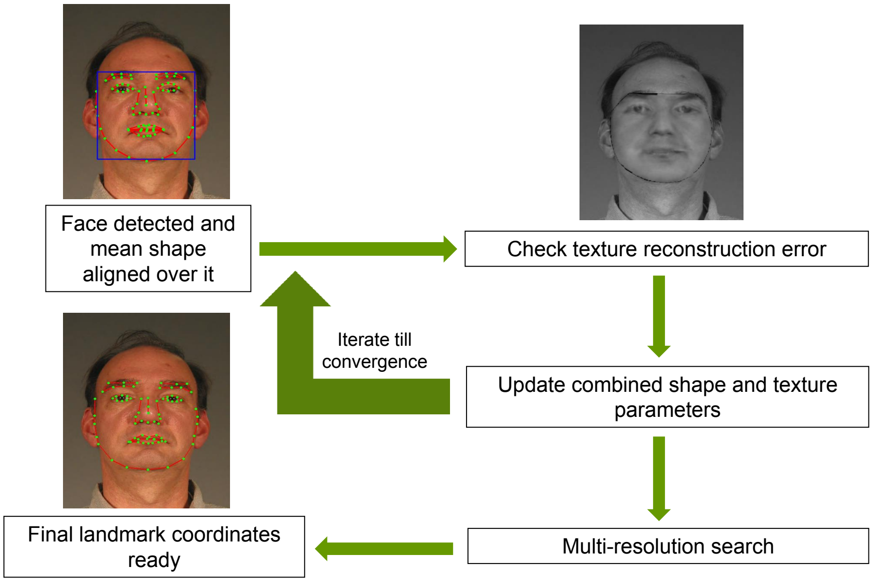

Face detected and mean shape aligned over it
Check texture reconstruction error
Iterate till convergence
Update combined shape and texture parameters
Multi-resolution search
Final landmark coordinates ready

Figure 2.1: The typical sequence of steps followed by a multi-resolution Active Appearance Model (AAM) for carrying out facial alignment. The facial image used to demonstrate the process is from the Multi Biometric Grand Challenge (MBGC) database [14], [15].

models of what the region enclosed by these landmarks looks like. The difference between the two is that ASMs build local texture models of what small 1D or 2D regions around each of landmarks look like, while AAMs build global texture models of the entire convex hull bounded by the landmarks. The AAM fitting process is governed by updates to a combined vector of shape and appearance (texture) parameters based on the difference between the underlying facial texture and the texture reconstructed using the parameters. Typically, a multi-resolution framework is also used to ensure higher fitting accuracies and faster convergence rates, as shown in Figure 2.1.

ASMs belong to a class of methods that can be broadly referred to as Constrained Local Models (CLMs) [67], [68], [69]. CLMs build local models of texture variation around landmarks (sometimes referred to as “local patch experts”) and allow landmarks to drift into the locations that best match training data using these patch experts. The shape is then regularized using the shape model to generate a final set of landmarks whose coordinates are in accordance with their typical locations for a human face. Again, a a multi-resolution framework is quite common. The typical steps followed by an ASM to carry out facial alignment are depicted in Figure 2.2. Several improve-

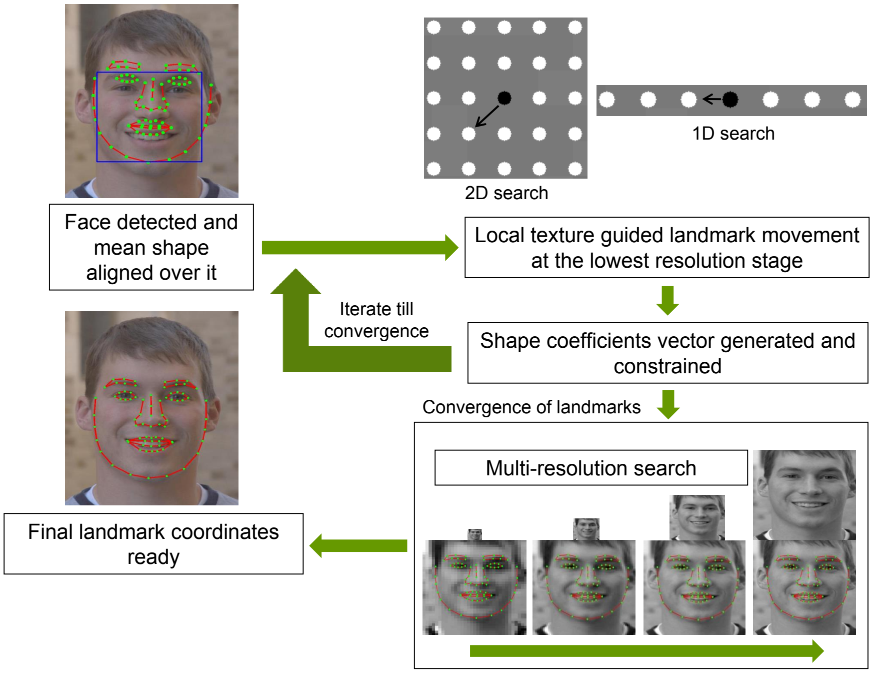

2D search
1D search
Face detected and mean shape aligned over it
Local texture guided landmark movement at the lowest resolution stage
Iterate till convergence
Shape coefficients vector generated and constrained
Convergence of landmarks
Multi-resolution search
Final landmark coordinates ready

Figure 2.2: The typical sequence of steps followed by a multi-resolution Active Shape Model (ASM) for carrying out facial alignment. The facial image used to demonstrate the process is from the Multi Biometric Grand Challenge (MBGC) database [14], [15].

ments have been made to ASMs over the years, such as those proposed in [55], [70], that have mainly focused on developing better local texture models. However, they still remain susceptible to occlusions, the problem of local-minima, and are very dependent on accurate initialization being provided, which is something we have previously investigated in [56]).

It has been demonstrated that ASMs are more suited to the task of precise facial landmarking than AAMs [56], [66], [71] as AAMs are generative, global texture based approaches and are more easily affected by variations in illumination, pose, and occlusions. AAMs also generalize poorly when dealing with unseen faces (faces outside their training set) compared to ASMs. However, there has been a lot of prior work on improving the objective function and update rules that AAMs use to deform the initial shape overlaying a face to better represent it [64], [72], [73], [74]. Recently, Tzimiropoulos and Pantic [75] proposed new optimizations for fast and accurate AAM fitting and demonstrated better fitting results on unseen images with a large range of pose variation using a more unconstrained training set drawn from the Labeled Face Parts in the Wild (LFPW) dataset [4], [5]. This is one of the approaches that we use in our evaluations in section 3.2.3, in which the landmark localization accuracies of various facial alignment algorithms are compared, and in section 4.2, in which the alignment results produced by the same algorithms are used in a face recognition experiment.

Though it is possible to build separate AAMs or ASMs to handle pose variation using view-based models, as carried out in [76] and [77], [78], and [79], respectively, the fact that they require very accurate initialization decreases their effectiveness, especially when dealing with real-world images. Thus, over the past few years, several efforts have been made to develop alternative shape regularization techniques to better cope with pose variation and partial occlusion of the face. Zhou *et al.* [80] proposed a Bayesian Tangent Shape Model (BTSM) to infer the shape parameters through a maximum a posteriori (MAP) estimation in a tangent space and obtained more accurate results than those obtained by the classic ASM algorithm. Gu and Kanade [81] proposed a shape regularization model that incorporated a nonlinear shape prior and the likelihood of multiple candidate landmarks in a three-layered generative model that demonstrated higher accuracy than BTSM on images from the AR face database [82], [83] and the Multi-PIE (MPIE) database [1], [2], [3]. Their method also demonstrated some tolerance to expression variations and occlusions in real-world images. However, both these approaches were not developed to deal with a wide range of yaw variation from $-90°$ to $+90°$.

2.2 State-of-the-art Work on Pose Tolerant Discriminative Facial Alignment

Over the last few years there has been a dramatic increase in work dealing with the automatic localization of landmarks in non-frontal faces. Everingham *et al.* [84] developed an algorithm that used a generative model of facial feature positions (modeled jointly using a mixture of Gaussian trees) and a discriminative model of feature appearance (modeled using a variant of AdaBoost and “Haar-like” image features [85]) to localize a set of 9 facial landmarks in videos with faces exhibiting slight pose variation. Dantone *et al.* [86] used conditional regression forests to learn the relations between facial image patches and the location of feature points conditioned on global facial pose. Their method also localized a sparse set of 10 landmarks in real-time and achieved accurate results when trained and tested on images from the Labeled Faces in the Wild (LFW) database [59], [60].

Belhumeur *et al.* [4] proposed a novel approach to localizing facial parts by combining the output of local detectors with a consensus of nonparametric global models for part locations, computed using training set exemplars in a Bayesian framework, that served as the surrogate for shape regularization. Their approach was able to localize a set of 29 facial landmarks on faces that exhibited a wider range of occlusion, pose, and expression variation than many previous approaches. Their work inspired other nonparametric exemplar based approaches, such as those proposed Zhou *et al.* [87] and Smith *et al.* [88]. Zhou *et al.* developed an Exemplar-based Graph Matching (EGM) approach to obtain the optimal landmark configuration by solving a graph matching problem using linear programming and improved on the localization accuracy that was obtained by Belhumeur *et al.* on the same set of 29 landmarks. Recently, Smith *et al.* proposed a data-driven approach that uses feature voting based landmark detection and nonparametric shape regularization to build an in-plane rotation, pose, expression, and occlusion tolerant facial alignment algorithm.

All of these approaches are capable of providing accurate fitting results on some challenging images but lack in a few areas (some of the desirable attributes in an alignment algorithm that we previously described in section 1.1) that our approach aims at addressing. With the exception of the nonparametric method proposed by Smith *et al.* , the other previously described approaches only localize a sparse set of landmarks which is unsuitable for many real-world applications, such as expression analysis or the building of 3D facial models, that require a slightly denser set of landmarks in order to establish point correspondences. Also, none of the approaches demonstrate the ability to localize landmarks in faces with absolute yaw angles greater than 60° and are thus incapable of automatically landmarking profile faces. Finally, even though a few of the previously mentioned approaches, such as that of Smith *et al.* , demonstrate some tolerance to facial occlusions, none of them provide a score or label that can be used to determine which landmarks are potentially misaligned or occluded.

The approach proposed by Roh *et al.* [89] is one that satisfies most of the criteria we feel are important for a facial alignment algorithm to be truly generalizable to any task. Their approach used local detectors to determine an initial set of plausible candidates for each facial point. However, in order to combat occlusion, a RANdom SAmple Consensus (RANSAC) [90] based hypothesize-and-test strategy was adopted to determine which set of landmarks to actually use, *i.e.*, which set of landmarks are potential inliers. Using the inliers, the full set of landmarks could be hallucinated and feedback can also be provided on the remaining landmarks that can be classified as outliers (potentially occluded). This approach was used to demonstrate a tolerance to fitting of purely frontal images with facial occlusions, however, a more general framework using pose-specific shape models could also be developed to handle a larger range of yaw variation.

Another approach proposed by Yu *et al.* [91] also satisfies many of the previously mentioned criteria and is similar to our approach in that positive (well aligned) and negative (occluded or misaligned) texture patches around landmarks are modeled during the training stage. Using a logistic regression framework to obtain shape coefficients, Yu *et al.* were able to demonstrate a tolerance to occlusions in frontal images drawn from the AR and LFPW datasets and could also predict the locations of the occlusions. However, their approach was only applied to purely frontal images and not benchmarked against recent state-of-the-art approaches on more challenging datasets containing non-frontal facial images.

In their recent seminal work, Zhu and Ramanan [6] proposed an elegant framework that built on the previously developed idea of using mixtures of Deformable Part Models (DPMs) for object detection [92] to simultaneously detect faces, localize a dense set of landmarks, and provide

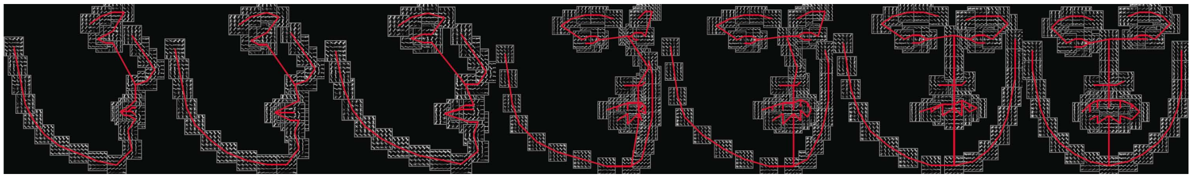
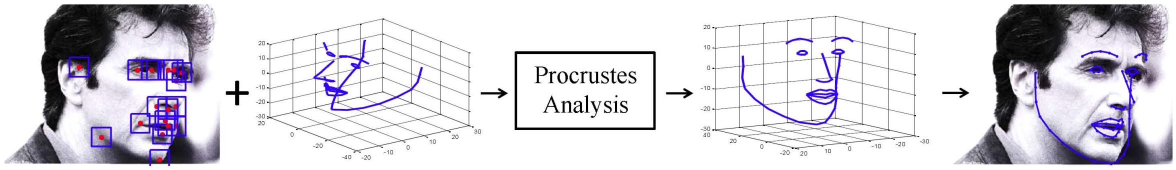


Figure 2.3: The mixture-of-trees model used in the work of Zhu and Ramanan [6]. Topological changes due to viewpoint are encoded by the different mixtures. The red lines denote the connections (springs) between the various parts (landmarks). Closed loops are not present in order to maintain the tree property. A common shared set (pool) of templates are utilized by the trees, thus making learning and inference quite efficient. This figure has been reproduced from [6].

a course estimate of facial pose (yaw) in challenging images. Their approach used a mixture of trees with a shared pool of parts to model sets of facial landmarks at various views (yaw angles), as depicted in Figure 2.3. Histogram of Oriented Gradients (HOG) [93] were used to model the local texture around each facial landmark, global mixtures were used to capture changes in facial shapes across pose (yaw), and the Tree-Structured Models (TSMs) were optimized quickly and effectively using dynamic programming, *i.e.*, inference was performed in an efficient manner to determine the best possible configuration of parts for each mixture which maximized a scoring function that took shape and appearance into account. The approach is quite effective and is tolerant to a range of yaw variation from -90° to $+90^\circ$, which is quite rarity in this field. However, it is not extremely accurate when dealing with occluded faces or faces that exhibit large in-plane rotation. Nevertheless, this groundbreaking facial alignment implementation is one against which all current pose-tolerant facial alignment algorithms are being compared and has inspired several other efforts aimed at pose and occlusion tolerant facial alignment, sometimes in a joint framework with face detection.

Yu *et al.* [16] built on the work of Zhu and Ramanan to automatically select a sparse set of salient landmarks to serve as initialization. They subsequently used a 3D facial model, mean-shift with CLMs, and face component-wise active contour models to produce a refined set of facial landmarks. An overview of their approach is provided in Figure 2.4. Recently, Ghiasi and Fowlkes [94] also built on the work of Zhu and Ramanan and proposed a hierarchical deformable part



Optimized Part Mixture 3D Reference Shape 3D Translated Shape Initialized Landmarks

Figure 2.4: An overview of the approach proposed by Yu *et al.* in [16]. This figure has been reproduced from [16].

model for face detection and landmark localization to explicitly model the occlusion of parts and hence achieved more accurate results on challenging occluded real-world images. However, their approach modeled three clusters and handled a pose variation of only $\pm 22.5°$. It is to be noted however that none of these three methods provide misalignment/occlusion labels for the fitted landmarks.

Asthana *et al.* [95] recently developed a discriminative regression based approach for the CLM framework that they referred to as Discriminative Response Map Fitting (DRMF). DRMF represents the response maps around landmarks using a small set of parameters and uses regression techniques to learn functions to obtain shape parameter updates from response maps. Their technique improves on the Regularized Landmark Mean-Shift (RLMS) approach in [69].

Xiong and De la Torre [96] recently formulated the Supervised Descent Method (SDM) and applied it to the task of detecting interior facial landmarks (excluding landmarks that lie along the facial boundary) to good effect. The SDM algorithm was formulated to minimize a Nonlinear Least Squares (NLS) function using descent directions learned from training data and without computing the Jacobian nor the Hessian. For the task of facial alignment, consider an image $\mathbf{d} \in \mathbb{R}^m$ consisting of m pixels with p facial landmarks and with $\mathbf{d}(\mathbf{x}) \in \mathbb{R}^p$ indexing these landmarks. Let $\mathbf{h}$ represent a nonlinear feature extraction technique or function such that $\mathbf{h}(\mathbf{d}(\mathbf{x})) \in \mathbb{R}^{np}$, where n is the dimensionality of the feature vector extracted around each facial landmark (128 dimensional Scale-Invariant Feature Transform (SIFT) [97] features are used in this case). If the initial configuration of facial landmarks (generally obtained using a mean shape) can be represented by $\mathbf{x}_0$, then the facial alignment problem is posed as the minimization of the function f, given by equation (2.1), over the variable $\Delta\mathbf{x}$.

$$f(\mathbf{x}_0 + \Delta\mathbf{x}) = \|\mathbf{h}(\mathbf{d}(\mathbf{x}_0 + \Delta\mathbf{x})) - \mathbf{\Phi}_*\|_2^2 \tag{2.1}$$

In equation (2.1), $\mathbf{\Phi}_* = \mathbf{h}(\mathbf{d}(\mathbf{x}_*))$ represents the features extracted from a manually labeled training image. $\mathbf{\Phi}_*$ and $\Delta\mathbf{x}$ are known for all training images and hence the goal of SDM is to use this information to learn a series of descent directions to produce a series of updates $\mathbf{x}_{k+1} = \mathbf{x}_k + \Delta\mathbf{x}_k$, starting from $\mathbf{x}_0$ and converging to $\mathbf{x}_*$, and then applying these update rules to minimize f when applied to a test image.

2.3 Regression Based Facial Alignment

Recently, a variety of approaches [11], [98], [99], [100], [101] that can be broadly grouped under the category of regression based approaches have emerged. In such approaches, an initial shape S^0 is aligned roughly with the face in an image and is progressively refined by estimating a shape increment ΔS using an iterative (stage-by-stage) framework. The shape increment at each iteration $t = 1, 2, \ldots, T$ is determined using a regression function (or a set of regressors) computed using the input image, the shape from the previous stage S^{t-1}, and the local texture features extracted from a region around each landmark.

Cao *et al.* [98] proposed a novel regression based approach in which a regressor is trained to explicitly minimize the alignment error over training data in a holistic manner, *i.e.*, all all facial landmarks are regressed jointly in a vectorial output. The shape constraints are encoded in non-parametric form by using the constraint that the regressed shape is always a linear combination of all the training set shapes. This approach provided accurate landmark localization results on the LFW, LFPW, and BioID [102], [103] databases, however, it is not extremely effective at dealing with partially occluded faces and faces that exhibit large shape variations.

To explicitly deal with occluded faces and provide feedback on which landmarks were occluded Burgos-Artizzu *et al.* [11] proposed the Robust Cascaded Pose Regression (RCPR) algorithm. They incorporated occlusion modeling explicitly into the training stage using facial images that were manually annotated to provide ground truth landmark coordinates as well as occlusion labels for each of the landmarks. RCPR was trained on images from the LFPW dataset and the newly annotated the Caltech Occluded Faces in the Wild (COFW) dataset [11], [12] in order to better equip it to deal with the problem of facial occlusion in real-world images. However, it is to be noted that the RCPR framework is extremely flexible and can be trained using any consistent landmarking scheme, *i.e.*, any set of images with the same number of landmarks, and can also be trained on images that do not have occlusion labels for the landmarks. RCPR uses shape-indexed features that are invariant to face scale and pose to enable robust shape estimation in real-world images. The features are referenced by linear interpolation between two landmarks, which results in improved shape fitting and faster computation. RCPR also incorporated a smart restart method to obtain higher landmark localization accuracies, compared to the explicit shape regression approach in [98], using different shape initializations. RCPR localized landmarks with high accuracy when trained and tested on similar images, *i.e.*, similar variations are manifested in the test images as in the training images. However, it does not generalize well to unseen variations, requires facial bounding boxes during the training stage to almost perfectly match those at the test stage, and can not automatically be used to annotate profile faces, that require a different set of landmarks to be localized compared to frontal faces.

To deal with the problem of sensitivity to initialization, Yan *et al.* [99] proposed a framework that generates multiple hypotheses (using a cascade regression based approach) by randomly shifting and re-scaling the bounding box provided by a face detector and then fuses these hypotheses to produce a final output. Recently, Ren *et al.* [100] proposed the use of computationally cheap local binary features and a linear regression framework to achieve fast and precise facial alignment on images from the LFPW, Helen [104], [105], and 300 Faces in-the-wild (300-W) challenge [8], [9] datasets.

Regression based approaches have become quite popular in the recent past due to their high fitting speed and the accuracies they achieve under the right conditions. However, many of these approaches are sensitive to initialization. It is also to be noted that none of these approaches can automatically (without a previous pose estimation step) deal with faces that exhibit absolute yaw angles in excess than $60°$.

2.4 Deep Learning Approaches to Facial Alignment

With the recent surge in the application of deep learning and the use of Convolutional Neural Networks (CNNs) for solving a variety of computer vision and machine learning problems, a few new approaches to facial alignment have also emerged. Such approaches have the advantage of using networks that are trained to localize all landmarks simultaneously thus implicitly modeling the geometric constraints between them without the need for an explicit shape model.

Sun *et al.* [106] used three-level cascaded convolutional networks where at each level, the outputs of multiple networks are fused for accurate localization of 5 facial landmarks. Zhou *et al.* [107] developed a a four-level course-to-fine convolutional network cascade in which each network level is trained to refine a subset of facial landmarks generated by previous network levels. Recently Zhang *et al.* [108] formulated a tasks-constrained deep model to optimize facial landmark detection along with correlated tasks such as head pose estimation, gender classification, *etc.*.

All of these methods provide highly accurate landmark localization results on widely varying real-world images. However, they too have focused on dealing with facial images that exhibit yaw variation only in the range between $-45°$ and $+45°$. Additionally, the fact that these methods require a large amount of training data has restricted them slightly. At the present time there are no large-scale datasets available with manual annotations for a dense set of landmarks and a wide variety of faces exhibiting different poses and expressions. Thus, many of the current deep learning based approaches tend to localize a sparse set of landmarks for which manual annotations are available on a large corpus of images. However, this is likely to change in the near future and such approaches are likely to be researched and used extensively, as discussed in section 7.1.4.

2.5 Synopsis

Many of the previously described facial alignment algorithms, such as those proposed in [84], [86], and [106] only localize a sparse set of landmarks which is unsuitable for many applications, such as expression analysis or the building of 3D facial models. Most of the approaches, with the exception of [6] and [16], are not equipped to automatically localize landmarks in faces that exhibit absolute yaw angles greater than $60°$ and are thus presently incapable of automatically localizing landmarks in such profile faces. Finally, only a few of the approaches, such as [11], provide feedback in the form of occlusion labels for the detected landmarks. Thus, there is still a need for a facial alignment algorithm that can provide a dense set of facial landmarks, deal with the full range of yaw variation from $-90°$ to $+90°$ and facial expressions, handle partial occlusion of the face, and provide misalignment/occlusion labels. It is our intention to draw attention to the all of these desirable attributes in a facial alignment algorithm, provide details on our own approach to facial alignment that is able to incorporate them, and finally, demonstrate the effectiveness of our approach over some of the more widely used state-of-the-art algorithms on challenging real-world datasets. In addition to this, our facial alignment algorithm is also suitably modified to deal with facial landmark tracking in video sequences, in chapter 5, and landmark localization on low-resolution images, in chapter 6. An overview of prior work in these fields is provided in these chapters for the sake of convenience.

Chapter 3

Our Approach to Facial Alignment

"How often have I said to you that when you have eliminated the impossible, whatever remains, however improbable, must be the truth?"
Sherlock Holmes in *The Sign of the Four* by Arthur Conan Doyle

This chapter provides details on our approach to pose, illumination, expression, and occlusion tolerant facial alignment [13] in section 3.1 and validates the same claims using a through set of experiments, in section 3.2, that demonstrate the effectiveness of our approach on many challenging datasets. It must be noted that illumination is dealt with by using models trained on images acquired under varying illumination conditions and by our local texture features (that are tolerant to illumination variations), and is also less focused on (and also less of a challenge to the facial alignment process) in our work than the other mentioned variations/degradations.

3.1 Overview of Our Approach

Our approach proceeds in a stage wise fashion. After a face in an image is detected and a standard sized crop of the face is generated, we first carry out a sparse landmarking step in which we use

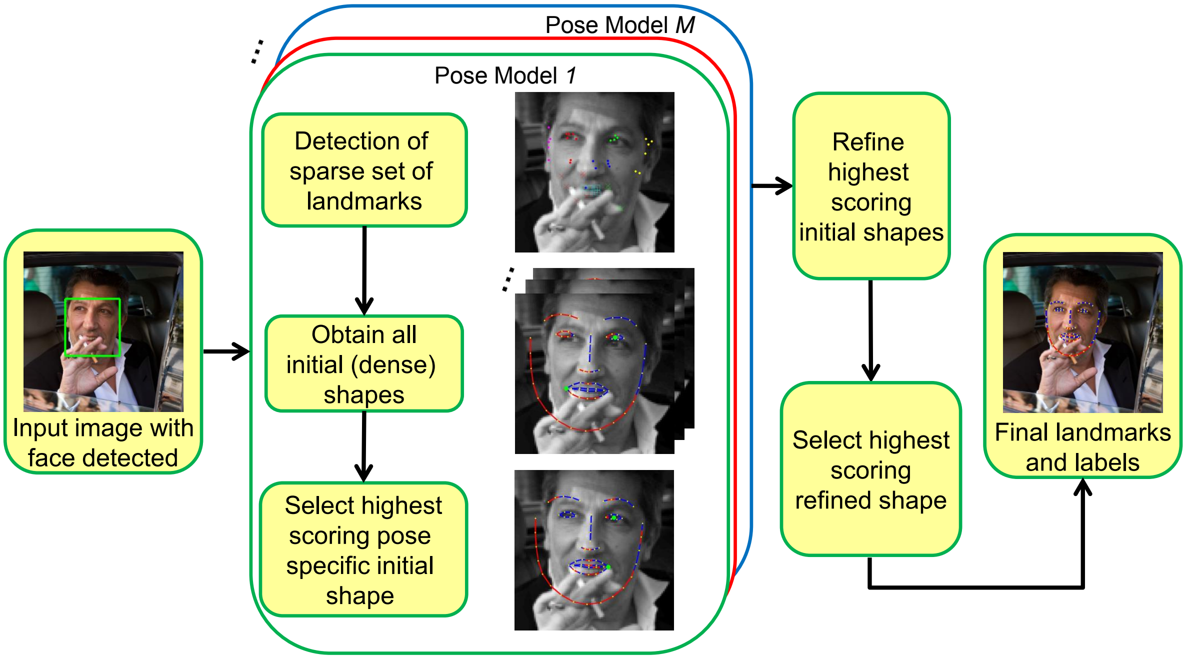
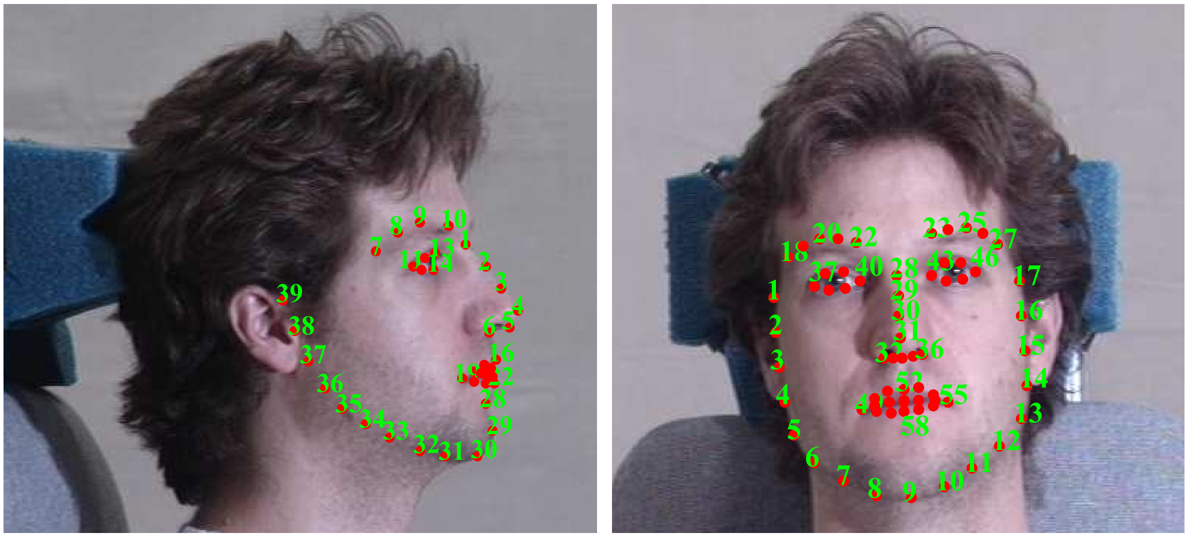

Pose Model *M*
Pose Model *1*
Input image with face detected
Detection of sparse set of landmarks
Obtain all initial (dense) shapes
Select highest scoring pose specific initial shape
Refine highest scoring initial shapes
Select highest scoring refined shape
Final landmarks and labels

Figure 3.1: An overview of our approach showing how each stage in the process works on an image from the test set partition of the Labeled Face Parts in the Wild (LFPW) dataset [4], [5]. In all facial images with landmarks overlaid on them, yellow dots are used to indicate the locations of facial landmarks, blue line segments indicate that the landmark at their center is accurately localized, and red line segments indicate that the landmark at their center is misaligned or potentially occluded. The same color scheme is maintained in all figures that show facial alignment results produced by our approach in this chapter. This figure appears in [13].

a sliding window based approach to search for only a few key facial landmarks (that we refer to as seed landmarks), such as the centers of the eyes, tip of the nose, tip of the chin, and the corners of the mouth, using pose-specific local detectors. It is important to note that we do not require all of these seed landmarks to be visible or accurately localized and only require that any combination of two of these landmarks be reliably localized. In our next step, we exhaustively evaluate denser pose-specific shapes that are obtained by taking all combinations (taken two at a time) of these seed landmarks and by using a similarity transformation to align a pose-specific mean shape of the full (dense) set of landmarks with them. These denser shapes are evaluated using a goodness of fit criteria based on whether each landmark in the dense set of landmarks is an inlier, *i.e.*, potentially resembling well aligned patches extracted from around that landmark at a



(a) (b)

Figure 3.2: MPIE landmarking (markup) schemes for (a) profile faces (39 facial landmarks) and (b) frontal faces (68 facial landmarks). The facial images in the figure are from the MPIE database. (a) and (b) appear in [13].

specific yaw at the training stage, or an outlier, *i.e.*, not resembling well aligned patches extracted from around that landmark at a specific yaw at the training stage and thus misaligned or possibly occluded. We are now able to retain a single (highest scoring) dense landmark based shape for each of the M discrete facial yaw ranges and transition from a step in which we located a sparse set of facial landmarks to a dense set of landmarks that best approximate the underlying textural information. The last step involves refining the top scoring shapes from among the M initial shapes and a ranking of the results to determine a single set of landmarks that are most likely well aligned with the facial image. This step also simultaneously uses and provides labels that indicate how many of the landmarks are localized with high confidence (inliers) or not (outliers). For carrying out a key part of this stage, we propose the use of an ℓ_1-regularized least squares based approach to regularize the deformed facial shapes using a dictionary of shapes. This technique is able to generate a more accurate regularized facial shape than the corresponding technique that is employed by ASMs. Figure 3.1 provides an overview of how our approach works and details on each of the stages in the algorithm follow.

3.1.1 Sparse Landmark Detection

As we have previously mentioned, our initial step in the alignment process is the detection of a sparse set of key facial landmarks that we refer to as seed landmarks. We train our models (see section 3.2.1 for details) using images with manual annotations available for 68 landmarks for for frontal faces (frontal, in this context, implies a facial yaw between $-45°$ and $+45°$) and 39 landmarks scheme for non-frontal (with an absolute yaw in excess of $45°$) faces. These landmarking schemes are shown in Figure 3.2 and were used (and hence popularized) to manually annotate some images in the MPIE database [1], [2], [3] by the database's curators. For frontal faces, we search for 8 seed landmarks, the centers of the two eyes, the tip of the nose, the corners of the mouth, the tip of the chin, and two opposite points on the facial boundary close to the ears (landmarks 2 and 16 in frontal faces and landmark 38 in profile faces in the landmarking schemes shown in Figure 3.2). The same corresponding set of seed landmarks is searched for in profile faces (faces that exhibit an absolute yaw angle greater than $45°$), however, the number of seed landmarks in such cases is 5, as a portion of the face is hidden from view in such cases.

During our training stage we construct landmark, expression, and pose-specific local appearance (texture) models for each landmark, including the seed landmarks. It is to be noted that we build and use $M = 10$ models for the various yaw ranges used in our approach. We also build 6 models for frontal yaw cases with open mouth expressions (scream and surprise). However, we do not use them at this stage, or our subsequent dense landmarking stage, to ensure a higher a fitting speed and since we found that only our final refinement step demanded the use of expression specific models to obtain highly accurate results. Section 3.2.1 provides details on how these models are built and the parameters used in their construction.

The first step in a model's construction is to generate a crop of a fixed size around the ground truth landmark locations. Following this step, a classifier is built for each landmark in every model

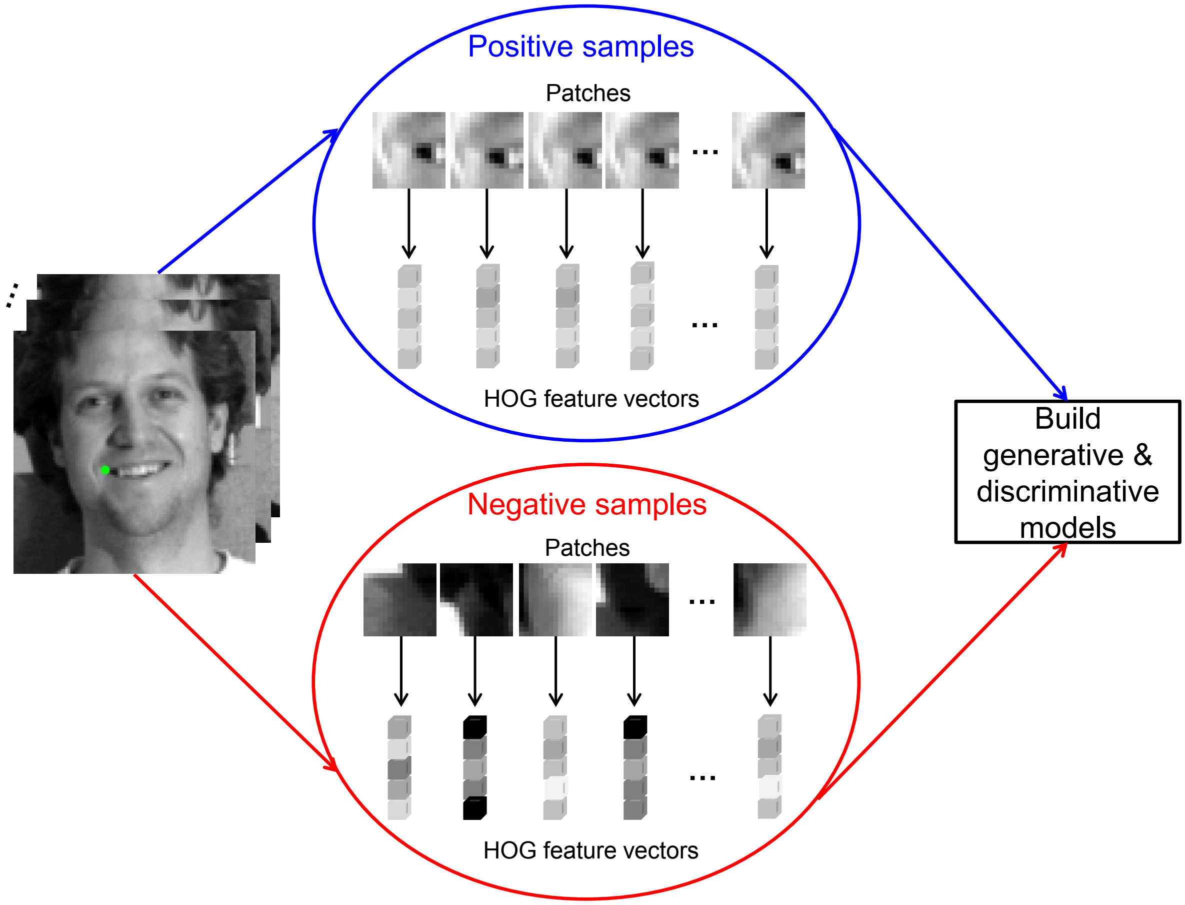


Figure 3.3: Process by which our local texture classifiers and linear subspaces are constructed for a specific landmark (mouth corner, marked with a green dot) and pose using a training set of various annotated images. The facial images in this figure are from the MPIE database.

to distinguish the local texture around the landmark in a particular feature space from the local texture of a different landmark or an occlusion. This is carried out by extracting features for positive samples, at the exact locations of the ground truth coordinates and from a small region around these locations, and negative samples, from random locations close to and far away from the ground truths, using all images for a specific yaw range and expression. We also construct separate linear subspaces (for the positive and negative classes using samples from the respective classes) using Principal Component Analysis (PCA) [109], [110], [111] as our dimensionality reduction technique. These subspaces are used in the next stage of our facial alignment pipeline (the dense landmark alignment stage). We use Histogram of Oriented Gradients (HOG) [93] as our feature descriptors as they have been proven to be quite discriminative in prior facial alignment algorithms, such as [6], [99], [112], [113], and are quite tolerant to illumination variation. As previously mentioned, this provides our approach with a level of illumination tolerance and we do not perform illumination compensation or further address the problem of illumination in a specific fashion in our work as it poses less of a challenge to the facial alignment process than factors such as pose and the presence of occlusions. This is because local texture based alignment approaches are less susceptible to this problem than global texture based approaches. Figure 3.3 illustrates how our local appearance models and subspaces are constructed for each landmark while section 3.2.1 provides exact details on how many such models are built and the parameters used in their construction.

Our local texture classifiers are constructed using an ensemble of classifiers (decision stumps) in a Real AdaBoost [114] framework. We chose the Real AdaBoost framework due to the minimal parameters that need to be specified for such a classifier (only the number of boosting rounds or number of classifiers in the ensemble need to be specified) and its resistance to overfitting [115], [116]. It must be noted that while any classifier that provides a measure of confidence in its classification output could be used in our approach, we determined that the Real AdaBoost implementation that was used in our work edged out implementations of other classifiers, such as Support Vector Machines (SVMs) [117], random forests [118], and random ferns [119] in the accuracy vs speed and memory trade-off. Real AdaBoost has also been used quite frequently and successfully in the past for carrying out facial alignment [113], [120], [121], [122]. The Real AdaBoost framework not only allows for the classification of a feature vector as positive or negative (misaligned or possibly occluded), but also returns a confidence score for the prediction. This allows us to greedily retain the highest scoring locations in the response map for a particular seed landmark when a search over the typical region where the landmark is likely to lie is performed on a test face crop. The search is repeated for rotated versions of the crop (typically for rotation

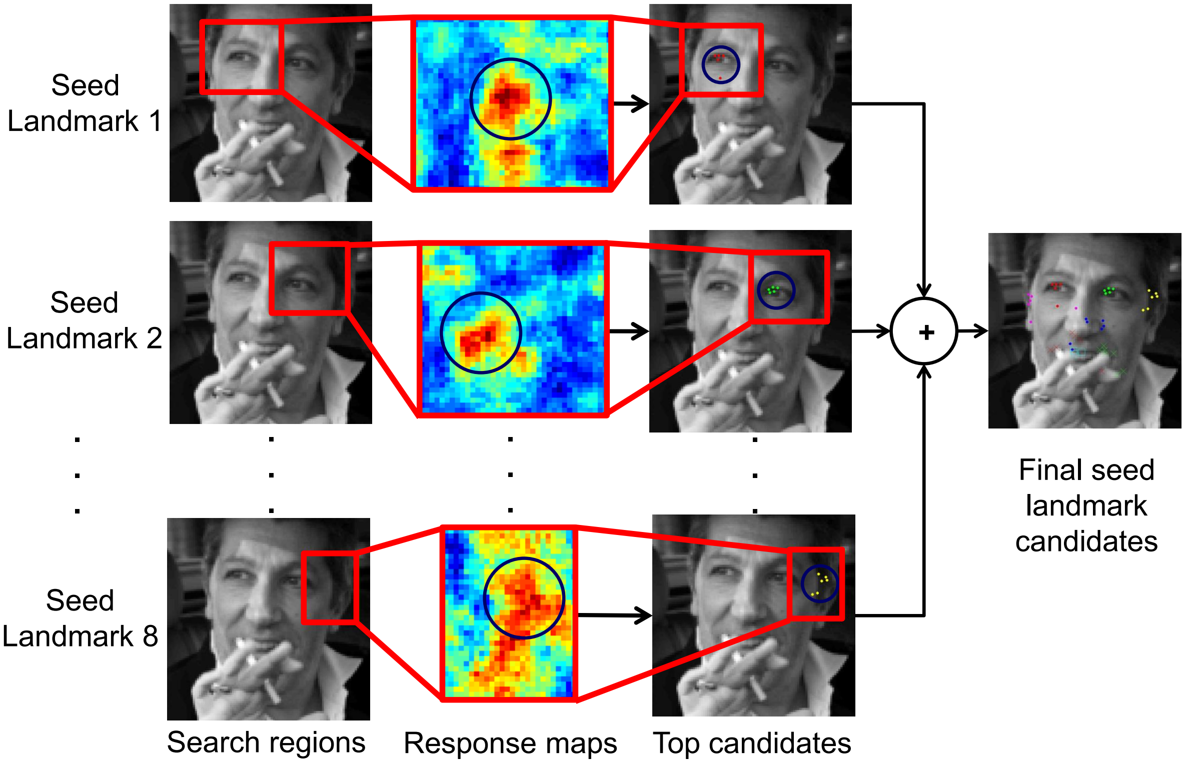


Figure 3.4: Process by which seed landmark candidates are retained by our approach when fitting a facial image from the test set partition of the LFPW dataset. The process is shown only for one of the pose models (for a yaw of $0°$ to $+15°$) and is repeated to retain seed landmark candidates specific to each pose model. This figure appears in [13].

angles between $-30°$ and $+30°$ in $15°$ increments) with clustering used to reduce the number of candidates if a number of them are found to lie within a small bandwidth. Figure 3.4 shows how we retain candidates for the various seed landmarks for a particular pose-specific model. An overview of the Real AdaBoost algorithm can be found in Appendix A while a detailed explanation of the method and related proofs can be found in [114].

3.1.2 Dense Landmark Alignment and Optimal Shape Initialization

Once we have pose-specific seed landmark candidates, the problem to be addressed is one of selecting a single combination of candidates for two different seed landmarks that allows for the optimal initialization of a pose-specific mean facial shape $\bar{\mathbf{s}}^m$ $(m = 1, \ldots, M)$ consisting of the full set of facial landmarks for that pose model, *i.e.*, alignment of a dense set of pose-specific landmarks. By aligning each pose-specific mean shape with a combination of seed landmarks we end up with a total set of J_m dense shapes $\mathbf{s}^{j,m}$ $(j = 1, \ldots, J_m)$ that must be ranked using a scoring function that numerically assesses their goodness of fit. This step is extremely important to the alignment process because poor initialization is something a facial alignment algorithm can seldom recover from [56]. Thus, our contribution in providing a framework to transition from a set of sparse landmarks (possibly containing some spurious detections) to a dense set of initial landmarks is quite important.

At this point, it is necessary to provide details on how shape models (also sometimes referred to as Point Distribution Models (PDMs)) for CLMs work. Each facial shape $\mathbf{s}$ in the training set is represented by its N x and y coordinates in vectorial form as $\mathbf{s} = [x_1 \;\; x_2 \ldots x_N \;\; y_1 \;\; y_2 \ldots y_N]^{\mathrm{T}}$. These shapes are aligned using Generalized Procrustes Analysis (GPA) [123], [124], to normalize for scale, rotation, and translation effects and bring them into a common reference frame. In this reference frame, conventional PDMs are built by obtaining a mean facial shape $\bar{\mathbf{s}}$ and by constructing a PCA subspace $\mathbf{\Phi}$ of facial shape variation. A facial shape can now be represented using equation (3.1).

$$\mathbf{s} = T_{s,\theta,x_t,y_t}(\bar{\mathbf{s}} + \mathbf{\Phi}\mathbf{b}) \tag{3.1}$$

In equation (3.1), T is a similarity transformation parametrized by a scaling factor s, a rotation angle θ, and translation parameters x_t and y_t. The result obtained when the transformation T is applied to a single point (x, y) is shown in equation (3.2).

$$T_{s,\theta,x_t,y_t}\begin{pmatrix} x \\ y \end{pmatrix} = \begin{bmatrix} s\,cos\theta & -s\,sin\theta \\ s\,sin\theta & s\,cos\theta \end{bmatrix} \begin{bmatrix} x \\ y \end{bmatrix} + \begin{bmatrix} x_t \\ y_t \end{bmatrix} \tag{3.2}$$

The entire shape fitting process centers around the determination of the optimal vector shape coefficients $\mathbf{b}$ that best represents (and regularizes) the current set of landmarks whose locations are

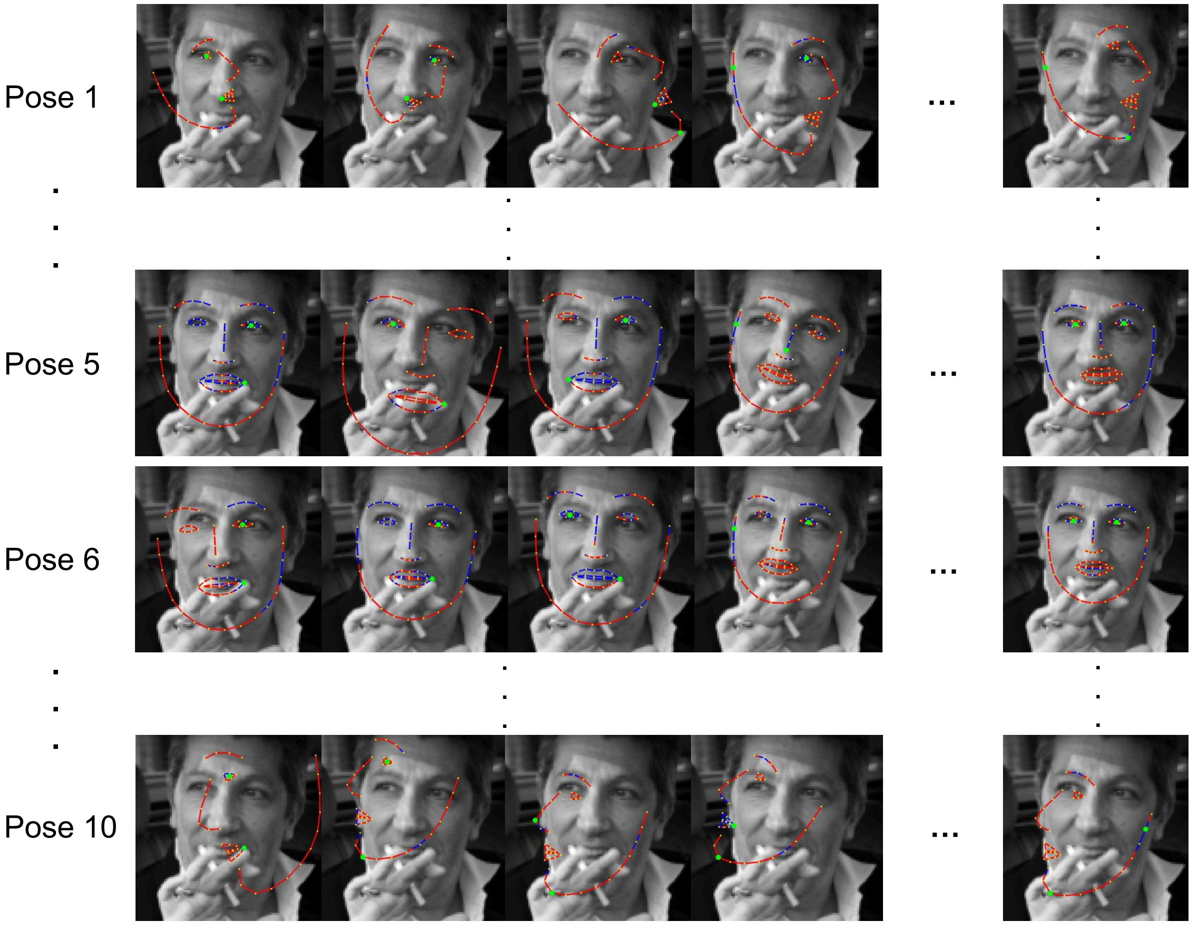


Figure 3.5: Samples of initially aligned shapes for the $M = 10$ pose models for a facial image from the test set partition of the LFPW dataset. In all facial images with landmarks overlaid on them, green dots are used to indicate the seed landmark candidates used to generate the aligned shapes, yellow dots are used to indicate the locations of facial landmarks, blue line segments indicate that the landmark at their center is accurately localized, and red line segments indicate that the landmark at their center is either misaligned or occluded.

determined using local texture based methods.

Since facial shape varies dramatically with pose and expressions, we construct 16 pose and expression specific PDMs in our approach. However, our approach to determine the shape coefficients is a novel method that does not use the conventional shape model equation in (3.1). Instead of using PCA to determine our set of basis vectors, we retain the entire set of shapes in a pose and expression specific dictionary that we later use in an ℓ_1-regularized least squares based approach to determine the shape coefficients. However, we do retain the building of pose and expression specific mean canonical shapes in our approach as well. It is the $M = 10$ pose-specific mean shapes $\bar{\mathbf{s}}^m$ $(m = 1, \ldots, M)$ that we use to determine the best initialization that can be provided for each pose range (the best fitting mean shape roughly aligned over the face for each pose range).

Figure 3.5 illustrates how each of 10 pose-specific mean shapes is aligned with every combination of seed landmark candidates for that pose and then scored using a scoring function. It is to be noted that before the scoring is performed, the region around the shape is cropped and derotated (since the angle of rotation required can be calculated using two fixed landmarks) in order to match the crops generated during our training process. For profile poses, a fewer set of shapes need to be evaluated as there are fewer seed landmarks. For example, in a frontal case with 8 seed landmarks taken 2 at a time with 10 candidates for each of them, the number of shapes that would need to be scored is $J = 10 \times 10 \times \binom{8}{2} = 2800$, while for a profile case with 5 seed landmarks, the corresponding number would only be $J = 1000$. All of these shapes $\mathbf{s}^{j,m}$ $(j = 1, \ldots, J_m)$ must be scored in a way that maximizes the joint probability of correct alignment of the landmark coordinates $\mathbf{x}_n^{j,m} = [x_n^{j,m} \; y_n^{j,m}]^{\mathrm{T}}$ $(n = 1, \ldots, N_m)$ in the shape. This joint probability of correct alignment for the full set of N_m landmarks in shape $\mathbf{s}^{j,m}$ for a particular pose model m is given by equation (3.3), assuming the conditional independence of the individual probabilities of correct alignment for the landmarks. In equation (3.3), $I_n^{j,m} \in \{-1, +1\}$ $(n = 1, \ldots, N_m)$ denotes whether landmark $\mathbf{x}_n^{j,m}$ is correctly aligned or not.

$$P(I_1^{j,m} = 1, I_2^{j,m} = 1, \ldots, I_{N_m}^{j,m} = 1 | \mathbf{s}^{j,m}) = \prod_{n=1}^{N_m} P(I_n^{j,m} = 1 | \mathbf{x}_n^{j,m}) \tag{3.3}$$

To use equation (3.3) as the objective function to maximize in order to find the highest scoring aligned shape would require the modeling of the individual probabilities for each landmark. This could be carried out by modeling the distributions of the texture features extracted around each landmark using parametric or non-parametric methods. However, there are simpler scoring functions that could be used as surrogates for this joint probability function that suit our purpose. The key point to take note of here is that only a finite set of shapes need to be evaluated and scored and that this is a different problem from one that involves the optimization of a continuous function. It is for this reason that we use a different scoring function in order to evaluate the set of shapes $\mathbf{s}^{j,m} \ \ (j = 1, \ldots, J_m)$ on their goodness of fit, which is assessed by determining how well a landmark's surrounding local texture matches pre-trained models of what this local texture looks like in a particular feature space.

We first project the feature vector $\mathbf{t}_n^{j,m}$ (obtained using the local texture around a landmark $\mathbf{x}_n^{j,m}$ in the shape) onto the positive subspace ${\mathbf{\Psi}_{\text{pos}_n}}^m$ (after subtracting the mean texture vector ${\bar{\mathbf{t}}_{\text{pos}_n}}^m$ for the subspace) for that landmark and pose to obtain coefficients ${\mathbf{c}_{\text{pos}_n}}^m$, using equation (3.4).

$$
{\mathbf{c}_{\text{pos}_n}}^m = ({\mathbf{\Psi}_{\text{pos}_n}}^m)^{\text{T}}(\mathbf{t}_n^{j,m} - {\bar{\mathbf{t}}_{\text{pos}_n}}^m) \tag{3.4}
$$

These coefficients are used to generate a reconstruction ${\mathbf{t}'_{\text{pos}_n}}^{j,m}$, using equation (3.5).

$$
{\mathbf{t}'_{\text{pos}_n}}^{j,m} = {\bar{\mathbf{t}}_{\text{pos}_n}}^m + {\mathbf{\Psi}_{\text{pos}_n}}^m {\mathbf{c}_{\text{pos}_n}}^m \tag{3.5}
$$

The reconstruction is in turn used to compute a reconstruction error vector whose norm $r_{\text{pos}}(\mathbf{x}_n^{j,m})$ is given by equation (3.6).

$$
r_{\text{pos}}(\mathbf{x}_n^{j,m}) = \|({\mathbf{t}'_{\text{pos}_n}}^{j,m} - \mathbf{t}_n^{j,m})\|_2 \tag{3.6}
$$

The same process is followed using the negative subspace for the specific landmark to obtain $r_{\text{neg}}(\mathbf{x}_n^{j,m})$. We then calculate the ratio of the reconstruction error norms $r_{\text{pos}}(\mathbf{x}_n^{j,m})$ and $r_{\text{neg}}(\mathbf{x}_n^{j,m})$ for a particular landmark using equation (3.7).

$$
r(\mathbf{x}_n^{j,m}) = \frac{r_{\text{pos}}(\mathbf{x}_n^{j,m})}{r_{\text{neg}}(\mathbf{x}_n^{j,m})} \tag{3.7}
$$

Next, the average $R^{j,m}$ of these ratios over all N_m landmarks in shape $\mathbf{s}^{j,m}$ is calculated using

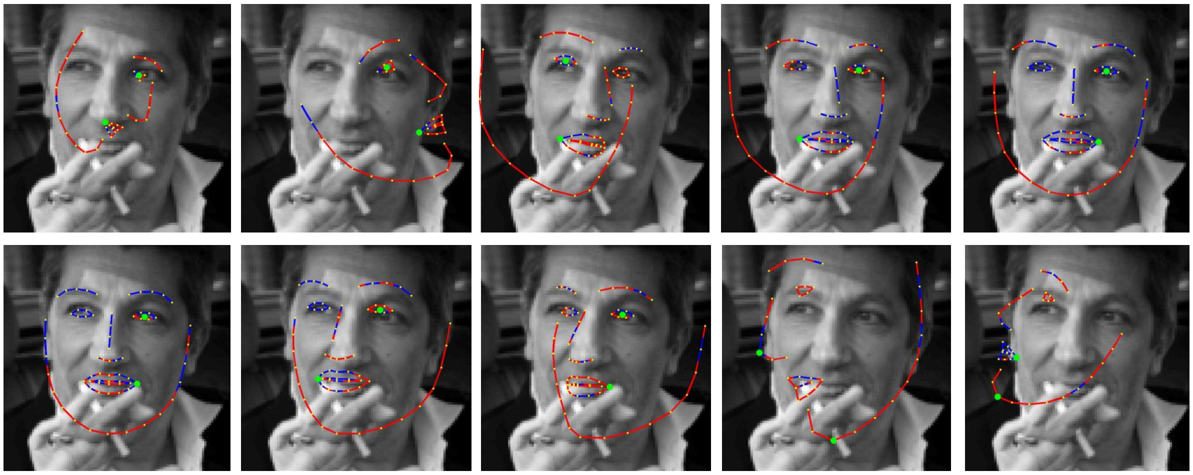


Figure 3.6: The highest scoring aligned initial shapes for each of the $M = 10$ pose models for a facial image from the test set partition of the LFPW dataset. This figure appears in [13].

equation (3.8).

$$R^{j,m} = \frac{1}{N_m} \sum_{n=1}^{N_m} r(\mathbf{x}_n^{j,m}) \tag{3.8}$$

Finally, we use this reconstruction error based metric in combination with knowledge of the number of inliers $N_{\text{inliers}}^{j,m}$ in shape $\mathbf{s}^{j,m}$, *i.e.*, the number of landmarks in the shape that are classified as accurately aligned by our local texture based classifier in our shape scoring function $f(\mathbf{s}^{j,m})$. $f(\mathbf{s}^{j,m})$ is determined using equation (3.9) and is the final metric we use to determine the "best" aligned initial shape from among the total set of J_m shapes for pose m.

$$f(\mathbf{s}^{j,m}) = \frac{N_{\text{inliers}}^{j,m}}{R^{j,m}} \tag{3.9}$$

The assumption here is that well aligned shape will contain more inliers than a poorly aligned one and hence will end up with a high value for the numerator and a low value for the denominator in equation (3.9). The highest scoring aligned shape $\mathbf{s}_{\text{init}}^{m}$ for each pose range from among the J_m evaluated shapes can be determined using equations (3.10) and (3.11) and used as initialization for the final step in our alignment process.

$$j_0 = \arg\max_j \; f(\mathbf{s}^{j,m}) \tag{3.10}$$

$$\mathbf{s}^m_{\text{init}} = \mathbf{s}^{j_0,m} \tag{3.11}$$

Figure 3.6 shows the highest scoring aligned shapes for each pose range for a sample test image.

3.1.3 Shape Refinement

The last stage of our alignment algorithm involves the refining (deforming and regularizing of a shape) of the highest scoring initial shapes that were obtained using the previous stage and the selection of one of these refined shapes as the final locations of the facial landmarks. To carry this out we use an iterative fitting process that has it roots in ASMs and CLMs. In practice, to allow for a gain in fitting speed, only a few ($M^{'} < M$) of the highest scoring fitting M initial shapes $\mathbf{s}^m_{\text{init}}$ $(m = 1, \ldots, M)$ are selected for refinement to obtain shapes $\mathbf{s}^{m^{'}}_{\text{ref}}$ $(m^{'} = 1, \ldots, M^{'})$. It is also to be noted that during the refinement process we also score results produced using the open mouth expression shape and texture models for the frontal pose ranges and the higher scoring of the open mouth and closed mouth fitted shapes are retained for each pose $m^{'}$.

A window around each landmark's current location is generated and the local texture around each pixel in the window is scored and classified using our local texture classifiers. The landmarks are independently moved into the highest scoring locations for them. The process is repeated for a few iterations until the landmarks converge. However, between each iteration, the facial shape produced as a result of landmark motion must be regularized in order to generate a shape that is consistent with what a typical facial contour looks like. We carry out this regularization using a novel technique that allows for a higher fitting accuracy compared to the regularization method employed by ASMs. Figure 3.7 illustrates how one iteration of this process is carried out. Finally,

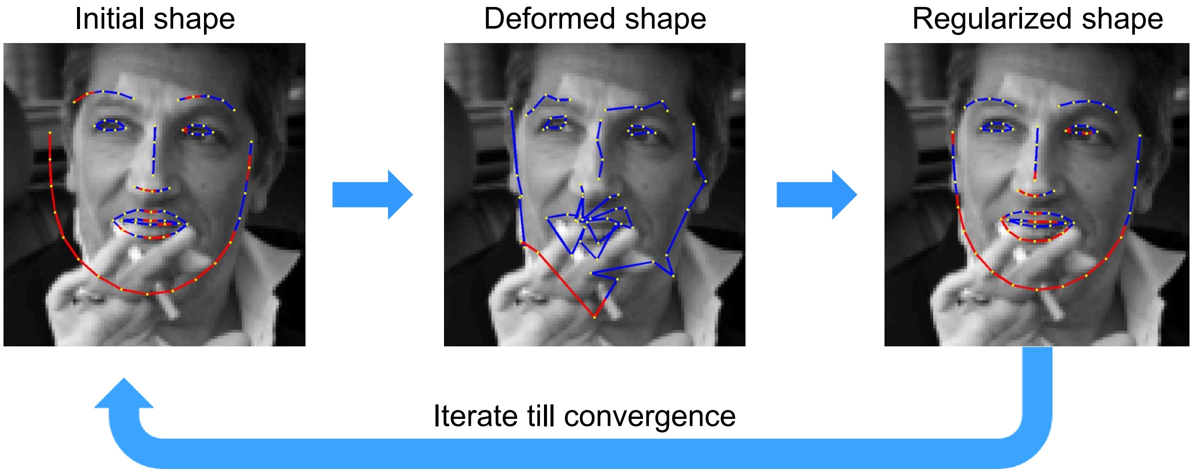


Figure 3.7: Iterative process used in our shape refinement step demonstrated on a facial image from the test set partition of the LFPW dataset. This figure appears in [13].

the highest scoring shape from among the refined shapes is identified and returned.

ℓ_1-Regularized Least Squares Based Shape Coefficients Determination

Shape regularization involves the determining and updating of a vector of shape coefficients. Consider an initial shape $\mathrm{s}_{\text{init}}^{m'}$ (we drop the superscript m' in this section for the sake of simpler notation). After each of the landmarks in the shape have been allowed to independently move into the optimal locations for them, the new shape obtained is denoted by s_{def}. In an ASM based approach, the inverse of the similarity transformation T that best aligns the mean shape $\bar{\mathrm{s}}$ with s_{def} is applied to s_{def} (in the image space) to generate s'_{def} (in the model space). The problem becomes one of determining the optimal set of shape coefficients $\mathbf{b}_{\text{init}}$ to minimize the objective function in equation (3.12).

$$\mathbf{b}_{\text{init}} = \arg\min_{\mathbf{b}} \ \|\mathbf{\Phi b} - (\mathbf{s}'_{\text{def}} - \bar{\mathrm{s}})\|_2^2 \tag{3.12}$$

In equation (3.12), $\mathbf{\Phi}$ is a previously trained orthonormal linear subspace of shape variation (all shapes being aligned using Procrustes analysis before the building of the subspace) with dimensions $d \times u$ ($d > u$) where $d = 2N$ is the dimension of each shape vector, u is the number of eigenvectors retained in order to account for $95 - 97\%$ of the shape variance (and also the dimensionality of the shape coefficients vector), and $\bar{\mathrm{s}}$ is the mean shape. The solution to the overdetermined Least Squares Problem (LSP) in equation (3.12) is given by equation (3.13).

$$\mathbf{b}_{\text{init}} = \mathbf{\Phi}^{+}(\mathbf{s}'_{\text{def}} - \bar{\mathbf{s}}) \tag{3.13}$$

In equation (3.13) $\mathbf{\Phi}^{+}$ denotes the left Moore-Penrose pseudoinverse of $\mathbf{\Phi}$. Since $\mathbf{\Phi}$ is an orthonormal basis, $\mathbf{\Phi}^{+} = (\mathbf{\Phi}^{\mathrm{T}}\mathbf{\Phi})^{-1}\mathbf{\Phi}^{\mathrm{T}} = \mathbf{\Phi}^{\mathrm{T}}$ and equation (3.13) gets simplified to equation (3.14).

$$\mathbf{b}_{\text{init}} = \mathbf{\Phi}^{\mathrm{T}}(\mathbf{s}'_{\text{def}} - \bar{\mathbf{s}}) \tag{3.14}$$

The values of $\mathbf{b}_{\text{init}}$ are constrained to lie within three standard deviations of their zero mean values (based on the assumption that these coefficient values are distributed according to a zero mean Gaussian distribution) in order to generate plausible shapes (regularization) and this results in a new vector of shape coefficients denoted by $\mathbf{b}_{\text{mod}}$. In practice, the shape coefficients and the similarity transformation parameters are determined by first initializing the shape coefficients to zero and iteratively determining a new set of values and transformation parameters simultaneously [66]. The entire process of landmark displacement, shape coefficient vector determination, constraining of shape coefficients, and generation of a new regularized set of landmark coordinates is repeated for a few iterations until the shape parameter values or the landmark coordinates do not change by much between iterations. A regularized shape s_{reg} is obtained when the final set of shape coefficients are applied and the resulting shape is aligned back into the image space using the transformation T, as shown in equation (3.15).

$$\mathrm{s}_{\text{reg}} = T(\bar{\mathrm{s}} + \mathbf{\Phi}\mathbf{b}_{\text{mod}}) \tag{3.15}$$

In our approach, rather than constructing a PCA subspace to model shape variation, we retain the entire dictionary of shapes for each pose model. Thus, the analogue to the previously defined $\mathbf{\Phi}$ is a dictionary of shape variation $\mathbf{D}$ of size $d \times v$ $(d < v)$, where $d = 2N$ is the dimension of each shape vector in the dictionary and v is the number of such training shapes for a specific yaw model and also the dimensionality of the shape coefficients vector. We recast the problem of shape regularization using equation (3.16), in which λ is a regularization parameter, and generate a regularized shape using equation (3.17).

$$\hat{\mathbf{b}} = \arg\min_{\mathbf{b}} \; \|\mathbf{D}\mathbf{b} - \mathbf{s}'_{\text{def}}\|_2^2 + \lambda\|\mathbf{b}\|_1 \tag{3.16}$$

$$\mathbf{s}_{\text{reg}} = T(\mathbf{D}\hat{\mathbf{b}}) \tag{3.17}$$

What we achieve by formulating the problem in this fashion is that simultaneous determination and regularization of shapes is now possible using a single objective function without the need for the additional step involved in ASMs to modify the shape coefficients based on the Gaussian assumption. Our formulation makes no assumptions about the distribution of the coefficients, is not a linear function of $\mathbf{s}'_{\text{def}}$ (as is the case in equation (3.12)), and allows for a data driven framework to achieve regularization, which is a key area of focus in in [4] and [88] as well.

The problem in equation (3.16) is commonly called the ℓ_1-regularized LSP whose general form is given by equation (3.18).

$$\underset{\mathbf{x}}{\text{minimize}} \; \|\mathbf{A}\mathbf{x} - \mathbf{y}\|_2^2 + \lambda\|\mathbf{x}\|_1 \tag{3.18}$$

In equation (3.18), $\mathbf{A} \in \mathbb{R}^{p\times q}$ is a data matrix, $\mathbf{y} \in \mathbb{R}^p$ is a vector of observations, $\mathbf{x} \in \mathbb{R}^q$ is a vector of unknowns, and $\lambda > 0$ is the regularization parameter. The problem in equation (3.18) is convex but not differentiable. It always has a solution, but there is no closed form equation to obtain it. However, it is possible to compute a solution numerically. The problem has been well studied and is also closely related to the problems of Basis Pursuit Denoising (BPDN) [125] and least absolute shrinkage and selection operator (Lasso) [126]. An ℓ_1-Regularized LSP can be transformed into a convex quadratic problem with linear inequality constraints and solved by standard convex quadratic methods, such as interior-point methods [127], [128], homotopy methods and variants [129], [130], [131], and also by subgradient methods [132], [133]. However, some of these solvers can be quite slow and also only efficient when the solution is very sparse. Providing details on each of of these solvers and analyzing their impact is beyond the scope of this thesis, however, we determined that a custom interior point based method for solving large scale ℓ_1-regularized LSPs that was developed by Kim *et al.* [134] was ideally suited for our purposes and is the solver we use in our work. [134] also provides details on the limiting behavior of the solution to the problem as $\lambda \rightarrow 0$ and $\lambda \rightarrow \infty$. A key result that is outlined in [134] that governs the choice of the regularization parameter λ is that for $\lambda \geq \lambda_{\max} = \|2\mathbf{A}^{\mathrm{T}}\mathbf{y}\|_\infty$ ($\|2\mathbf{D}^{\mathrm{T}}\mathbf{s}'_{\mathrm{def}}\|_\infty$, in our problem setup) an all zero vector becomes the optimal solution. A value of $\lambda = 10^{-4}\lambda_{\max}$ was recommended by Kim *et al.* (when using an open source MATLAB [135] implementation of their code [136]) and such a value was empirically found to be suitable for the purposes of our problem as well. An overview of the solver is provided in Appendix B and the reader is referred to [134] for further details that are not provided in the appendix.

Shape regularization can be carried out more accurately if only inliers are used in the process. Since this is possible in our approach, using the results produced by local texture classifiers, we exclude all outliers from participating in the shape regularization process and only use the rows of $\mathbf{D}$ ($\mathbf{\Phi}$ in the case of the previously described PCA based approach that is used by ASMs) that correspond to these inlier landmarks. The shape coefficients obtained using this process can be used to reconstruct a full set of landmarks and hallucinate the locations of the outliers.

An important set of results that we highlight in section 3.2.5 is that even when only the inliers are used for shape regularization, our ℓ_1-regularized approach outperforms the previously outlined approach used in ASMs to obtain more accurate landmark localization results on several datasets. In addition, we also demonstrate that the ℓ_1-regularized approach provides a higher level of accuracy than using an ℓ_2-regularized (Tikhanov regularization) based approach (details on this problem can also be found in [134]), when the same value of λ is used. In such an ℓ_2-regularized approach, a closed form solution to the problem in equation (3.19) is provided by $\mathbf{x} = (\mathbf{A}^{\mathrm{T}}\mathbf{A} + \lambda\mathbf{I})^{-1}\mathbf{A}^{\mathrm{T}}\mathbf{y}$.

$$\underset{\mathbf{x}}{\text{minimize}} \;\; \|\mathbf{A}\mathbf{x} - \mathbf{y}\|_2^2 + \lambda\|\mathbf{x}\|_2^2 \tag{3.19}$$

Our intuition behind these results is that the ℓ_1-regularized approach results in a deformed shape either being modeled using a smaller set of training shapes (in the training shapes dictionary) that best approximate their locations or by using smaller weights for unsuitable training shapes in the dictionary. This sparsity promotion results in a deformed shape being better approximated using a set of coefficients that is not a linear function of the deformed shape itself, and thus in higher landmark localization accuracies than those obtained using an ℓ_2-regularized approach.

Final Shape Scoring and Selection

The last step in our alignment process is the selection of a single shape from among the set of $M^{'}$ ($3 \leq M^{'} \leq 5$ in our work) refined shapes that best fits the underlying facial structure. The shape $\mathbf{s}_{\text{fin}}$ with the highest percentage of inliers expressed as a percentage of the total number of landmarks in the shape, using the scoring function $g(\mathbf{s}_{\text{ref}}^{m^{'}})$ in equation (3.20), is chosen to obtain a final set of landmark coordinates using the following equations.

$$g(\mathbf{s}_{\text{ref}}^{m^{'}}) = \frac{N_{m^{'}}^{\text{inliers}}}{N_{m^{'}}} \quad \text{where } m^{'} = 1, \ldots, M^{'} \tag{3.20}$$

$$m_0^{'} = \underset{m^{'}}{\arg\max} \;\; g(\mathbf{s}_{\text{ref}}^{m^{'}}) \tag{3.21}$$

$$\mathbf{s}_{\text{fin}} = \mathbf{s}_{\text{ref}}^{m_0^{'}} \tag{3.22}$$

The same scoring function is also used to determine the best fitting expression specific shape for each pose $m^{'}$. Now that we have provided a detailed description of each of the steps in our

Algorithm 1 Our facial alignment approach.

```
Input: Image $\mathbf{I}$, face detection bounding box $\mathbf{b}_b$, pre-trained yaw and expression specific models $\{\mathbf{M}_m\}_{m=1}^{16}$
Output: Final landmarks (shape) $\mathbf{s}_{\text{fin}}$, landmark misalignment/occlusion labels for $N$ landmarks $\{\mathbf{o}_n\}_{n=1}^{N}$

for $m = 1, \ldots, M$ do
    Retain top seed landmark candidates $\{\mathbf{p}_1^m, \mathbf{p}_2^m, \ldots, \mathbf{p}_{10}^m\}_{i=1}^{N_m^l}$ for each of the $N_m^l = 5$ or $N_m^l = 8$ seed landmarks for pose $m$
end for

for $m = 1, \ldots, M$ do
    Score all dense shapes $\mathbf{s}^{j,m}\ \ (j = 1, \ldots, J_m)$ using equations (3.7) - (3.9)
    Retain highest scoring initial shape $\mathbf{s}_{\text{init}}^{m}$ using equations (3.10) and (3.11)
end for

Retain top scoring $M^{'}$ pose-specific initial shapes
for $m^{'} = 1, \ldots, M^{'}$ do
    Refine $\mathbf{s}_{\text{init}}^{m^{'}}$ to obtain $\mathbf{s}_{\text{ref}}^{m^{'}}$ using model $\{\mathbf{M}_{m^{'}}\}$
end for

Retain highest scoring refined shape $\mathbf{s}_{\text{fin}}$ using equations (3.20) - (3.22)
Output final landmarks (shape) $\mathbf{s}_{\text{fin}}$ and landmark misalignment/occlusion labels for $N$ landmarks $\{\mathbf{o}_n\}_{n=1}^{N}$
```

approach, a summary of the steps is provided in Algorithm 1.

3.2 Experiments and Results

In this section we provide details on how our approach was trained and describe the experiments that we carried out in order to demonstrate the effectiveness of our algorithm when it was tested on challenging real-world datasets.

3.2.1 Training our Algorithm

We trained a set of models using a subset of the CMU Multi-PIE (MPIE) database [1], [2], [3]. Our shape and texture models were trained using a total of $6,495$ images of almost all of the 337 subjects drawn from across all 4 sessions in the database. The images contained faces imaged under differing illumination conditions, showing various expressions (neutral, disgust, smile, squint, scream, and surprise), and acquired from 13 different viewpoints from -90° to $+90^\circ$ in steps of 15°. Manually annotated ground truths for all these images were available to us as a small subset of the MPIE database was annotated using 68 landmarks for frontal faces and 39 landmarks for profile faces by the database's curators. We clustered the data into $M = 10$ bins with overlapping yaw ranges and the same number of facial landmarks for every image in the bin, *i.e.*, -90° to -75°, -75° to -60°, -45° to -30°, -30° to -15°, -15° to 0°, and 5 more similar bins for the positive yaw cases. These 10 partitions were created using facial images with the mouth slightly open or closed (neutral, disgust, smile, and squint expressions). A similar set of 6 partitions (for frontal poses with a yaw range from -45° to $+45^\circ$ only) were created to model the shape and texture of facial landmarks across pose in expressions when the mouth is completely open (scream and surprise expressions).

Our models were built using the process described in section 3.2.1 with Real AdaBoost as our choice of classifier. However, when we tested on a subset of the MPIE dataset, we only trained on three-fourths of the training data with a test set drawn from the remaining images. Thus, we always tested our algorithm on unseen images and subjects. We used an open source MATLAB [135] toolbox [137] to extract the HOG features and implement the training and testing of the Real AdaBoost framework. A standard facial crop size of 100×100 and a patch size of 15×15, that were found to be optimal on a validation set, were used by us to extract HOG feature descriptors and build the local texture models (classifiers).

Table 3.1: Details on the datasets used in our experiments.

Dataset	Training Set Size	Test Set Size	Yaw Variation	Expression Variation	Facial Occlusion Level	Number of Landmarks
MPIE	___	850	$-90°$ to $+90°$	Yes	Not Present	$39/68$
LFPW	811	224	$-45°$ to $+45°$	Yes	Low	68
Helen	$2,000$	330	$-45°$ to $+45°$	Yes	Low	68
AFW	___	337	$-45°$ to $+45°$	Yes	Medium	68
ibug	___	135	$-45°$ to $+45°$	Yes	Medium	68
COFW	507	500	$-45°$ to $+30°$	Yes	High	29

3.2.2 Datasets Used in Our Experiments

Details on the various datasets which were used in our experiments (for benchmarking our approach, understanding what each stage of our approach contributed, *etc.*) are provided below and summarized in Table 3.1.

(1) MPIE: A set of 850 images were held back from our training set and served as a test set. These images contained faces with varying expressions and with yaw angles in the range from $-90°$ to $+90°$. This test set was created to demonstrate that our algorithm could deal with such variations in unseen images from outside its training set and was also used to benchmark our approach against the TSMs algorithm, which could also handle this range of yaw variation.

(2) LFPW: The Labeled Face Parts in the Wild (LFPW) dataset [4], [5] originally consisted of 1132 training images and 300 test images of various people (mainly celebrities) that were collected from the Internet and manually annotated with 29 landmarks. Many of the URLs for the images in the dataset have expired, however, a set of 811 training images and 224 test images were recently made available along with landmark annotations [138], [139] for the 68 landmarks in the MPIE markup as part of the the 300 Faces in-the-wild (300-W 2013) challenge [8], [9]. All algorithms were tested on the 224 images in the test set partition of the dataset. The faces in the images exhibit slight pose variation (absolute yaw of up to 45° and slight in-plane rotation), varying expressions, and low levels of occlusion.

(3) Helen: The Helen dataset [104], [105] consists of 2000 training images and 330 test images of various people that were collected from the Internet and manually annotated with 194 landmarks. Landmark annotations [138], [139] for the 68 landmarks in the MPIE markup for all images were recently made available as part of the the 300-W 2013 challenge. All algorithms were tested on the 330 images in the test set partition of the dataset. The faces in the images exhibit slight pose variation (absolute yaw of up to 45° and slight in-plane rotation), varying expressions, and low levels of occlusion.

(4) AFW: The Annotated Faces in-the-Wild (AFW) dataset [6], [7] consists of 205 images with 468 faces (some images contain multiple faces) drawn from Flickr images. Facial bounding boxes, manual annotations for 6 landmarks (the centers of the eyes, the tip of the nose, and the two corners and center of the mouth), and discretized pose information were originally made available along with the images. As part of the 300-W 2013 challenge, 68 point annotations for 337 faces in the images were made available [138], [139], which served as a test set. The faces in the images exhibit pose variation (absolute yaw of up to 45° and slight in-plane rotation), varying expressions, and facial occlusions.

(5) ibug: The Intelligent Behavior Understanding Group (ibug) dataset [8], [9], [10] consists of 135 facial images with annotations for 68 landmarks for each of the faces. The dataset was made publicly available as part of the 300-W 2013 challenge. The images in this dataset are very challenging due to the high pose variation exhibited (both in pitch and yaw), as well as the presence of varying expressions and facial occlusions.

(6) COFW: The Caltech Occluded Faces in the Wild dataset [11], [12] consists of 500 training and 507 test images that were downloaded from the Internet. All images were manually annotated with the same 29 landmarks that were used in the exemplar based facial alignment method proposed by Belhumeur *et al.* [4]. The faces in the images exhibit slight pose variation (yaw between $-45°$ and $+30°$ and sometimes severe in-plane rotation), varying expressions, and high levels of occlusion (the average level of occlusion, *i.e.*, the number of landmarks labeled as occluded as a percentage of the total number of landmarks, of faces, due to hats, sunglasses, food, *etc.*, in the dataset is 23%). The dataset was mainly proposed to push the boundaries of occlusion tolerance by facial alignment algorithms and thus also provides occlusion labels for each landmark. All algorithms were evaluated on the 507 images in the test set partition of the dataset.

3.2.3 Benchmarking our Approach

We compared the fitting accuracy of our approach against that of various existing state-of-the-art methods (an overview of these methods has been provided in section 2) on the previously mentioned datasets. The approaches we compared our approach (denoted/abbreviated as Ours in figures and tables in this thesis) against were those proposed by Tzimiropoulos and Pantic [75] (denoted/abbreviated as AAM-Wild in this thesis), Zhu and Ramanan [6] (denoted/abbreviated as TSMs in this thesis) using their pre-trained and best performing *Independent-1050* model, Yu *et al.* [16] (denoted/abbreviated as CDSM in this thesis), Asthana *et al.* [95] (denoted/abbreviated as DRMF in this thesis), Xiong and De la Torre [96], (denoted/abbreviated as SDM in this thesis), and Burgos-Artizzu *et al.* [11] (denoted/abbreviated as RCPR in this thesis). These approaches were chosen because of their wide use in literature for benchmarking purposes, use of similar training data and landmark annotation schemes to ours, and availability of open source code implementations - AAM-Wild [140], TSMs [7], CDSM [141], DRMF [142], SDM [143], and RCRPR [12]. In addition to this, the algorithms each use different approaches to facial alignment that covered some of the broad categories of approaches that we described in chapter 2, such as AAMs, CLMs, regression based approaches, *etc.*.

It is to be noted that while some of the approaches (TSMs, DRMF, and CDSM) were trained on MPIE database images, similar to the images our approach was trained on, some of the other approaches were at an advantage as they were trained on more unconstrained real-world data. For example, the AAM-Wild approach was trained on the training set partition of the LFPW dataset, RCPR was trained on the training set partitions of the LFPW and COFW datasets (to ensure the best results on the challenging test set partition of the COFW dataset), and the SDM implementa-

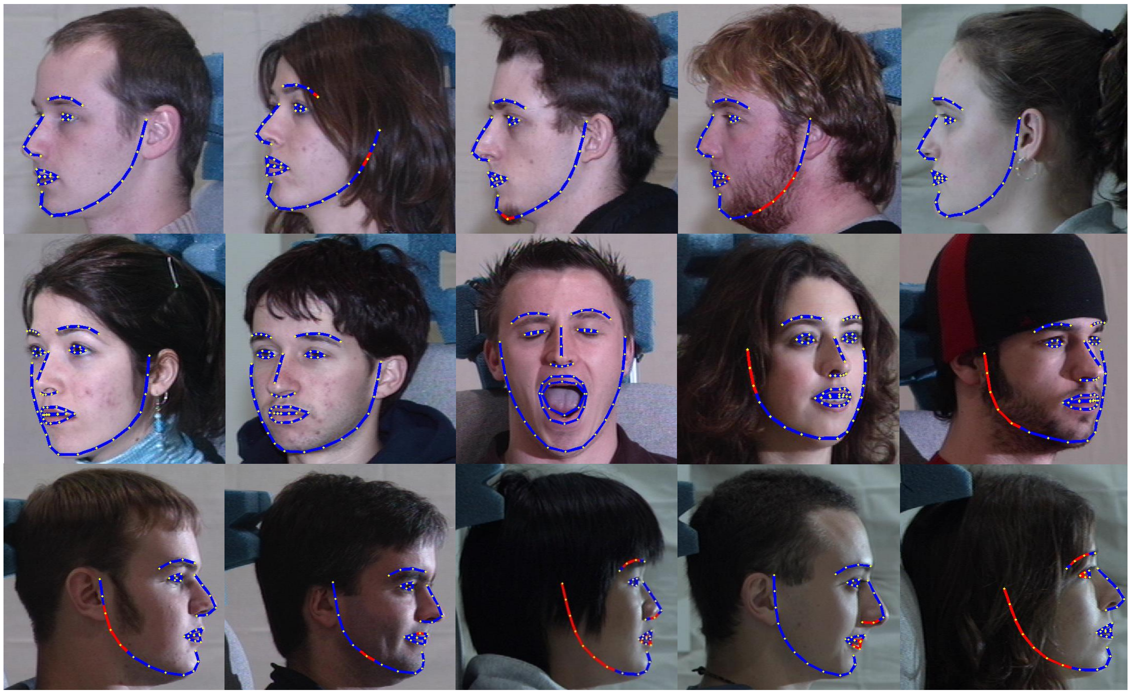


Figure 3.8: Qualitative landmark localization results produced by our approach (trained on images from the MPIE dataset) on some images from the MPIE dataset.

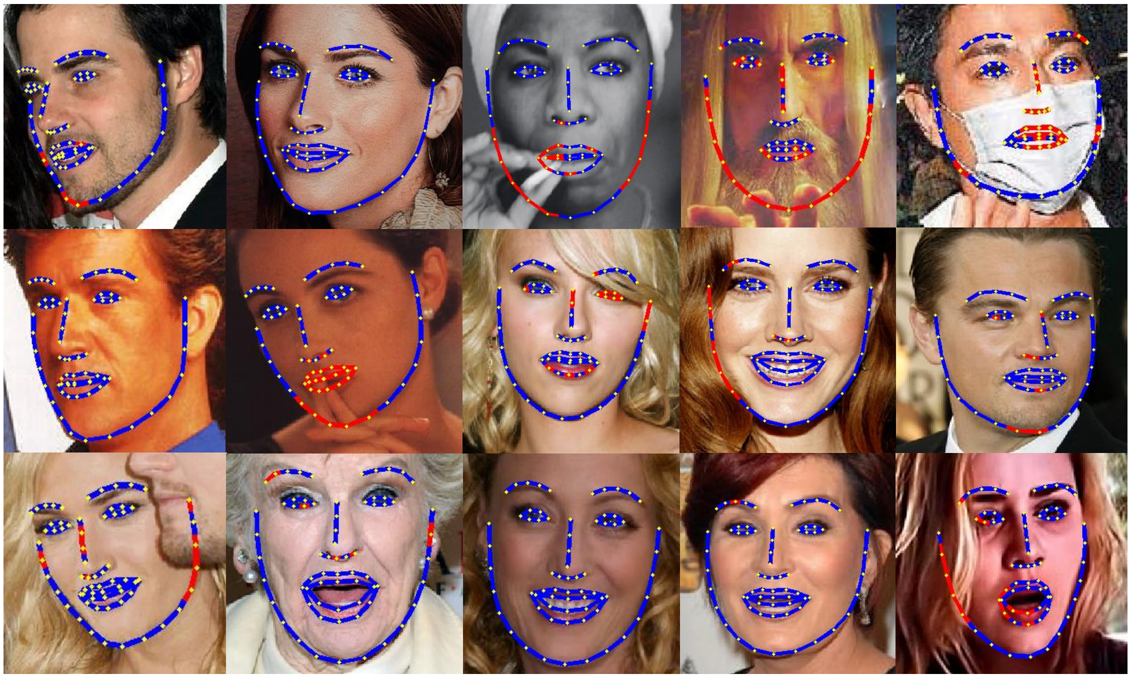


Figure 3.9: Qualitative landmark localization results produced by our approach (trained on images from the MPIE dataset) on some images from the LFPW dataset.

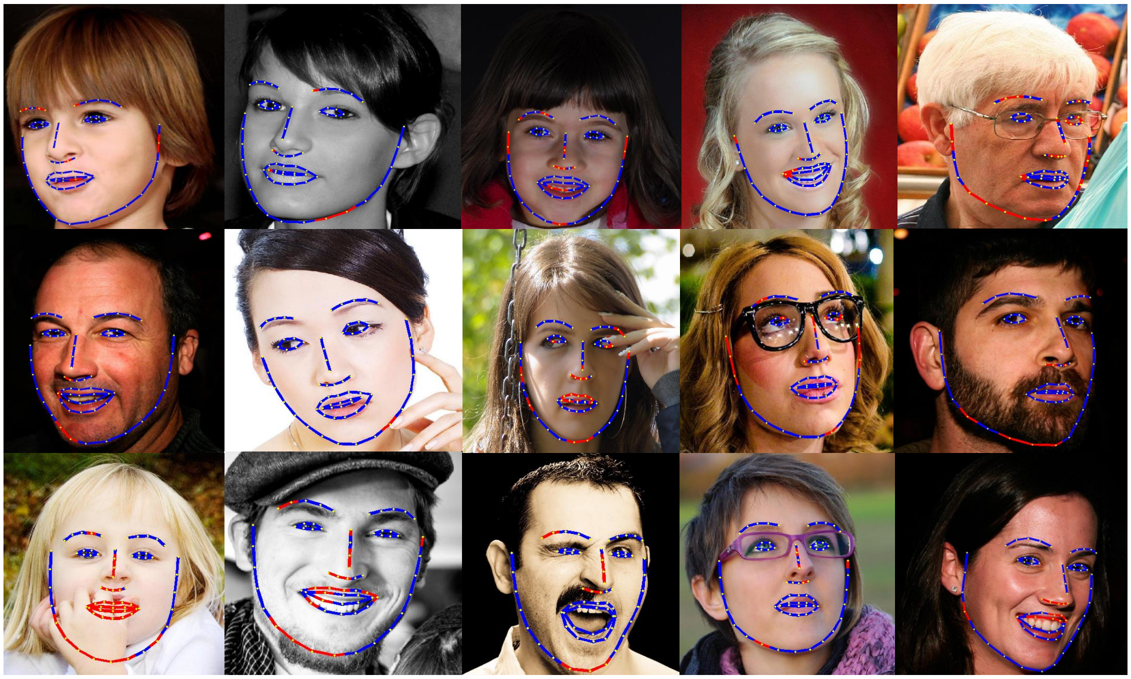


Figure 3.10: Qualitative landmark localization results produced by our approach (trained on images from the MPIE dataset) on some images from the Helen dataset.

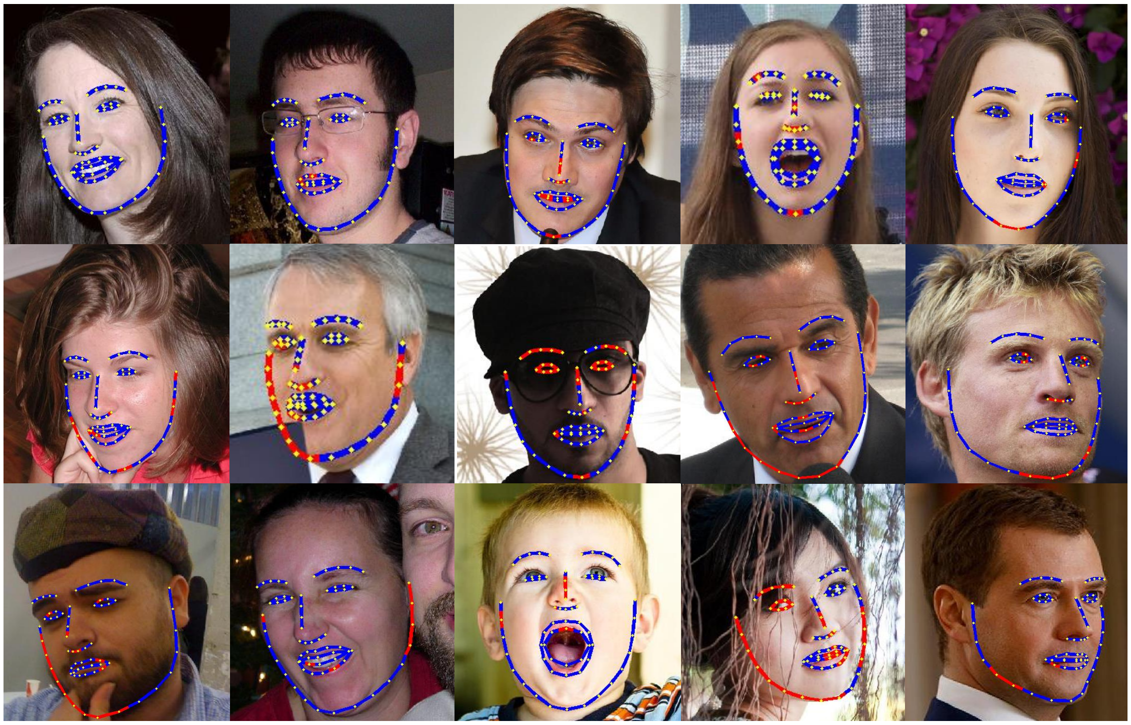


Figure 3.11: Qualitative landmark localization results produced by our approach (trained on images from the MPIE dataset) on some images from the AFW dataset.

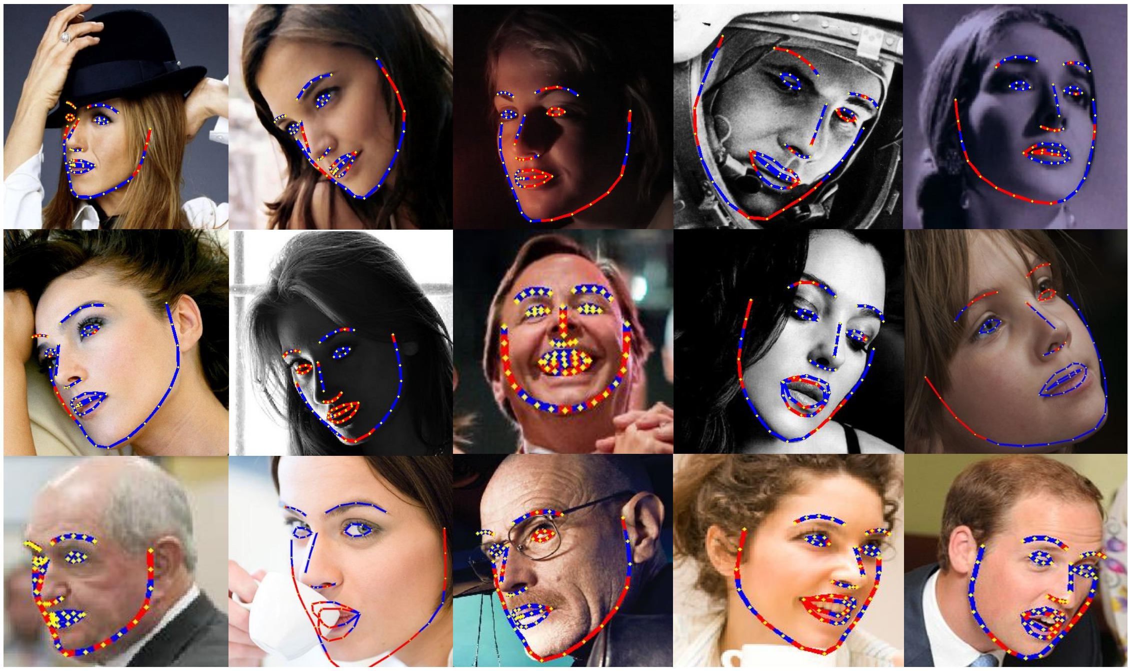


Figure 3.12: Qualitative landmark localization results produced by our approach (trained on images from the MPIE dataset) on some images from the ibug dataset.

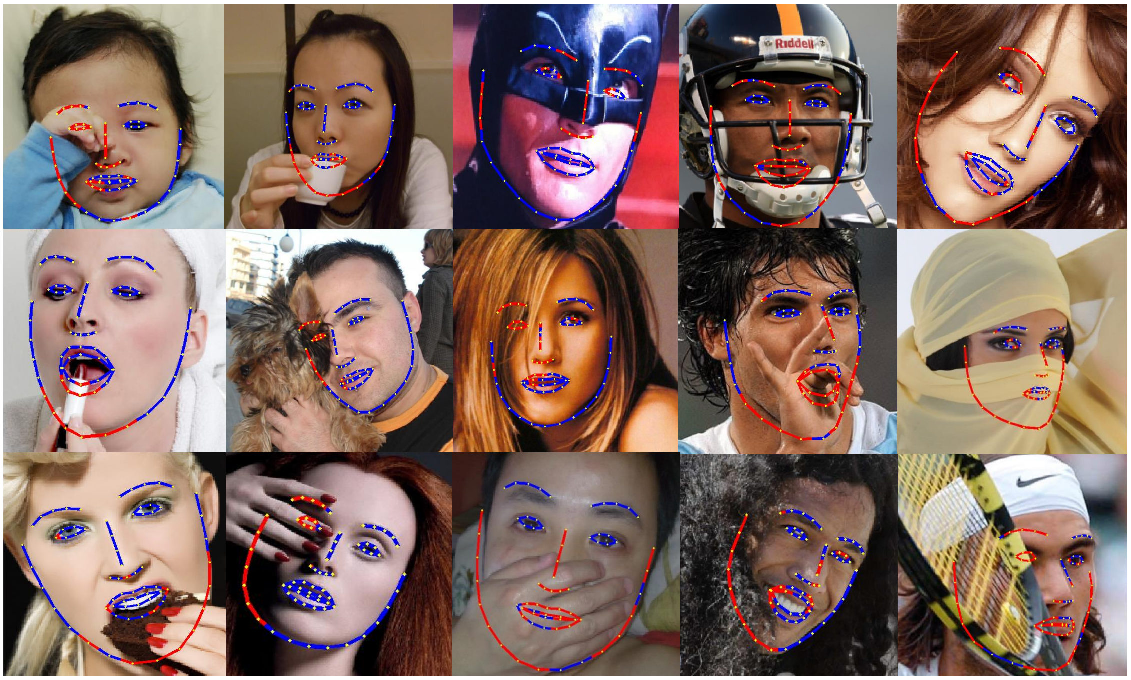
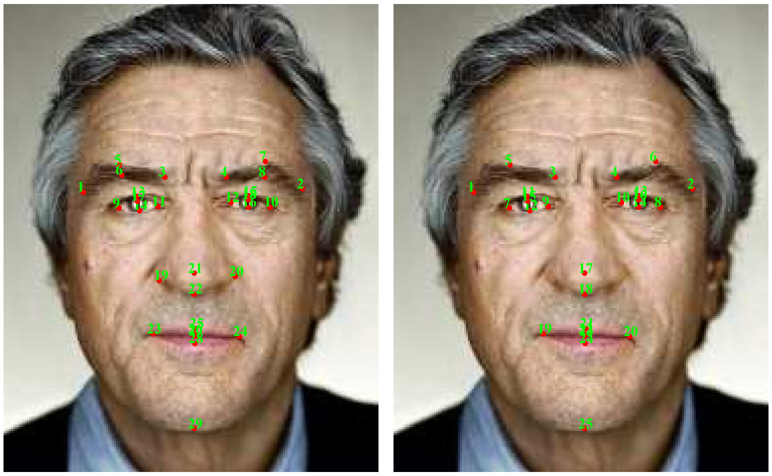


Figure 3.13: Qualitative landmark localization results produced by our approach (trained on images from the MPIE dataset) on some images from the COFW dataset.

tion we used was trained on images from the MPIE and the Labeled Faces in the Wild (LFW) [59], [60] databases. Thus, in order to perform a fair comparison and to demonstrate the importance of training data when dealing with real-world images, we report results obtained by our approach when it was trained using the MPIE images we previously mentioned, the 811 images in the training set partition of the LFPW dataset (when fitting test images from the LFPW, Helen, AFW, and ibug datasets), and the 845 LFPW and 500 COFW training set images with the 29 landmarks and occlusion labels that RCPR was trained on (when testing on the COFW test set images). Our models were built in the same fashion as previously described (using clustering of images into appropriate pose and expression groups) with appropriate changes to account for a different set of landmarks and a lack of images to model absolute yaw variation in excess of 45°. We also report results obtained by training the RCPR algorithm (trained using the optimal parameter values specified by the authors for training on un-occluded images) on the same set of MPIE images (images with yaw variation between -45° and $+45^\circ$ and 68 ground truth annotations) as our approach and using the 68 point landmarking scheme. As we will show, our approach performs admirably when trained on only MPIE images and provides much higher accuracy levels than all the other approaches when trained on the real-world LFPW and COFW training set images. Some results obtained using our approach (when trained on MPIE images only) on images from the various datasets are shown in Figures 3.8 - 3.13. As can be seen, our approach is able to generalize well to accurately localize facial landmarks in these unseen images and is quite tolerant to the challenging pose, occlusion, and expression variations in the images in spite of being trained on images that do not contain these variations.

The other key aspect to consider when reporting a fair performance comparison of facial alignment algorithms is initialization (facial bounding boxes provided as input). The DRMF and SDM algorithms are extremely sensitive to these inputs and it was observed that they produced extremely poor results when we used the bounding box initializations for the LFPW, Helen, AFW, and ibug datasets [9] provided by the organizers of the 300-W 2013 challenge that were obtained using what was referred to as their "in-house face detector." Thus, for these approaches we used an OpenCV implementation of the Viola-Jones [144] face detection algorithm to provide bounding box initializations for these datasets whenever the face detection results produced were in close agreement with the provided bounding boxes and by reverting to suitably modified versions (square bounding boxes) of the provided bounding boxes whenever spurious/no detections were made by the OpenCV face detector. For initializing the RCPR algorithm (trained on MPIE images), that is also sensitive to bounding boxes provided, on these three datasets, we used the same process as during training and provided bounding boxes that were crops around then ground truth landmark locations grown by 15%. For the AAM-Wild algorithm, our approach, CDSM, and the TSMs approach, we used the used the bounding box initializations provided by the organizers of the 300-W 2013 challenge with the crop grown by a factor of 1.5 to enclose the facial region in the latter three cases. This was carried out because though the CDSM and TSMs approach function as face detectors, a fair comparison based on landmark localization accuracy demands that an appropriate region of interest be provided. Our approach does not detect faces and can fail in the event of extremely poor face detection results, however we did not train our approach by assuming specific details about the bounding boxes available during testing. Thus, specifying a large region of interest for our initial seed landmark detection stage is sufficient to deal with slight scale and translation differences in face detection bounding boxes and we were thus in a position to use bounding boxes that we had no information about during our training stage.

For the COFW test set we observed that the facial bounding boxes provided along with the ground truth landmarks were generally not optimal for many of the alignment algorithms (DRMF, AAM-Wild, and SDM) as they only partially enclosed the facial region in many cases. In order to ensure better initialization, a square crop was generated to enclose the ground truth landmark locations and then grown by 15% before being provided as initialization (for the DRMF and SDM algorithms, these bounding boxes were only used when OpenCV implementation of the Viola-Jones face detection algorithm could not provide accurate bounding boxes). For CDSM, TSMs,



(a) (b)

Figure 3.14: (a) The 29 point landmarking scheme for the COFW dataset and (b) The 25 landmarks common to both the 68 point MPIE landmarking scheme and the ground truth annotations available for the COFW dataset. The facial image in the figure is from the training set partition of the COFW dataset. (a) and (b) appear in [13].

and our approach, this crop region was further expanded by a factor of 1.5 to enclose the facial region and yet not provide any initialization advantage. The same initialization protocol that was used when testing RCPR (trained on MPIE images) on the other test sets was used when it was run on the COFW test set images. We also evaluated our approach (trained in identical fashion to the RCPR approach on the same set of LFPW and COFW images with 29 landmarks) and RCPR using the bounding boxes provided along with the COFW test set images for initialization. Finally, for the MPIE dataset, a square crop around the convex hull of the ground truth landmark locations was extracted and then grown by a factor of 1.5 before being provided as initialization to TSMs and our approach.

Most of the approaches use the same 68 point landmarking scheme to annotate frontal facial images, making a fair comparison possible on the LFPW, Helen, AFW, and ibug datasets. However, the SDM algorithm localizes 49 facial landmarks (does not localize landmarks $1-17$ (facial boundary points) and landmarks 61 and 65 (interior points near the corners of the mouth) in Figure 3.2 (b). Thus, our results are reported for both these cases and by utilizing the maximum possible common landmarks localized by the various algorithms. For the COFW dataset, where only 29 manually annotated landmarks are available, we measured the landmark localization accuracy of the algorithms using a set of 25 landmarks (and 24 for the SDM method) which are common to both the 29 point and 68 point markups. This set of landmarks is shown in Figure 3.14 (b). However, we also provide results that compare our approach against the RCPR approach on the full set of 29 landmarks, shown in 3.14 (a). The MPIE dataset is the only one that contains facial images that exhibit an absolute yaw in excess of $45°$ and is hence used to only compare our algorithm against the TSMs approach, as none of the other algorithms provide localization results using the same 39 point landmarking scheme that are shown in Figure 3.2 (a) or handle such yaw variation.

To compare the landmark localization accuracies of the various algorithms, the fitting error (the Euclidean distance between the automatically fitted landmarks and their corresponding manually annotated ground truth locations) was normalized for each image using the distance between the outer corners of the eyes (landmarks 37 and 46 for the frontal landmarking scheme in Figure 3.2 (b), landmarks 9 and 10 in Figure 3.14 (a), and landmarks 7 and 8 in Figure 3.14 (b)) in the ground truth annotations, as was carried out in the 300-W 2013 challenge, to enable a fair comparison across all images (of varying resolution and facial sizes) in the datasets. For the MPIE dataset the average eye center to mouth corner distance was used for normalization as this dataset contained images with profile views. These distances were averaged over all landmarks to produce a normalized fitting error for each image in the dataset. The Mean Normalized Fitting Error (MNFE) of these fitting errors, calculated by averaging the normalized fitting error over all images in the test dataset and expressed as a percentage, is the common metric commonly employed to determine the accuracy of a facial alignment algorithm. Another metric that is used to compare the approaches is the percentage of failures. This is computed as the percentage of the total images fitted that have a normalized fitting error value of over 10%, a measure that was proposed in [86]. These same metrics are used when reporting results for the various test sets in question in future sections of this chapter and future chapters as well.

Table 3.2: Performance of various algorithms on various test sets with MNFE values computed using 68 (or 39 for the MPIE test set) common landmarks, except in cases where an alternative number of landmarks (indicated in brackets) were used.

Algorithm	Test Set					
	MPIE	LFPW	Helen	AFW	ibug	COFW
	MNFE (%)	MNFE (%)	MNFE (%)	MNFE (%)	MNFE (%)	MNFE (%)
Ours (Best Models)	___	**4.98**	**5.46**	**7.10**	**9.95**	**6.00 (29)**
Ours (MPIE Tr Set)	5.37	6.68	7.47	8.79	13.18	8.53 (25)
DRMF	___	6.77	8.97	11.81	19.40	10.32 (25)
CDSM	___	7.63	10.08	10.30	19.57	9.73 (25)
TSMs	6.68	8.99	8.47	10.72	25.46	9.58 (25)
RCPR (COFW + LFPW Tr Sets)	___	___	___	___	___	6.16 (29)
RCPR (MPIE Tr Set)	___	8.10	9.87	12.54	20.14	12.47 (25)
AAM-Wild	___	12.41	12.81	17.75	27.88	12.03 (25)

Table 3.3: Performance of various algorithms on various test sets with failure percentages computed using 68 (or 39 for the MPIE test set) common landmarks, except in cases where an alternative number of landmarks (indicated in brackets) were used.

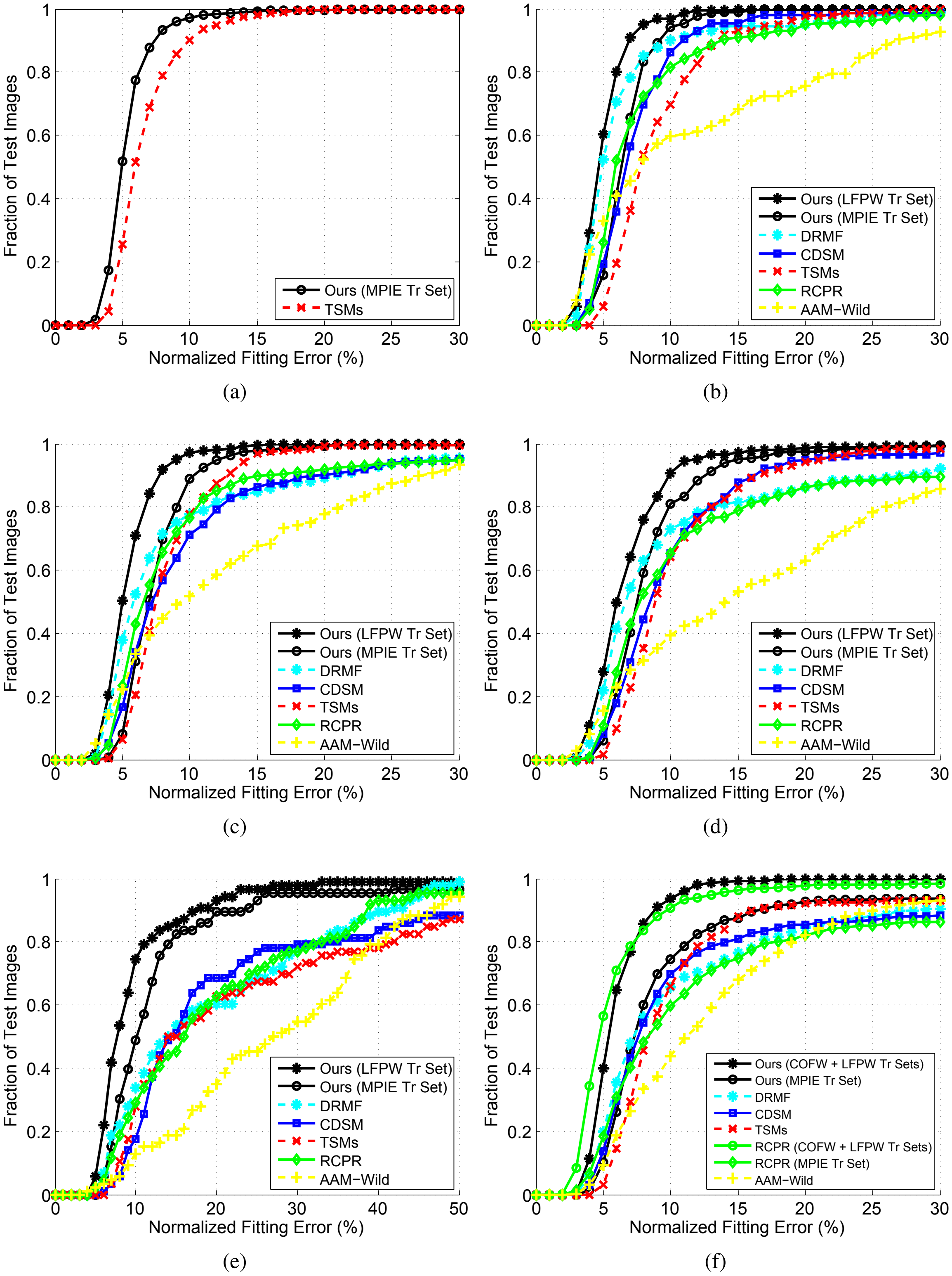
Algorithm	Test Set					
	MPIE	LFPW	Helen	AFW	ibug	COFW
	Failure (%)	Failure (%)	Failure (%)	Failure (%)	Failure (%)	Failure (%)
Ours (Best Models)	___	**3.2**	**2.9**	**9.4**	**25.6**	**6.3 (29)**
Ours (MPIE Tr Set)	2.9	5.9	11.2	19.2	51.2	20.9 (25)
DRMF	___	10.0	22.4	27.0	66.3	30.3 (25)
CDSM	___	13.2	27.4	34.1	81.5	23.8 (25)
TSMs	9.8	30.3	22.4	35.7	72.1	30.1 (25)
RCPR (COFW + LFPW Tr Sets)	___	___	___	___	___	9.3 (29)
RCPR (MPIE Tr Set)	___	18.6	23.3	34.7	70.9	36.6 (25)
AAM-Wild	___	40.3	48.2	60.7	87.2	53.6 (25)

Table 3.4: Performance of various algorithms on test sets with MNFE values computed using 24 (for the COFW test set), 49 or 27 (for the MPIE test set), and 49 (for all other test sets) common interior landmarks localized by the various algorithms.

Algorithm	Test Set					
	MPIE	LFPW	Helen	AFW	ibug	COFW
	MNFE (%)	MNFE (%)	MNFE (%)	MNFE (%)	MNFE (%)	MNFE (%)
Ours (Best Models)	—	**4.19**	**4.59**	**6.00**	**8.34**	**5.90**
Ours (MPIE Tr Set)	4.38	6.09	6.71	8.02	11.97	8.34
SDM	—	5.00	6.60	9.83	15.90	6.65
DRMF	—	5.92	7.71	10.68	16.80	10.14
CDSM	—	6.10	8.22	8.40	16.78	9.49
TSMs	5.08	7.30	7.00	9.19	23.11	9.27
RCPR (COFW + LFPW Tr Sets)	—	—	—	—	—	6.00
RCPR (MPIE Tr Set)	—	7.79	9.77	12.67	20.11	12.33
AAM-Wild	—	12.07	12.36	17.80	28.42	11.57

Table 3.5: Performance of various algorithms on test sets with failure percentages computed using 24 (for the COFW test set), 49 or 27 (for the MPIE test set), and 49 (for all other test sets) common interior landmarks localized by the various algorithms.

Algorithm	Test Set					
	MPIE	LFPW	Helen	AFW	ibug	COFW
	Failure (%)	Failure (%)	Failure (%)	Failure (%)	Failure (%)	Failure (%)
Ours (Best Models)	—	**0.9**	**1.6**	**5.2**	**19.8**	**5.3**
Ours (MPIE Tr Set)	1.3	3.6	7.3	14.9	36.1	19.9
SDM	—	4.1	12.1	15.3	36.1	11.9
DRMF	—	8.1	17.3	18.5	54.7	29.7
CDSM	—	6.4	16.6	19.0	54.3	22.3
TSMs	2.6	14.5	10.5	22.7	58.1	27.6
RCPR (COFW + LFPW Tr Sets)	—	—	—	—	—	8.7
RCPR (MPIE Tr Set)	—	16.3	18.9	31.2	65.1	36.0
AAM-Wild	—	39.9	44.7	58.1	83.7	50.8



(a)

(b)

(c)

(d)

(e)

(f)

Figure 3.15: Cumulative Error Distribution (CED) curves for various algorithms obtained by averaging the normalized fitting errors (%) over all common landmarks (29 or 25 for the COFW test set, 68 or 39 for the MPIE test set, and 68 for all other test sets) on the (a) MPIE, (b) LFPW, (c) Helen, (d) AFW, (e) ibug, and (f) COFW test sets.

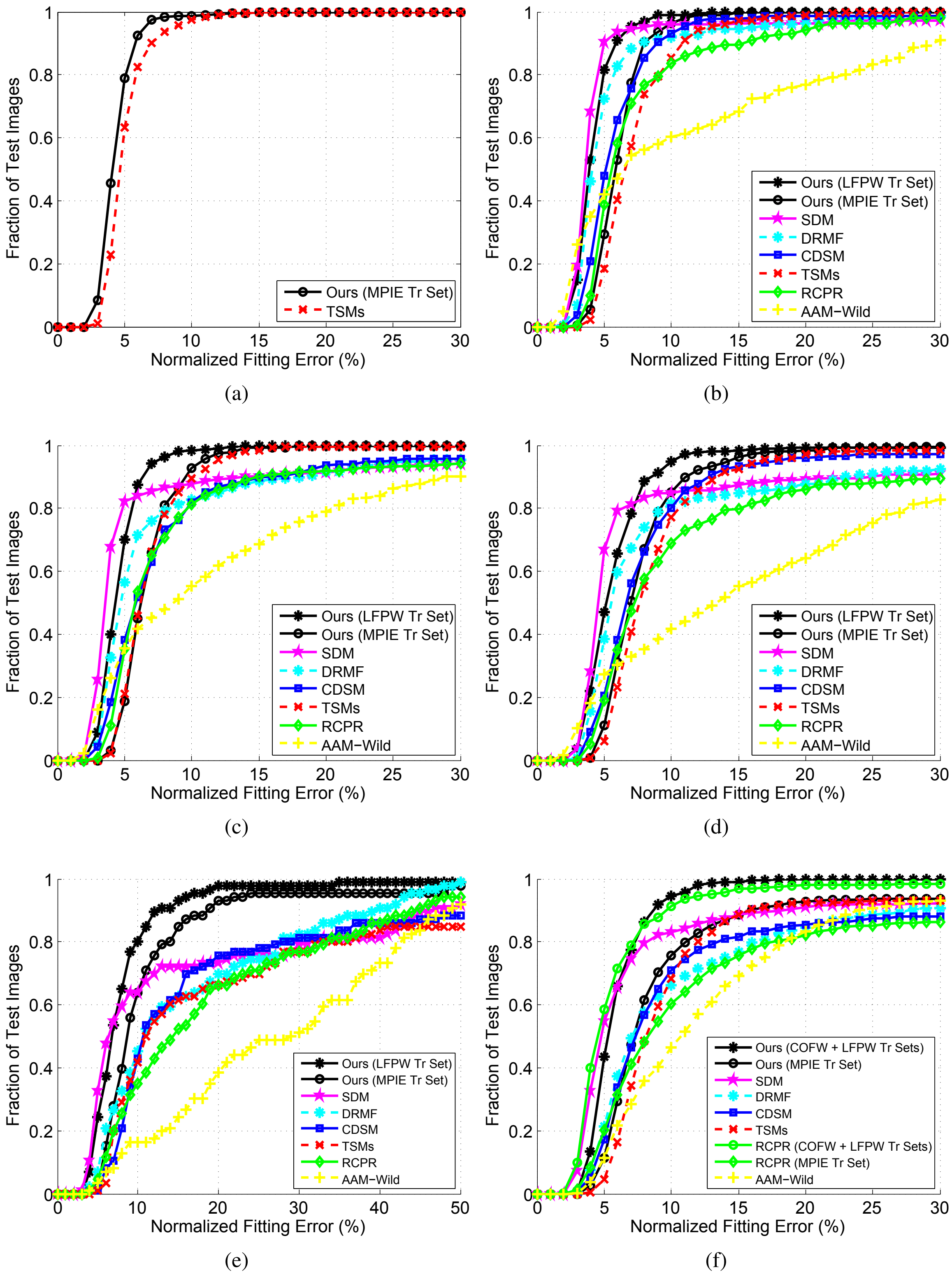
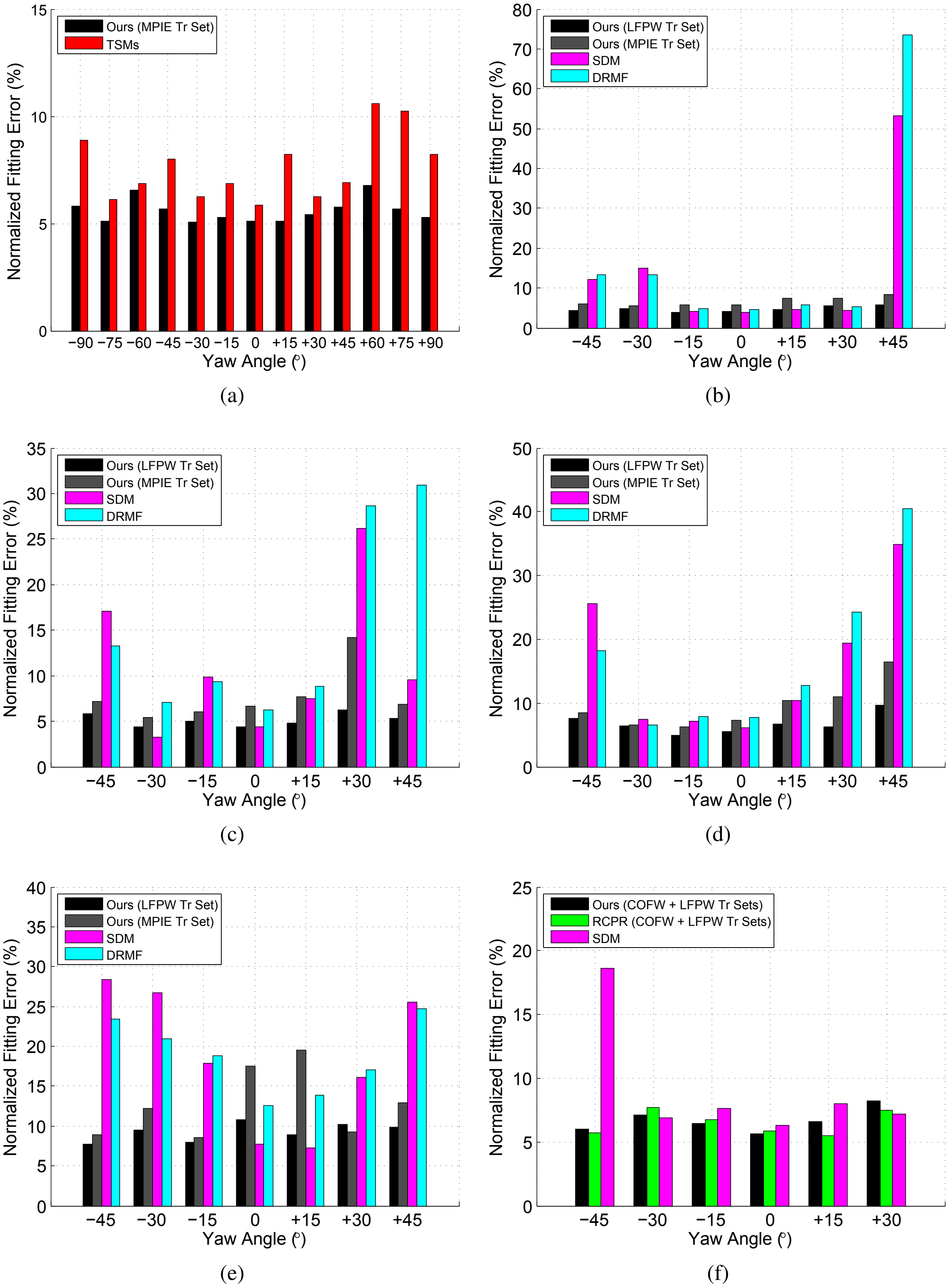


Figure 3.16: Cumulative Error Distribution (CED) curves for various algorithms obtained by averaging the normalized fitting errors (%) over common interior landmarks (24 for the COFW test set, 49 or 27 for the MPIE test set, and 49 for all other test sets) on the (a) MPIE, (b) LFPW, (c) Helen, (d) AFW, (e) ibug, and (f) COFW test sets.

Due to a lack of correspondence between landmarks in the 68 and 39 point landmarking schemes, we report results over (average over) only those images where the TSMs method determined a set of 68 landmarks. Table 3.2 and Table 3.3 respectively list the MNFE and failure percentage values obtained by the various approaches on the different test sets over the largest number (most commonly 68 landmarks) of common landmarks while Table 3.4 and Table 3.5 list the same values when only interior facial landmarks (mostly commonly 49 landmarks) are considered and also has results obtained by the SDM algorithm implementation, which does not localize landmarks along the facial boundary. For the COFW dataset case, this corresponded to the exclusion of just the tip of the chin. Predictably, all methods demonstrated a higher accuracy when localizing only the interior landmarks. As can be seen from the tables, our approach performed quite well even when trained only on images from the MPIE database. However, the best performance (indicated by Best Models) for our approach was achieved on the LFPW, Helen, AFW, and ibug test sets when trained on the LFPW training set images and on the COFW test set when trained on the LFPW and COFW training set images (in a similar fashion to the RCPR algorithm). Our best performing models provided more accurate results than the other algorithms on all the test sets and this serves to demonstrate the efficacy of our approach as well as the role the training set plays when reporting such accuracy rates.

An alternative way of comparing the accuracy of the methods is using Cumulative Error Distribution (CED) curves that plot the fraction of facial images (plotted along the y-axis) found to have a normalized fitting error (%) value lower than a certain value (plotted along the x-axis). CED curves summarizing the performance of the various methods on the various datasets are shown in Figure 3.15 and Figure 3.16 using the same number of landmarks as those in Table 3.2 and Table 3.4, respectively. From these figures it is again clear that our approach (best performing models) localized landmarks more accurately than the other algorithms on all the test sets.

We also determined how pose, expression, and occlusion factors influenced the top performing algorithms on each test set. Figure 3.17 shows how the normalized fitting error (%) values (with



(a)

(b)

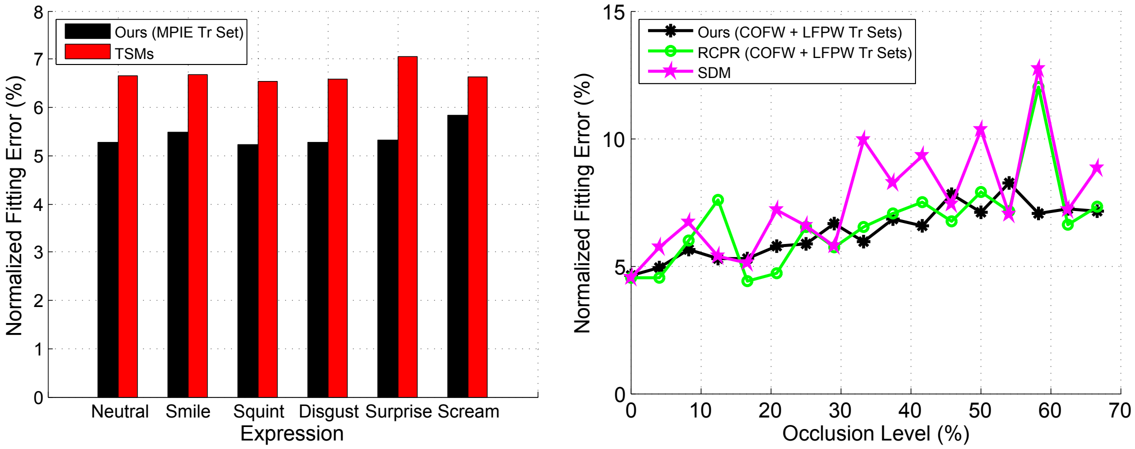
(c)

(d)

(e)

(f)

Figure 3.17: Normalized fitting errors (%) as a function of yaw for the top performing algorithms calculated using a common set of landmarks (24 for the COFW test set, 68 or 39 for the MPIE test set, and 49 for all other test sets) on the (a) MPIE, (b) LFPW, (c) Helen, (d) AFW, (e) ibug, and (f) COFW test sets. (b), (d), (e), and (f) appear in [13].



(a) (b)

Figure 3.18: Normalized fitting errors (%) obtained (using a common set of 68 or 39 landmarks on the MPIE test set and 24 landmarks on the COFW test set) using various algorithms on faces with various expressions and occlusion levels on the (a) MPIE and (b) COFW test sets, respectively. (b) appears in [13].

all images in the test sets considered) vary as a function of the facial yaw angle for the various algorithms. The images in each test set were clustered into various bins (by comparing the ground truth landmarks to those of images in the MPIE database) and the average of the normalized fitting errors for all the images belonging to that yaw bin were calculated and plotted as a function of yaw. In similar fashion, graphs were obtained to determine how the fitting errors varied by expression on the MPIE test set (see Figure 3.18 (a)), for which expression labels were available, and as a function of level of occlusion (% of landmarks labeled as occluded out of the full set of landmarks) on the COFW test set (see Figure 3.18 (b)). As can be seen in Figure 3.17, our approach (best models or models trained using MPIE images) provided a consistent level of performance across the various yaw angles and demonstrated a higher tolerance to faces with an absolute yaw in excess of 30° than SDM and DRMF. This tolerance to pose is particularly evident on the AFW and ibug datasets that contain a larger number of images with high yaw compared to the LFPW, Helen, and COFW test sets. Similarly, Figure 3.18 (a) shows how our approach provided consistent results on the MPIE test set for varying expressions, while the same point is made by Figure 3.18 (b) for the

Table 3.6: Occlusion prediction performance of RCPR and our algorithm (both trained using COFW and LFPW training set images) when localizing 29 landmarks on the COFW test set. This table appears in [13].

Algorithm	Accuracy (%)	True Positive Rate (%)	False Positive Rate (%)
RCPR (thresh = 0.6100)	80.62	20.32	1.51
RCPR (thresh = 0.2445)	82.88	59.25	10.11
Ours	81.25	52.08	10.11

Table 3.7: Average time required by various algorithms to process a single face in an image. This table appears in [13].

Algorithm	Avg. Fitting Time Per Image (secs)
SDM	$\approx$ 0.10
RCPR	≈ 1
DRMF	≈ 2
AAM-Wild	≈ 4
CDSM	≈ 5
TSMs	≈ 18
Ours	≈ 33

varying level of facial occlusion in the COFW test set.

We also provide details on the occlusion prediction performance (over 29 landmarks) of RCPR and our approach (when both were trained on the same set of LFPW and COFW training set images and provided real valued occlusion labels that had to be thresholded to produce binary occlusion labels) on the COFW test set in Table 3.6. The metrics used in the table are the accuracy ($(TP + TN)/(P + N)$, where TP is the number of true positive detections, TN is the number of true negative detections, P is the total number of positive samples, N is the total number of negative samples), true positive (TP/P), and false positive (FP/N) rates. As can be seen, our approach provided a higher accuracy rate than RCPR when the default (precomputed using the training data) occlusion detection threshold value of 0.61000 was used for RCPR. However, RCPR provided marginally more accurate results for a different threshold value of 0.24445 (determined

Table 3.8: MNFE (%) values obtained on test sets by each stage of our approach (MPIE training set) by averaging over the maximum number of landmarks localized at that stage.

Test Set	Ours (MPIE Tr Set)				
	Stage 1 MNFE (%)	Stage 2 MNFE (%)	Stage 3 MNFE (%)	With Pose Estimation MNFE (%)	Best Result MNFE (%)
MPIE	5.09	9.12	5.11	5.34	4.91
LFPW	7.32	9.17	6.60	6.89	6.06
Helen	7.37	9.77	7.56	7.54	6.77
AFW	9.70	11.57	8.86	9.50	7.80
ibug	13.33	14.92	12.08	11.80	10.04
COFW	10.38	9.83	8.32	8.15	7.07

from a Receiver Operating Characteristic (ROC) curve using the same false positive rate as our approach).

Finally, to complete the benchmarking analysis, we also provide a timing analysis of the various approaches in Table 3.7. The table lists the average time required by the various approaches to fit an image on a desktop computer with an Intel Xeon X5680 processor with a clock rate of 3.33 GHz running Windows 7. While our approach presently requires a larger amount of time to process an image, it is to be noted that this our implementation is currently purely MATLAB [135] based and is not heavily optimized for speed, which is something that we are in the process of addressing.

3.2.4 Stage-by-stage Breakdown of Our Approach

Tables 3.8 and 3.9 show a breakdown of the MNFE (%) values obtained at each stage of our approach across all datasets when fitting errors were averaged across all landmarks localized at each respective stage. Stage 1 corresponds to the seed landmark localization stage (when only 8 landmarks are localized for frontal images and 5 for profile images) and the best localized candidates for each seed point were compared against their respective ground truth locations. Stage 2 corresponds to the dense landmark alignment and optimal shape initialization stage, and finally Stage 3 is the refinement stage. The seed landmarks detection stage provided much higher accuracy rates

Table 3.9: MNFE (%) values obtained on test sets by each stage of our approach (best models) by averaging over the maximum number of landmarks localized at that stage.

Test Set	Ours (Best Models)				
	Stage 1 MNFE (%)	Stage 2 MNFE (%)	Stage 3 MNFE (%)	With Pose Estimation MNFE (%)	Best Result MNFE (%)
MPIE	—	—	—	—	—
LFPW	5.00	7.16	4.84	5.09	4.44
Helen	5.21	7.88	5.31	5.64	4.97
AFW	6.78	9.50	6.55	7.20	6.00
ibug	10.24	13.64	10.03	12.44	8.70
COFW	8.16	7.50	5.83	6.68	5.69

on the test sets with low occlusion levels and was impacted most heavily on the COFW test set, when high occlusion levels resulted in high fitting errors for the occluded seed landmarks. Stage 2 transitions from a sparse set to a dense set of landmarks. However, this stage only provides a course alignment using two seed landmark candidates and is thus substantially improved upon by the refinement stage, that provides the final landmark coordinates as output. The importance of the role of Stage 2 in the alignment pipeline (that we alluded to in section 3.1.2) is easy to understand using Tables 3.8 and 3.9, especially in cases with high levels of facial occlusion and when it is not apparent which seed landmarks have been accurately localized.

Tables 3.8 and 3.9 also list the final stage landmark localization accuracies that were obtained when a rough pose estimate value for the facial images was made available using a commercial face detection algorithm (the Pittsburgh Pattern Recognition (PittPatt) face detection algorithm). All numbers in the tables were obtained by averaging over the set of images in each test set for which the appropriate face was detected and a pose estimate was available. This pose estimate value was binned into one of our M bins and used to select the most appropriate initial dense shape for the refinement stage. As the tables show, our scoring method to select the best final refined shape from among a pool of M' shapes often provided lower MNFE values than those obtained using a single refined shape that was pre-selected based on a pose estimate value. Thus, though the speed of fitting dramatically decreases if such contextual pose information is available,

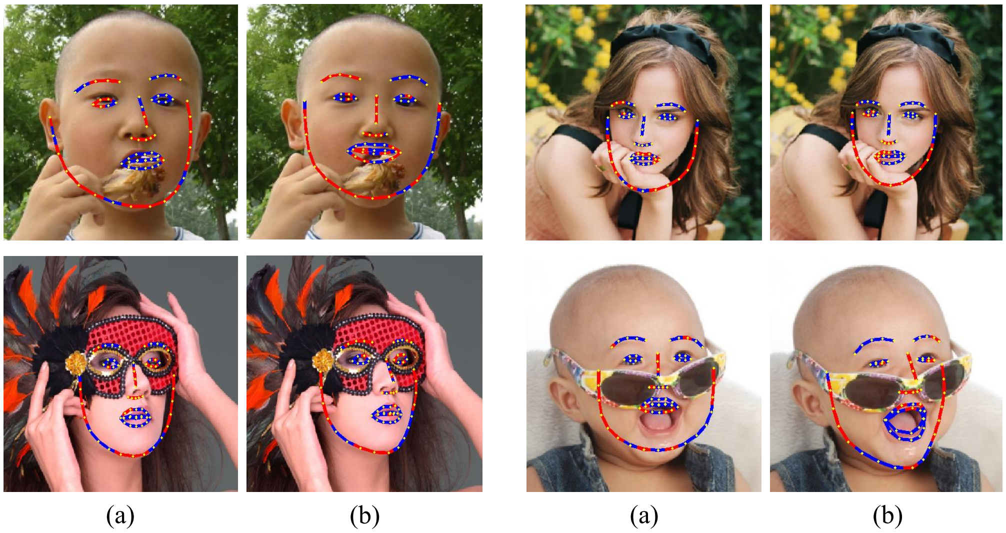


Figure 3.19: Examples from the COFW dataset of where the highest scoring facial shape determined using our approach, shown in columns (a), was not the best fitting shape and could have been replaced with a better fitting facial shape that was not as highly scored, shown in columns (b).

it can lead to a propagation of error if the estimate is inaccurate and incorrect shape and texture models are used from the outset as a result of this estimate.

We have also provided MNFE values for the best case scenario that can be obtained using our approach. These values were obtained by comparing all finally refined shapes (not just the finally picked highest scoring one) against the ground truths and choosing the shape with the lowest fitting error. This indicates that an issue that can arise in our approach is at the final step when a single set of landmarks has to be chosen from among the $M^{'}$ refined shapes and is an area for possible improvement. Figure 3.19 shows some examples where the top ranked shape picked by our method was less accurate than a slightly lower ranked shape that could have been picked. However, it is important to note that since facial alignment is never carried out in isolation and is generally used as a stage in a pipeline for carrying out a subsequent task, such as face recognition, building of 3D facial models, *etc.*, it could also prove advantageous to have multiple facial alignment results available (especially in cases involving high levels of occlusion, where a certain amount of subjectivity is involved in the determination of the "optimal" locations for the landmarks) that can all be used in the subsequent stage so that the best possible result can be obtained using one of these

Table 3.10: Comparison of MNFE (%) values obtained by averaging over 25 landmarks on the COFW test set, 68 or 39 landmarks on the MPIE test set, and 68 landmarks on all other test sets using various shape regularization techniques with models trained on MPIE images.

Test Set	**Ours** **(ℓ_1-Regularized)** **MNFE (%)**	Ours (ℓ_2-Regularized) MNFE (%)	ASM Method (PCA Based) MNFE (%)
MPIE	**5.15**	5.24	5.81
LFPW	**6.36**	6.92	7.24
Helen	**6.80**	7.46	7.69
AFW	**7.06**	7.70	8.01
ibug	**8.03**	8.20	8.64
COFW	**6.69**	7.33	7.50

alignment results or manually selected through visual inspection.

3.2.5 Impact of Shape Regularization

We have already discussed the importance of shape regularization at the final stage of our approach in 3.1.3 and in this section we provide experimental justification for our addition to the shape regularization stage. For each of the previously used datasets, we selected all images with an MNFE (%) lower than 10% (computed using 25 landmarks on the COFW test set, 68 or 39 landmarks on the MPIE test set, and 68 landmarks on all other test sets) and re-fit these images at the final shape refinement stage using the shape regularization technique that has been used in prior ASM implementations and described in section 3.1.3, as well as by using an ℓ_2-regularized approach to solve the problem in equation (3.19), instead of the problem in equation (3.18). Table 3.10 summarizes the results obtained using these different shape regularization techniques and it is evident from the table that our ℓ_1-regularized approach consistently provides more accurate results, albeit at an increased computational cost, than both these approaches and serves to justify why our ℓ_1 based shape fitting approach is an important contribution to the facial landmark localization procedure.

3.2.6 Impact of Initialization on the Performance of Our Approach

We have previously mentioned the importance of the role of initialization (facial bounding boxes provided as input) in the facial alignment process. While our approach is not equipped to perform simultaneous face detection and landmark localization as TSMs and CDSM are, it does exhibit a certain amount of tolerance to differences in the kind of bounding boxes used as input. In order to demonstrate this, we used our approach to localize facial landmarks on images in the previously described AFW and LFPW test sets using 4 different initialization techniques: (1) using a square crop around the manually annotated ground truths, (2) using the bounding boxes provided along with these datasets by the organizers of the 300-W 2013 challenge that were obtained using what was referred to as their "in-house face detector" (the same results were used for benchmarking purposes in section 3.2.3), (3) using the face detection results provided by the face detector described in [17] for which an open source DPM-based implementation is available [145], and finally (4) using the face detection results provided by the Pittsburgh Pattern Recognition (PittPatt) face detection algorithm.

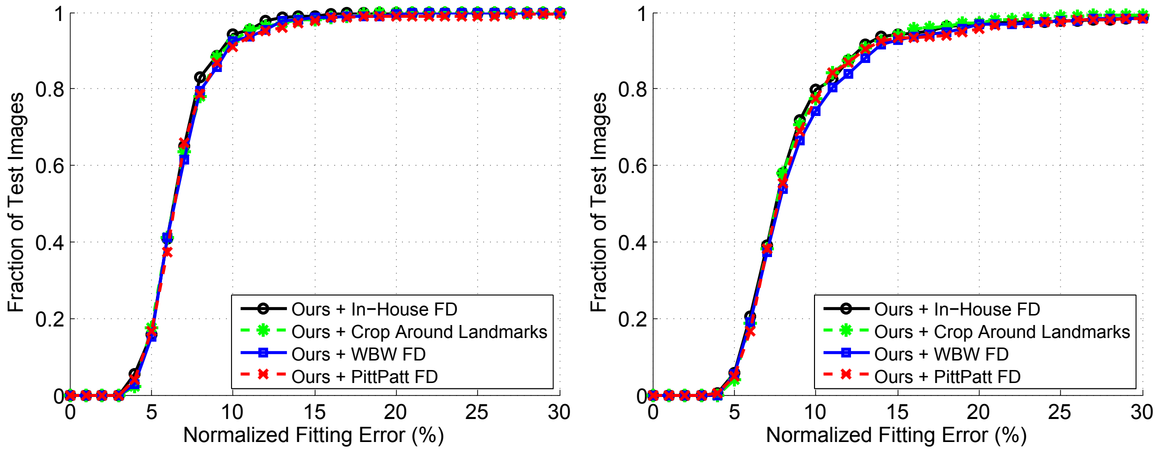
When the experiment was conducted, a large region around the face of interest in each image was cropped using the ground truth landmark annotations prior to bounding box generation by WBW FD and the PittPatt face detector to avoid any ambiguities during the face detection process and to ensure that a bounding box was generated around the "correct" face in the image (the images in the AFW and LFPW test sets often contain more than one face and this approach). A square region was generated using the appropriately translated bounding boxes and scaled by a factor of 1.5 to approximately match the crop generation process during the training stage of our approach. The bounding boxes generated by PittPatt mandated a slight change to the seed landmark search regions used in our implementation. Making this minor change to our code enabled it to deal with varying bounding box initializations. For many alignment algorithms that require face bounding boxes as input, such as DRMF, SDM, and RCPR, a fairly precise transformation (scale and translation parameters) to transform the provided bounding boxes in order to better match

Table 3.11: Comparison of the MNFE (%) and failure % values obtained by our approach (with models trained on MPIE images) by averaging over 68 landmarks on the AFW and LFPW test sets using various initialization techniques (facial region bounding boxes).

Initialization Method	Test Set			
	LFPW		AFW	
	MNFE (%)	Failure (%)	MNFE (%)	Failure (%)
Crop around ground truth landmark locations	6.81	7.6	8.78	22.7
In-House face detector [8], [9], [138], [139]	6.70	5.8	9.07	20.2
Face detector in [17]	6.92	7.6	9.26	25.9
PittPatt face detector	7.25	9.0	9.09	22.7

those used at the training stage would need to be determined and applied at the testing stage in order to ensure their effectiveness when dealing with alternative bounding boxes (compared to the ones used at the training stage). This is accomplished more easily in our approach. While our approach cannot cope with extremely poor initialization (a crop around the face where the face is too small or too large in relation to the crop, resulting in texture signatures that are completely different from those acquired during the training process), a certain amount of tolerance is built in due to the sequence of stages in our approach that are structured to specifically not make too many assumptions regarding the bounding boxes.

Table 3.11 summarizes the results that were obtained using our approach in conjunction with the various initialization methods and CED curves obtained are shown in Figure 3.20. As can be seen from the results, our approach provides fairly consistent results using these different initialization techniques in spite of the fact that only one of them (the crop generation using the ground truth landmark locations) results in crops that are extremely close to those used at the training stage. Beyond a point of course, the problem of poor initialization starts to become the problem of face detection itself and one that a joint framework for accurate face detection and alignment would be best suited to deal with, as discussed in section 7.1.5.



(a) (b)

Figure 3.20: Cumulative Error Distribution (CED) curves obtained using our approach with various different initializations (facial region bounding boxes) on the (a) LFPW and (b) AFW test sets, respectively. The initialization schemes used are: (1) a square crop around the manually annotated ground truths, (2) the bounding boxes provided along with these datasets by the organizers of the 300-W 2013 challenge that were obtained using what was referred to as their "in-house face detector", (3) using the face detection results provided by the face detector described in [17] (denoted by WBW FD), and (4) using the face detection results provided by the Pittsburgh Pattern Recognition (PittPatt) face detection algorithm.

3.3 Summary of Results and Contributions

An understanding of the roles of initialization and shape regularization to the landmark localization process was gained through our initial work on automatic facial landmark localization [55], [56]. This prior work aided in the evolution of the facial alignment approach that we have presented in this chapter. The facial alignment framework we have described is able to jointly deal with the problems posed by facial pose, illumination, and expression variations, and the presence occlusions. We also proposed the use of a shape dictionary and an ℓ_1-regularized LSP based approach for shape regularization that ensured higher accuracy rates than those achieved by previously used shape regularization techniques. Our approach is capable of handling a larger range of pose variation than many existing alignment algorithms and also provides misalignment/occlusion labels for each fitted facial landmark, which is a desirable attribute that is quite uncommon in prevalent work.

We demonstrated the superiority of our approach over several existing state-of-the-art algorithms on challenging real-world datasets and also provided proof of its consistent performance across varying facial pose, expressions, and occlusion levels. Details on the facial alignment approach that we have described in this chapter and the results obtained using it can be found in [13].

It is to be noted that our approach is modular in nature and is built using a few stages that follow in sequence. While we have demonstrated the efficacy of using a specific feature extraction technique, a particular classifier, certain evaluation metrics, and a particular regularization technique, it is quite possible to use substitutes for them within the same framework to achieve acceptable performance.

We now go on to detail results obtained when our alignment approach was used in conjunction with a facial recognition algorithm (in chapter 4), used to enable and aid in analysis of challenging naturalistic driving videos (in chapter 5), and finally, also used to deal with the challenge of localizing landmarks on low-resolution faces (in chapter 6).

Chapter 4

Role of Facial Alignment in Face Recognition

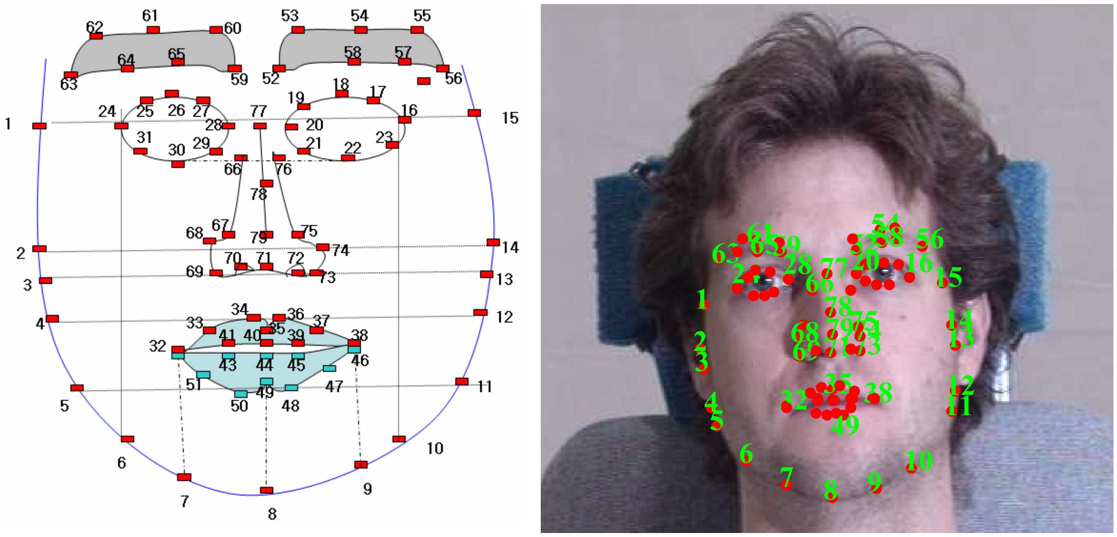
"Who is this guy anyway? Should I pass him on to facial recognition?"

Arlo Glass on 24 - *Day* 8*:* 10 : 00 *P.M.* - 11 : 00 *P.M. (Season* 8 *Episode* 7*)*, and a recurring theme on the show

The field of facial alignment has become extremely important primarily because of its absolute necessity as a pre-processing step for carrying out facial recognition in a completely automated scenario. As the area of face recognition has advanced from dealing with constrained frontal images of subjects exhibiting neutral expressions and acquired under good lighting conditions, to dealing with more unconstrained real-world images, the field of facial alignment has had to grow simultaneously or at an even faster rate in order to lay the groundwork to be able to recognize such faces. For example, most state-of-the-art facial recognition engines can not match profile images of a person (with an absolute yaw in excess of 60°) against a gallery of images with subjects exhibiting a different pose, however, several facial alignment algorithms have already emerged that can deal with such pose variation [6], [13], [16]. In this chapter we provide some context for our work on facial alignment by demonstrating how the results obtained by it are quite acceptable for a face reconstruction and recognition algorithm to perform facial matching on a large database consisting of real-world images. In addition, we also determine how the face recognition accuracies are affected when alignment results obtained by state-of-the-art alignment algorithms (that we benchmarked our approach against in section 3.2.3) are used as input. This serves to demonstrate the impact that facial alignment results can have on face recognition, given a constant facial recognition algorithm that, for all intents and purposes, can be treated as a black box.

4.1 Overview of Facial Recognition Algorithm Used for Evaluation

Before we describe the experimental setup used to evaluate facial alignment results in a face recognition scenario, it is necessary to provide a brief overview of the recognition algorithm used in this context. The facial recognition algorithm used in our experiment was recently developed by Prabhu and Savvides [18] and focused on developing a suitable unified facial representation to deal with factors such as pose variation, partial facial occlusion, varying illumination conditions, and image resolution, rather than the actual face matching process, which it accomplished using previously existing techniques. The central premise of the work was to treat the problem of dealing with real-world face acquisition conditions/degradations (pose, illumination, expression, facial occlusions, and image resolution) as a missing or corrupted data problem in a suitable weighted representation of the face. In this representation, appropriate weights could be assigned to missing or untrustworthy data and thus these portions of the face could either be reconstructed using a massive pre-trained dictionary and used in a texture based matching process, or the coefficients used to obtain the reconstruction could instead be directly used in the matching process, thus negating the need for a synthesized reconstruction. Thus, the three steps in the facial matching process are facial representation, recovery, and matching. We provide a summary of each of these steps using



(a) (b)

Figure 4.1: (a) The 79 point landmarking scheme used by the face recognition algorithm in [18] and (b) The 79 landmarks overlaid on a facial image from the MPIE database.

notation and some technical terms in [18].

4.1.1 Facial Representation

A high resolution 3D facial scan of a subject with a neutral expression exhibiting frontal pose and no illumination artifacts or facial occlusions can be considered to be a complete facial representation with complete textural and structural information. Any image not acquired under these ideal conditions can be considered to be missing certain information. Hence, the idea of addressing the challenges posed by image acquisition conditions or degradations as a missing data problem.

The first assumption made in this representation process is that the coordinates of a specific set (referred to as a sparse set in [18]) of facial landmarks are available (either as manual annotations or through use of an alignment algorithm). The facial landmarking scheme used in [18] is a 79 point scheme that is depicted in Figure 4.1. The same set of landmarks was used in previous publications by us on Modified Active Shape Models (MASM) and their applications [25], [27], [30], [43], [44], [45], [55], [56], [146]. [18] reports extensive results using both manually annotated data as

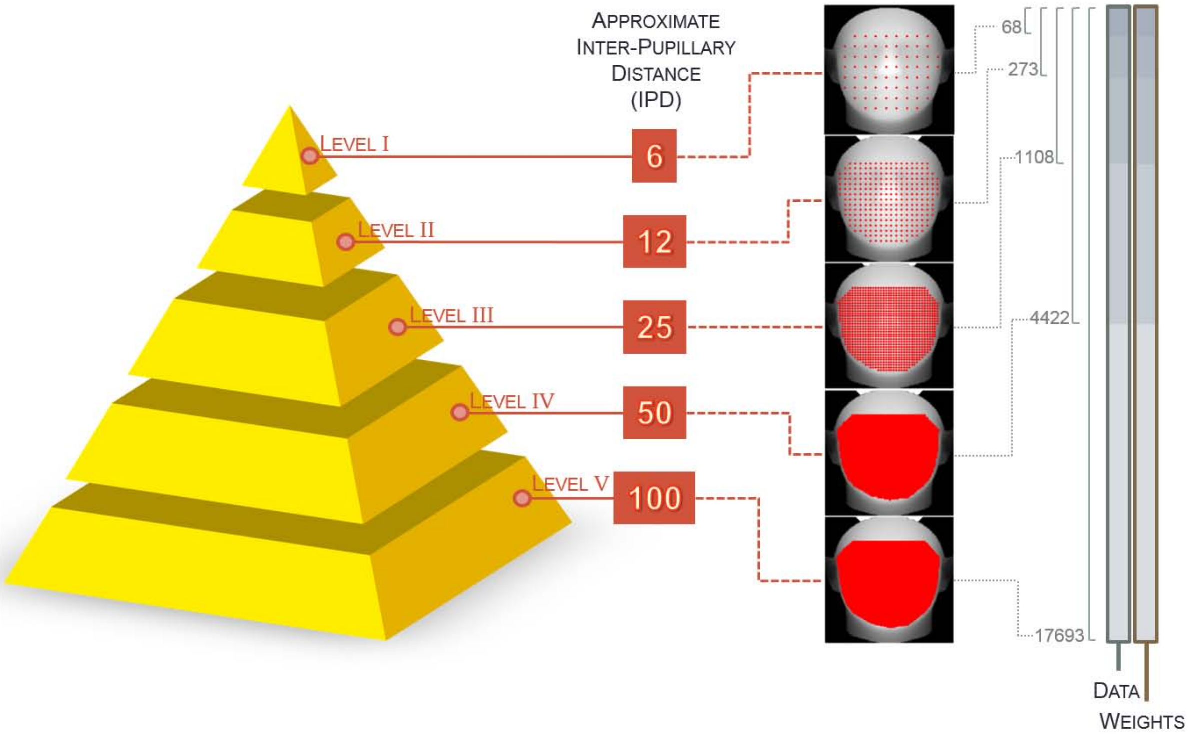


Figure 4.2: The multi-resolution modeling framework used in [18]. The Inter-Pupillary Distance (IPD) of the face determined the grid density to be used to obtain vertices. Vertices that were not sampled were filled with dummy values and set to have a weight of 0. This figure has been reproduced from [18].

well automatically aligned data. A denser set of facial landmark coordinates (mesh densification) was obtained using a Thin-Plate Spline (TPS) [147], [148] based interpolation approach. This TPS based mesh densification improves on the conventionally used Loop Subdivision [149] based densification by allowing for a more uniform distribution of vertices on the face and by ensuring that the coordinates of the initial sparse set of fiducial landmarks are not altered.

The facial representation consisted of a measurement vector $\mathbf{m} = [\mathbf{x}^{\mathrm{T}}\ \mathbf{y}^{\mathrm{T}}\ \mathbf{z}^{\mathrm{T}}\ \mathbf{g}^{\mathrm{T}}]^{\mathrm{T}}$ in which $\mathbf{x}$, $\mathbf{y}$, and $\mathbf{z}$ denote column vectors containing the x, y, and z coordinates of the vertices in the representation, respectively, and $\mathbf{g}$ is a column vector of the corresponding texture indices (grayscale pixel values are directly used in the work in order to represent texture). A weight vector $\mathbf{w} = [\mathbf{w}_x^{\mathrm{T}}\ \mathbf{w}_y^{\mathrm{T}}\ \mathbf{w}_z^{\mathrm{T}}\ \mathbf{w}_g^{\mathrm{T}}]^{\mathrm{T}}$ of the same length as the measurement vector was also maintained and

Table 4.1: Details on the databases used to train the reconstruction system in [18]. This table has been reproduced from [18].

Database	Size	Type	Resolution
USF HumanID 3D Face Database [150]	218	3D	High
Texas 3D Face Recognition Database [151], [152], [153]	$1,149$	3D	High
FRGC v2.0 Database [154], [155]	$34,696$	2D	High
Online Mugshots Database [156]	$1,000$	2D	Medium

contained confidence values (that lie between 0 and 1) of the corresponding observations in $\mathbf{m}$. If a particular measurement was missing due to a degradation, its measurement vector value in $\mathbf{m}$ was set to a dummy value and its confidence value in $\mathbf{w}$ was set to 0. Varying methods (depending on the type, degree, and measurement of the degradation) were used to compute these confidence values. Details on how this confidences were computed can be found in [18].

The final detail that needed to be addressed in the representation was the issue of handling different facial resolutions. The Inter-Pupillary Distance (IPD), computed using the initial sparse set of landmarks, was used to determine facial resolution. For low-resolution images, a pyramid based approach was used to determine the appropriate level of the pyramid that best corresponded to the calculated IPD for the data, which determined the maximum number of entries that could be populated in the measurement vector and the corresponding weight vector. The set of vertices present at a particular pyramid level was a superset of the set of vertices in the pyramid level immediately above it. The entire multi-resolution representation framework used is depicted in Figure 4.2.

4.1.2 Recovery

The powerful representation technique that was previously described was augmented using suitable algorithms that could deal with the weighted data problem to recover the missing measurement vector elements while preserving the confidently measured values. Domain knowledge was preserved in these algorithms by training them on a large set of complete and incomplete faces using unsupervised learning. The training set used in [18] consisted of approximately $37,000$ data items obtained from the USF HumanID 3D Face Database [150], the Texas 3D Face Recognition Database [151], [152], [153], the FRGC v2.0 Database [154], [155], and the Online Mugshots Database [156]. Details on the data available in each database can be found in Table 4.1. All training data was manually annotated according to the 79 point landmark convention and the finally trained representation had a dimensionality of $70,772$ (composed of the grayscale pixel values, x, y, and z coordinates of $17,693$ unique vertices).

Since most of the training data used was incomplete 2D data, the problem that needed to be solved was one of learning from a large amount of high dimensional data but with significant missing data elements, *i.e.*, the goal was to learn a linear basis model $\mathbf{B}$ from the incomplete training data to enable an accurate computation of a set of coefficients $\mathbf{C}$ in order to approximate a data matrix $\mathbf{M} \approx \mathbf{BC}$, assuming a linear model. By improving on the family of techniques referred to by the term Generalized Hebbian Algorithm (GHA) (a stochastic descent algorithm that converges to the principal eigenvectors of unweighted data), [18] proposed a Weighted GHA (GHA) algorithm in order to solve the Weighted Low-Rank Approximation (WLRA) problem at hand. In addition to this approach, an alternative sparsity based recovery technique that built on theory from the K-SVD [157] approach was also proposed. This technique was called the Weighted K-SVD (W-KSVD) approach. The WGHA basis in [18] consisted of 5000 basis vectors while the W-KSVD basis contained 1000 basis vectors. The reader is referred to [18] for further details on the construction of basis vectors using the proposed WGHA and W-KSVD techniques. It must be noted that the W-KSVD approach was the one finally used in computing the basis vectors for reconstruction in our experiment in section 4.2.

4.1.3 Facial Recognition

The final step in the facial recognition pipeline was to obtain matching scores between the gallery and probe images using their respective recovered coefficients. [18] computed these matching scores using three different schemes. The first such scheme used was the Normalized Cosine Distance (NCD) measure. This is a completely unsupervised distance technique that simply computes the distance $D(\mathbf{c}_1, \mathbf{c}_2)$ between 2 sets of coefficient vectors, $\mathbf{c}_1$ and $\mathbf{c}_2$, using equation (4.1).

$$D(\mathbf{c}_1, \mathbf{c}_2) = 1 - \frac{\mathbf{c}_1^{\mathrm{T}}\mathbf{c}_2}{\|\mathbf{c}_1\|\|\mathbf{c}_2\|} \tag{4.1}$$

The other two techniques used to compute similarity scores were the Class-Dependent Feature Analysis (CFA) [158] and Large Margin Nearest Neighbor (LMNN) [159], [160] approaches. However, these two techniques are not unsupervised in nature and were trained using the same training data used for the construction of the basis vectors. The NCD measure (between basis coefficient vectors of the probe and gallery images) was the one that was used in our experiment in section 4.2.

4.2 Impact of Facial Alignment on a Large-Scale Face Recognition Experiment

The previously described face recognition algorithm was used by us in a large-scale face recognition experiment to demonstrate the role that facial alignment plays in a face recognition scenario. The database used in this evaluation was the Labeled Faces in the Wild (LFW) database [59], [60]. The LFW database contains $13,233$ images of $5,749$ individuals (mainly celebrities, such as actors, politicians, and sports personalities), with $1,680$ of these individuals appearing in more than two images. The images are of size 250×250 and consist of roughly centered faces, with a facial region (a tight crop around the facial landmarks) roughly of size 115×115, acquired under real-world conditions with pose (moderate roll variation, yaw variation from $-60°$ to $+60°$, and some slight pitch variation), illumination, and expression variations as well some images with facial occlusions. The database has been extensively used over the last few years for benchmarking various facial recognition algorithms. There are aligned versions of the database as well, such as the funneled version [161], the LFW-a version, which uses an unpublished approach for image alignment, and the deep funneled version [162], that have been observed to produce superior face verification results than the original images for many algorithms. However, for our experiment , we utilized the original database images as the purpose of our experiment was to obtain alignment results using various techniques and then observe the impact that these results had when using a fixed facial recognition algorithm.

Our alignment approach as well as the previously mentioned (see section 3.2.3) AAM-Wild [75], TSMs [6] (using the pre-trained and best performing *Independent-1050* model), CDSM [16], DRMF [95], and RCPR [11] algorithms were again used to localize facial landmarks in all the LFW database images. Our models were trained on images from the LFPW (811 images from the training set partition), Helen (2000 and 330 images from the training set and testing set partitions, respectively), AFW (337 images), and ibug (135 images) datasets in addition to the previously used $6,495$ MPIE training images (see 3.2.1). The same set of images were also used to train the RCPR algorithm (trained using the optimal parameter values specified by the authors for training on un-occluded images), in order to perform a fair comparison against this approach. The other approaches could not be re-trained using these images due to a lack of availability of training code and were thus deployed using their best performing pre-trained models.

As we did in our experiments in section 3.2.3, we provided appropriate initialization to each of these approaches to ensure that the most accurate alignment results could be obtained using them. For CDSM, the TSMs approach, and our approach, a face detection result was not required for initialization as the LFW images are square in aspect ratio and generally contain the subject of interest roughly in the center. Thus, our alignment approach was actually run in an extremely unconstrained fashion compared to some of the other approaches and was still able to localize landmarks with high accuracy, as we will show. The AAM-Wild approach was initialized using a bounding box that was obtained using the convex hull of the landmarks localized by our ap-

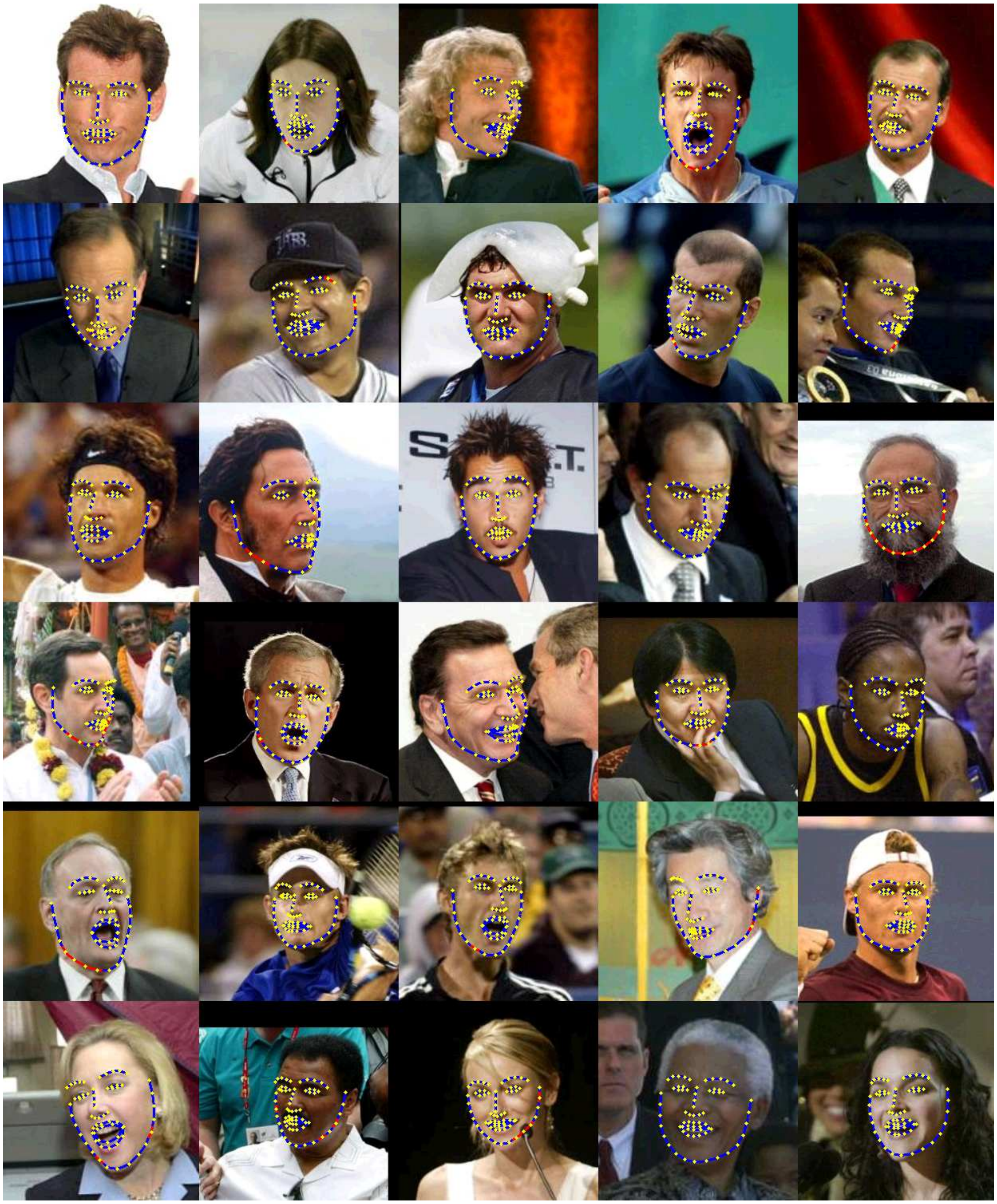


Figure 4.3: Qualitative landmark localization results produced by our approach on images from the LFW database. In all facial images with landmarks overlaid on them, yellow dots are used to indicate the locations of facial landmarks, blue line segments indicate that the landmark at their center is accurately localized, and red line segments indicate that the landmark at their center is misaligned or potentially occluded.

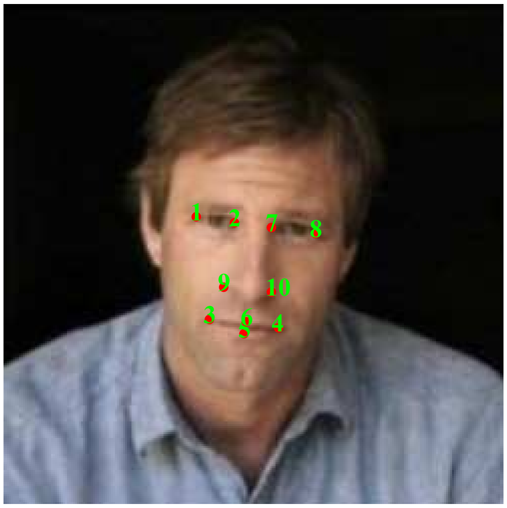


Figure 4.4: An image from the LFW database showing the locations of the 10 landmarks for which manually annotated ground truths are available.

Table 4.2: Performance of various algorithms on the LFW database with MFE, MNFE (%), and failure % values computed using 10 landmarks.

Algorithm	MFE	MNFE (%)	Failure (%)
Ours	**3.96**	**6.50**	**3.69**
RCPR	4.59	7.56	5.15
DRMF	4.57	7.50	6.00
CDSM	5.23	8.61	17.53
TSMs	6.69	11.15	39.99
AAM-Wild	11.35	18.73	78.68

proach as a reference, DRMF was again initialized by using an OpenCV implementation of the Viola-Jones [144] face detection algorithm to provide bounding box initializations, and RCPR was initialized using a bounding box around the convex hull of the landmarks localized by our approach that was grown by 15% to match the bounding boxes that were used during its training process. Results produced by our landmark localization technique on some images from the LFW database are shown in Figure 4.3.

Ground truth (manually annotated) landmark locations for all images in the LFW database have been made available by Dantone *et al.* [163] as part of their recent work on facial landmark

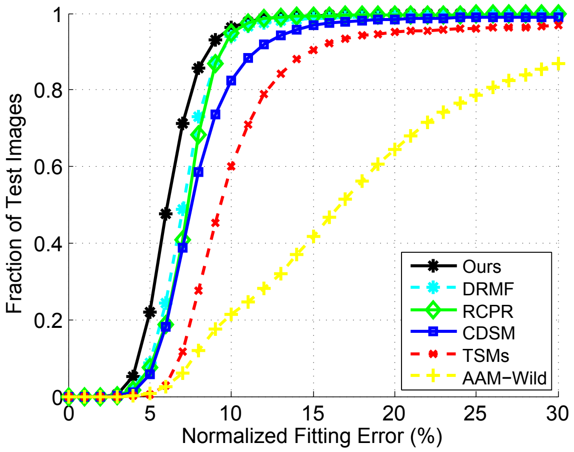
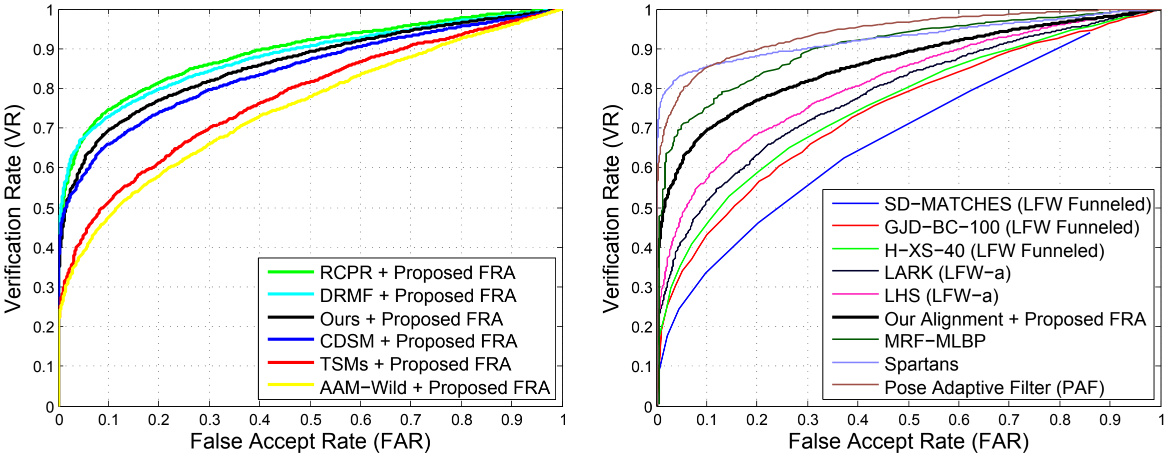


Figure 4.5: Cumulative Error Distribution (CED) curves for various algorithms on the LFW database obtained by averaging normalized fitting errors (%) over 10 common landmarks.

localization [86], [164]. However, only a sparse set of 10 landmarks were manually annotated. The landmarks for which manual annotations are available are shown in Figure 4.4. As in section 3.2.3, the fitting errors (the Euclidean distance between the automatically fitted landmarks and their corresponding manually annotated ground truth locations) produced by each approach were normalized for each image using the distance between the outer corners of the eyes in the ground truth annotations in order to report easily interpretable results that could be compared with those obtained in 3.2.3 and with prior work. However, we also report the un-normalized Mean Fitting Error (MFE) values (averaged over all images and landmarks in each image) as well, since these values can provide some context too, as all images have the same resolution. As was the case in section 3.2.3, we report results over (average over) only those images where the TSMs method determined a set of 68 landmarks. The Mean Normalized Fitting Error (MNFE) of these fitting errors, calculated by averaging the normalized fitting errors over all these images was again employed by us to serve as a metric for comparing the approaches. We also determined the failure percentage (the percentage of the images fitted that have a normalized fitting error value of over 10%) for all approaches. Table 4.2 summarizes the performance of all the alignment approaches using these metrics while Figure 4.5 shows the Cumulative Error Distribution (CED) curves (a plot of the fraction of facial images (plotted along the y-axis) found to have a normalized fitting error (%) value lower than a certain value (plotted along the x-axis)) obtained for the various approaches. As can be seen, the landmark localization accuracy obtained by our approach again surpassed the accuracies obtained by the other approaches. However, since the number of landmarks for which ground truth annotations are available is quite small, these results are presented more to serve as a reference than to evaluate the landmark fitting accuracies of the approaches, which we already carried out in section 3.2.3 in a thorough fashion on datasets specifically created to evaluate landmark localization accuracies.

In order to use these alignment results in conjunction with the previously described facial recognition algorithm, the landmarks localized had to be used to obtain a standard set of 79 landmarks prior to being used as input. A TPS based warping and interpolation algorithm was used to accomplish this transformation from one landmarking scheme to the other. We do not report results obtained using the SDM algorithm because this interpolation technique would place it at a disadvantage compared to the other approaches as it only localizes 49 interior facial landmarks and determining the locations of the landmarks along the facial boundary using these interior landmarks is prone to error. In addition to this, the open source SDM implementation [143] used in section 3.2.3 was trained on images from the LFW database, unlike the other approaches, which were not trained on these images. The additional input that was required for the face recognition algorithm was an estimate of the facial yaw in each image. While some approaches (ours, TSMs, CDSM, and DRMF) can provide coarse or fine estimates of the yaw and roll of annotated faces, some of the other approaches do not provide these estimates (AAM-Wild and RCPR). However, an open source implementation of a pose estimation algorithm that uses a 3D facial shape model that is aligned with the coordinates of 66 (all landmarks in Figure 3.2 (b) except for landmarks 61 and 65) 2D facial landmarks in order to compute an estimate of yaw, pitch, and roll, is available [165]. This algorithm was incorporated into the DRMF facial alignment algorithm by its authors in order to enable the approach to also provide more accurate pose estimates. The use of this algo-



(a)

(b)

Figure 4.6: Receiver Operating Characteristic (ROC) curves obtained using (a) the Face Recognition Algorithm (FRA) proposed in [18] with alignment results produced by various facial alignment algorithms as input and (b) various face existing recognition algorithms.

rithm allowed for a standard evaluation of all alignment approaches when used along with the face recognition algorithm. The appropriate resolution model for the face representation was chosen using the Inter-Pupillary Distance (IPD).

The recognition algorithm was now operating in a completely automated manner and was dealing with pose, illumination, and expression variations as well as the presence of facial occlusions in a completely unsupervised setting (as as no labeled training data was used at any stage (facial alignment, representation, recovery, or coefficient generation) in the recognition pipeline). Thus, the LFW evaluation protocol used was the unsupervised one.

The averaged (over 10 different folds) Receiver Operating Characteristic (ROC) curves obtained using the different facial alignment approaches and the same fixed face recognition algorithm are shown in Figure 4.6 (a). As can be seen, there is a gap in the recognition performance obtained using the better performing alignment algorithms (RCPR, DRMF, and ours) and those obtained using the other facial alignment approaches (AAM-Wild, TSMs, and CDSM). While the use of DRMF and RCPR resulted in marginally better ROCs than the one obtained using our align-

Table 4.3: Performance of various face recognition algorithms on the LFW database using the unsupervised protocol. The performance measure is the Area under the ROC Curve (AUC).

Algorithm	AUC
SD-MATCHES (LFW Funneled) [166]	0.5407
GJD-BC-100 (LFW Funneled) [166]	0.7392
AAM-Wild + [18]	0.7462
H-XS-40 (LFW Funneled) [166]	0.7547
TSMs + [18]	0.7734
LARK (LFW-a) [167]	0.7830
LHS (LFW-a) [168]	0.8107
CDSM + [18]	0.8405
Our Facial Alignment + [18]	0.8585
DRMF + [18]	0.8772
RCPR + [18]	0.8880
MRF-MLBP [169]	0.8994
Spartans [54]	0.9228
Pose Adaptive Filter (PAF) [170]	0.9405

ment approach, the gap in performance between our approach and these two approaches lies in a very narrow band (approximately 2%). While we have already demonstrated that the landmark localization accuracy obtained by our approach is superior to the accuracies obtained by these approaches, the tolerance of the facial recognition algorithm to facial alignment results (see [18]) meant that gains made on the accuracy on this front did not translate to exactly proportional gains in facial recognition rates. In addition to this, it must also be kept in mind that the LFW unsupervised face recognition protocol does not result in all images in the database being used for evaluation. Finally, it must also be mentioned that DRMF and RCPR were provided with an initialization advantage, while our approach was able run on the images without the requirement of a face detection bounding box.

For reference, we also provide a plot of the average ROC curve obtained using our alignment approach in conjunction with previously used face recognition algorithm along with the average ROC curves obtained using the best performing unsupervised techniques (some of which use the LFW funneled or LFW-a data) in literature in Figure 4.6 (b). As can be seen, the combination of using our facial alignment results and the face recognition algorithm in [18] is quite competitive with state-of-the-art algorithms and is only outperformed by the Spartans [54], Pose Adaptive Filter (PAF) [170], and the MRF-MLBP [169] techniques, and this would not be the case if alignment results obtained using TSMs or CDSM were used. For the unsupervised protocol, the Area Under the ROC Curve (AUC) value, and not the mean classification accuracy, serves as a metric to measure performance as there is no legitimate way to select a threshold for the results without using labels or label distributions. These AUC values obtained by the various state-of-the-art recognition algorithms as well as the face recognition algorithm in [18] in conjunction with the different facial alignment algorithms can be found in Table 4.3. The values in the table again demonstrate how much of an impact facial alignment results can have on a particular (fixed) face recognition algorithm.

Chapter 5

Application of our Facial Alignment Algorithm to Analysis of Naturalistic Driving Videos

"My sensors indicate you're somewhat disturbed, Michael."
The Knight Industries 2000 (K.I.T.T) on *Knight Rider - K.I.T.T the Cat (Season* 2 *Episode* 7*)*, and a recurring theme on the show

In this chapter, we apply our facial alignment algorithm (described in chapter 3) to the problem of landmark tracking across the frames of challenging videos that have been recently released. The videos used for evaluation were acquired as part of a Naturalistic Driving Study (NDS) commissioned by the Federal Highway Administration (FHWA) [61] in order to aid with research targeted at assessing driver behavior and improving driver safety using computer and vision and machine learning algorithms. As part of our efforts to assist with this goal, we carried out experiments to demonstrate the efficacy of our facial alignment algorithm when applied to videos, the task of head pose estimation, and the determination of whether a cell phone was being used or not by subjects in the various videos.

5.1 Introduction

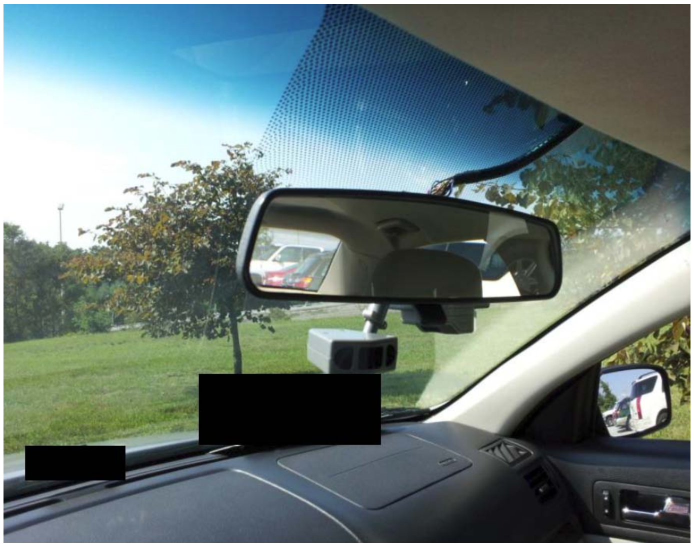
The number of deaths due to distractions caused during driving are on the rise, not just in the US but across the world. In 2013, $3,154$ people lost their lives and an estimated $424,000$ were injured in the US due to a distracted driver [171]. Distraction due to cell phone usage constitutes a sizable portion of the statistic with 18% of the incidents involving cell phone usage in 2009. In order to study the more general problem of driver behavior, the Federal Highway Administration (FHWA) recently recorded over $3,100$ videos of volunteer drivers under naturalistic driving scenarios [172] over a period of 2 years under the Strategic Highway Research Program 2 (SHRP2) [173] program using a custom Data Acquisition System (DAS) developed by the Virginia Tech Transportation Institute [174], [175]. The database [176], [177] size exceeds 2 Petabytes and contains over one million hours of footage that is unmatched in its size, scale, and real-world acquisition conditions in the transport community. However, it poses several challenges to researchers in the field. Firstly, the data suffers from low-resolution artifacts and widely varying illumination conditions. Secondly, due to its size, manual analysis and the providing of ground truths for all frames in the video footage is infeasible. To address both issues the FHWA commissioned an exploratory project that challenged researchers in university and industry to develop computer vision and machine learning based algorithms that were capable of processing such challenging naturalistic driving videos and detecting signs of tiredness in drivers, cell phone usage by drivers, tracking head pose, monitoring if the driver had both hands on the steering wheel, *etc.* [62]. The driver monitoring algorithms developed could be useful in automating the process of annotating the videos that have already been collected or collected during a future study or for deployment in a real-world scenario for driver monitoring as part of a law enforcement effort or to automate

It is in this context that our work aims at addressing the specific problems of head pose estimation and of detecting whether a driver is holding a cell phone in one hand and using only one hand to control the steering wheel of a vehicle. These tasks are both easily carried out if accurate facial landmark localization is carried out on a frame by frame basis in such videos. This ties in quite well with the work we have described in chapter 3, as we are able to easily extend our facial alignment algorithm from annotating still images to localizing the same set of facial landmarks in videos by using information from previous frames to aid the localization process in future frames.

The rest of this chapter is organized as follows. Section 5.2 provides details on the specific database that was the focus of attention in our experiments and evaluations. Following this, section 5.3 goes into details of how our facial alignment was suitably modified for localizing facial landmarks in the frames extracted from the videos in the database. This section also reports the facial landmark accuracies that were obtained using our approach and benchmarks them against results obtained using a commercial face detection and landmark localization algorithm. Sections 5.4 and 5.5 respectively provide context for these results by using the landmark localization results for the tasks of head pose estimation and the determination of whether the subjects in segments of the various videos were using a cell phone, *i.e.*, holding it up to one of their ears and thus keeping only one hand on the steering wheel of a car, or not. Finally, section 5.6 provides a summary of our contributions in this chapter, offers some concluding remarks, and highlights some possible research directions to pursue in future work.

5.2 Details on the Data Used in Our Studies

Full sharing of the previously mentioned SHRP2 NDS data is difficult due to privacy constraints regarding the possible identification of the subjects who participated in the data acquisition from the GPS coordinates of the start and end of trips and the face view videos. However, an alternative dataset is available at no charge to researchers under a less restrictive data sharing agreement. This dataset, referred to as the Head Pose Validation (HPV) dataset, is similar to the SHRP2 NDS data



[redacted]

[redacted]

Figure 5.1: The setup of the DAS head unit and cameras that was used for acquisition of the Head Pose Validation (HPV) data. This image has been reproduced, with some minor changes, from a document providing an overview of the head pose validation data that was obtained after signing a data sharing agreement. Certain portions of the image have been covered with black patches in order to prevent the dissemination of any information that is not to be made public under the terms of the data sharing agreement.

and was recorded by VTTI to measure head pose in drivers and the evaluate different head pose estimation techniques, such as VTTI's mask system.

The platform for collecting the HPV data was a 2001 Saab $9-3$ equipped with two proprietary Data Acquisition Systems (DAS). The collected data included digital video, GPS position and heading, acceleration, rotation rates, and ambient lighting collected at rate that varied from varied from 1Hz to 15Hz. The DAS units also collected data produced by the mask system. The participant was seated in driver's seat of the car and an experimenter (equipped with a laptop) was present with the participant. The experimenter supervised data collection and provided guidance to the participant. A hand-held trigger connected to one of the DAS units allowed the experimenter to annotate the DAS data stream whenever an event of interest occurred. In order to collect the participant's face view videos, a camera was mounted below the rear view mirror, as shown in Figure 5.1.

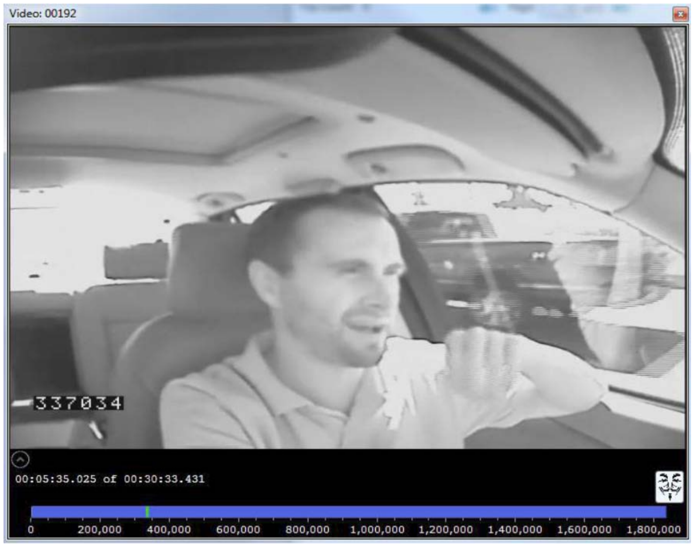


Figure 5.2: A frame from one of the 720×480 "full face view" videos that is not a part of the finally released HPV dataset and can be found on the InSight data access website [19]. This image has been reproduced from a document providing an overview of the HPV data that was obtained after signing a data sharing agreement.

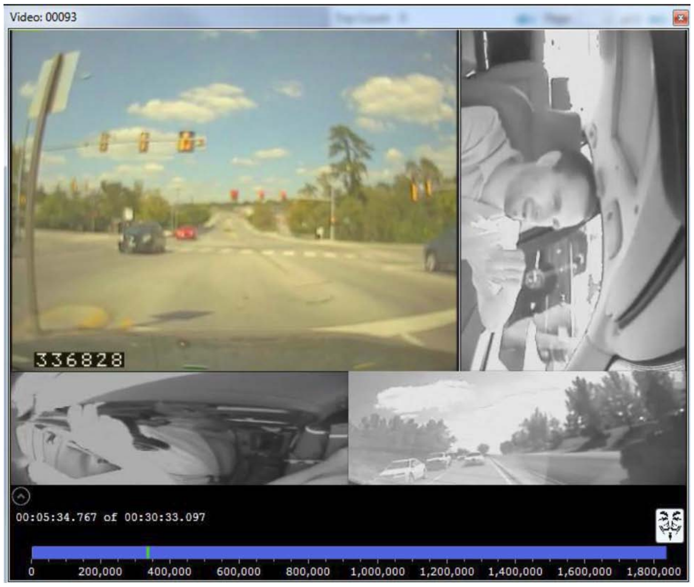
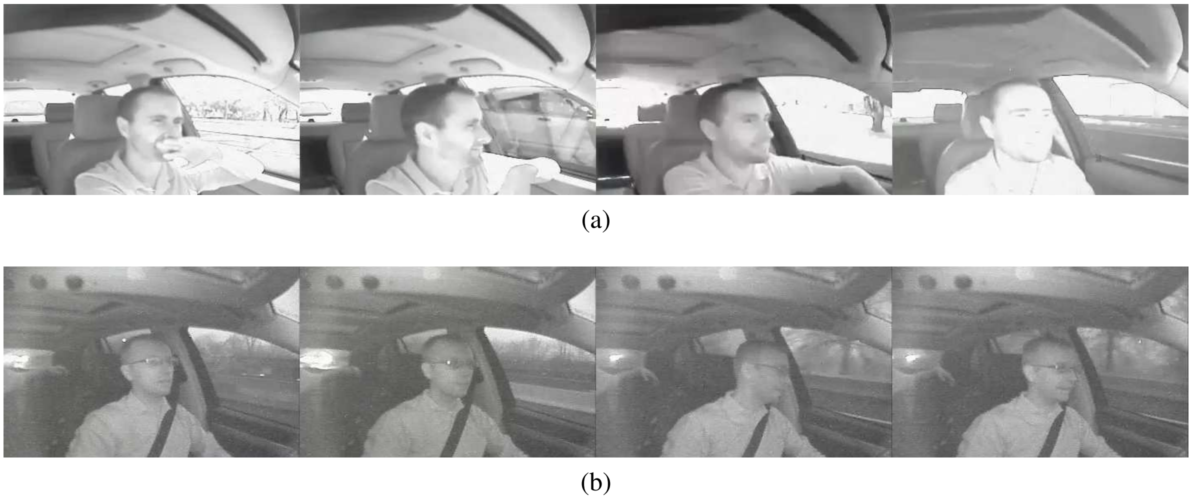


Figure 5.3: A sample frame showing the standard SHRP2 video views recorded by the SHRP2 configured Data Acquisition System (DAS). The frame in this figure was taken from videos that are not a part of the finally released HPV dataset. This image has been reproduced from a document providing an overview of the HPV data that was obtained after signing a data sharing agreement.



(a)

(b)

Figure 5.4: Sample SHRP2 face view video frames from videos acquired during the (a) daytime, (b) night. The frames are from videos that are similar to, though not part of, the videos in the HPV dataset and can be found on the InSight data access website [19]. The frames serve to illustrate the challenging nature of the HPV videos.

One of the DAS units collected a single channel of minimally compressed (resolution of 720×480), full face digital video at 15 frames per second. These videos are referred to as “full face view” videos and a frame from one such video which is not a part of the finally released data is shown in Figure 5.2. The other DAS unit collected standard SHRP2 videos. The two video streams were aligned using GPS timestamps that were recorded. The SHRP2 videos comprise of four channels of video, forward view, face view (resolution of 356×240 with the typical face region (a square region enclosing the convex hull of the facial landmarks of interest) in the frame of size 65×65), lap and hand view, and rearward view, recorded at 15 frames per second and cropped and compressed into a single quad video, as shown in Figure 5.3. It is the SHRP2 face view videos in the HPV dataset that we focused on in our work.

Some of the SHRP2 videos were acquired when the participant was seated in a stationary vehicle (static trials), while others were acquired when the participant was driving (dynamic trials). The environmental conditions (time of day) also varied in the videos and some were captured during the day, others at night, and the remaining during a transition period from daylight conditions to nightfall. Sample SHRP2 frames from videos that are not part of the released HPV dataset but are of the same resolution and acquired using a similar setup are shown in Figure 5.4. As can be seen, the videos pose quite a challenge to face detection and landmark localization algorithms due to the varying illumination conditions, presence of facial occlusions (hands covering the face, presence of glasses and sunglasses, *etc.*), and the extreme pose variation (roll, yaw variation from $-90°$ to $+90°$ and sometimes even beyond this range, and pitch variation during the performance of certain tasks). While the videos acquired at night contain significantly less light, as illumination is provided by infrared LEDs on the DAS in these videos, the daytime and transition videos pose a significant challenges due to the sometimes harsh glare present due to sunlight. In the static trials, the data was acquired in a research lot at VTTI with each of the 24 participants asked to perform a series of glances to predefined locations (such as the left window or mirror, forward windshield, center console, *etc.*) or to simulate a brief cell phone conversation. Each static trial participant was asked to wear four pairs of eyeglasses (including a pair of sunglasses) and a baseball cap and complete the glancing and cell phone simulation tasks under these varying conditions. The dynamic trials were conducted on a predefined route that was approximately 15 miles long and included a variety of road types around Blacksburg, Virginia. Over the course of the drive, each of the 24 participants were asked to perform various tasks that included reporting the vehicle's speed, turning the radio on and off, locating a cell phone in the center console and completing a brief simulated cell phone conversation, *etc.*. The prompted tasks were completed at roughly the same location on the route for each of the participants and were completed only if the participant felt safe in carrying them out.

This video data as well as additional data, such as kinematic data, static and dynamic vehicle segments, details on the participants (sex, skin tone, presence of facial hair, *etc.*), manually labeled ground truth locations for seven facial landmarks for the video frames in several trip segments, head pose estimates for frames in some video segments, details on the tasks performed by the participants during the static and dynamic trials (on certain segments of the videos), *etc.* is what constitutes the full HPV dataset. Out of the 48 videos, data for 2 of the static trials and 2 dynamic trials are being withheld (to be possibly released at a future date), bringing the total number of full face view videos (and associated data) in the released dataset to 44. However, the total number of SHRP2 face view videos in the released data which we worked with, referred to as the "clipped" data, is 41 with 20 videos (and associated data) acquired from static trials and 21 videos (and associated data) acquired from dynamic trials. It is to be noted that only the data that does not contain personally identifying information has been released publicly. Access to personally identifying data, such as the SHRP2 videos, is governed by a data sharing agreement. For this reason, any figure in this chapter containing the face of a subject who participated in the trials and whose video appears in the full face view videos set or the set of SHPR2 videos in the HPV dataset has been masked out using black patches. To illustrate our approach, we sometimes use frames from similar (SHRP2 quality) videos obtained from the InSight data access website [19], which do not have such restrictions governing use.

5.3 Facial Landmark Localization in Video Sequences

5.3.1 Related Work

Tracking of facial landmarks in video sequences has been carried out in the past alongside facial alignment in still images using suitable extensions or modifications. The common tools used for this were AAMs, ASMs and Kalman filters [178], [179], which have also been used in conjunction for tracking of different objects in video sequences, most commonly, human contours. Baumberg and Hogg [180] proposed a method for such an application by tracking the shape coefficients of an ASM independently and attained a speed up in fitting due to this. Baumberg built on this work and proposed a more efficient tracking method that utilized knowledge gained from the Kalman filtering process in order to not only initialize an ASM but also improve the search direction when determining the most suitable location for a landmark [181]. Lee *et al.* [182] achieved real time tracking of human contours using a hybrid algorithm that predicted the initial human outline using Kalman filtering in combination with block matching and a hierarchical ASM to perform model fitting.

In the field of tracking facial landmarks, Ahlberg [183] proposed a near real-time face tracking method that used an AAM. Pu *et al.* [184] reported results obtained using an ASM in combination with a mean shift based method [185] and a Kalman filter to obtain a bounding box around the face and hence initialize the ASM in every frame. Prabhu *et al.* [146] built on some of these ideas to perform tracking of individual facial landmarks, not just the face bounding box, using a Modified Active Shape Model [55] and Kalman filters. They proposed two tracking approaches. The first approach used a constant acceleration model, described in [179], to track the locations and velocities of the individual facial landmarks, once MASM had localized them in the first frame of the video sequence. The second approach tried to account for the correlated motion of the landmarks and rather than rather than tracking the landmark locations themselves, aimed at tracking parameters that affect these positions. These parameters consisted of the translation of the mean of all landmarks in the image, the rotation angle of the face, the size of the face, and some of the dominant PCA coefficients of the facial structure. Using estimates of the tracked components and the remaining PCA coefficients, the coordinates of all landmarks were reconstructed.

In recent years, some of the state-of-the-art facial alignment algorithms that we have described in chapter 2, such as SDM [96] and CDSM [16], have also demonstrated the capability to process video feeds by simply using alignment results from the previous frame as initialization for subsequent frames. While this approach, and indeed all the previously mentioned approaches, can be acceptable for constrained videos, there is a fundamental flaw that exposes them when dealing with unconstrained videos in which a subject exhibits rampant pose variation or when facial occlusions are present. This flaw is error propagation. Without a mechanism in place to determine when initialization from the previous is poor, when a face has been completely lost or is not present in the frame at all, or when the limits of pose tolerance have been reached (SDM, for example, does not handle absolute yaw in excess of $45°$), it is not possible to know when a re-initialization (face detection followed by all steps in the respective facial alignment pipeline) is required. This is something that our approach is able to address using the occlusion/misalignment labels that are determined for each landmark in every frame. In addition to this, our approach is also equipped to seamlessly shift between determining a particular set of landmarks (68 points) for frontal faces and an alternative set of landmarks (39 points) for profile faces. This is quite a challenging problem for most tracking approaches to deal with as the fundamental assumption made is that the number of points being tracked is the same. Thus, our framework for video-based facial alignment is an extremely general one that make minimal assumptions regarding the presence of occlusions, changes in scene or aspect ratio (zooming in or out), pose changes of the subject, *etc.*and is equipped to deal with all of these factors because of the minimal assumptions it makes when dealing with still images. Our approach is described in section 5.3.2 and we go on to detail results produced using it in section 5.3.3.

5.3.2 Our Approach

We are able to extend our previously described facial alignment algorithm (described in chapter 3) to enable landmark localization in frames extracted from videos. The main feature that we take advantage of to enable this is the fact that our approach provides confidence scores and misalignment/occlusion labels for all localized landmarks. This allows us to ascertain the goodness of fit of all landmarks on a particular frame and makes it easy to determine whether these landmarks can be used as initialization for the face in the next frame or not. Consider a frame $\mathbf{I}_t$ in a video sequence $\{\mathbf{I}_t\}_{t=1}^{T}$ with T frames. The goodness of fit g_{t-1} for the previous frame $\mathbf{I}_{t-1}$ is calculated using equation (5.1), in which N_t^{inliers} is the number of inliers (accurately localized) landmarks among

Algorithm 2 Facial alignment in video sequences using our approach.

```
Input: Video frames $\{\mathbf{I}_t\}_{t=1}^{T}$ and pre-trained yaw and expression specific models $\{\mathrm{M}_m\}_{m=1}^{16}$
Output: Final landmark locations for all frames $\{\mathbf{s}_t\}_{t=1}^{T}$ and associated misalignment/occlusion labels $\{\mathbf{o}_t\}_{t=1}^{T}$

for $t = 1, \ldots, T$ do
  if $(t == 1)$ then
    Run face detector on frame $\mathbf{I}_t$
    Run all stages of alignment algorithm using all models
  else
    if $(g_{t-1} > \mathrm{THRESH\_1})$ then
      Run refinement stage of alignment algorithm using models $m_{t-1} - 1$, $m_{t-1}$, and $m_{t-1} + 1$
    else if $(\{g_{t'}\}_{t'=t-5}^{t} > \mathrm{THRESH\_2})$ then
      Run face detector on frame $\mathbf{I}_t$
      Run all stages of alignment algorithm using models $m_{t'} - 1$, $m_{t'}$, and $m_{t'} + 1$
    else
      Run face detector on frame $\mathbf{I}_t$
      Run all stages of alignment algorithm using all models
    end if
  end if
  Save landmark localization results $\mathbf{s}_t$, misalignment/occlusion labels $\mathbf{o}_t$, pose index $m_t$, and goodness of fit $g_t$ (calculated using equation (5.1)) from best fitting model for frame $\mathbf{I}_t$
end for
Output $\{\mathbf{s}_t\}_{t=1}^{T}$ and $\{\mathbf{o}_t\}_{t=1}^{T}$
```

the set of landmarks N_t localized in frame $\mathbf{I}_t$.

$$g_t = \frac{N_t^{\mathrm{inliers}}}{N_t} \tag{5.1}$$

If g_{t-1} for frame $\mathbf{I}_{t-1}$ exceeds a certain threshold $\mathrm{THRESH\_1}$ (set to 0.55 in our work), then the results from this frame can be used for shape initialization on frame $\mathbf{I}_t$. This saves a lot of computation time during the fitting process on this frame as only the shape refinement stage (and not the sparse landmark determination step and dense shape evaluation steps) of our alignment pipeline needs to be carried out. In addition, since information on the most suitable pose model m_{t-1} is available for frame $t-1$, only models $m_{t-1} - 1$, m_{t-1}, and $m_{t-1} + 1$ need be evaluated on frame $\mathbf{I}_t$. In case the goodness of fit for frame $\mathbf{I}_{t-1}$ falls below $\mathrm{THRESH\_1}$, then it is assumed

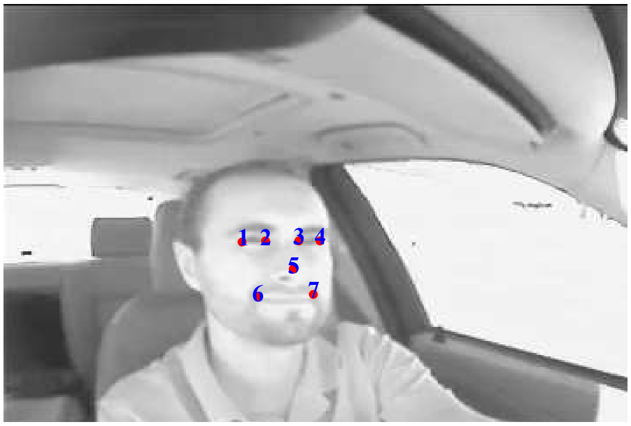


Figure 5.5: An SHRP2 face view video frame from a video that is not a part of the released HPV dataset and can be found on the InSight data access website [19] that shows the locations of the 7 landmarks for which manually annotated ground truths are available on some frames from the HPV SHRP2 face view videos.

that the presence of facial occlusions have caused this or that misalignment has occurred for other reasons. In such cases, a new shape initialization (using a face detection result) must be determined for frame $\mathbf{I}_t$. If a high confidence alignment (with a goodness of fit value that exceeds $\mathrm{THRESH\_2}$, set to 0.60 in our work) can be found among the previous 5 to 10 frames, then the pose model $m_{t'}$ from this frame $t^{'}$ is saved and used on frame $\mathbf{I}_t$ with the the sparse landmark determination, dense shape evaluation, and shape refinement stages of our approach carried out on frame $\mathbf{I}_t$ using only models $m_{t'} - 1$, $m_{t'}$, and $m_{t'} + 1$. In the event that this too is not possible, the algorithm resets to treating the current frame like it did the first frame, with no knowledge regarding facial bounding box location or pose information, and a face detection step followed by all stages in our facial alignment framework are carried out in order to localize the appropriate landmarks. Algorithm 2 summarizes this video fitting process.

5.3.3 Results

Our approach was tested on all 41 SHRP2 face view videos in the HPV dataset. In order to ensure better trained shape and texture models, our models were trained on images from the LFPW

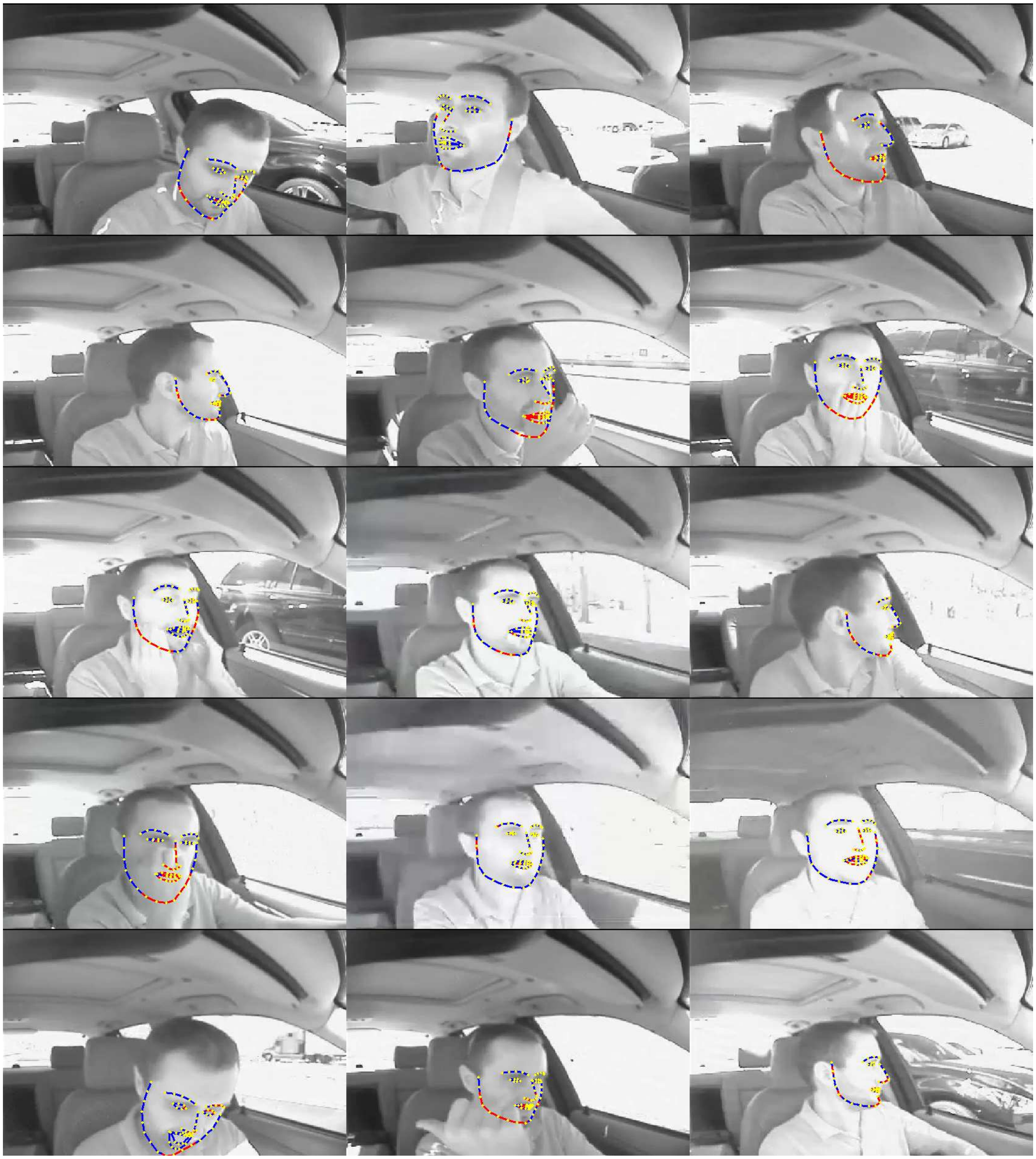


Figure 5.6: Qualitative landmark localization results produced by our approach on frames from a video that is similar to, though not part of, the videos in HPV dataset and can be found on the InSight data access website [19]. In all facial images with landmarks overlaid on them, yellow dots are used to indicate the locations of facial landmarks, blue line segments indicate that the landmark at their center is accurately localized, and red line segments indicate that the landmark at their center is misaligned or potentially occluded (goodness of fit feedback). The same color scheme is maintained in all figures that show facial alignment results produced by our approach in this chapter. Zoom in to see details.

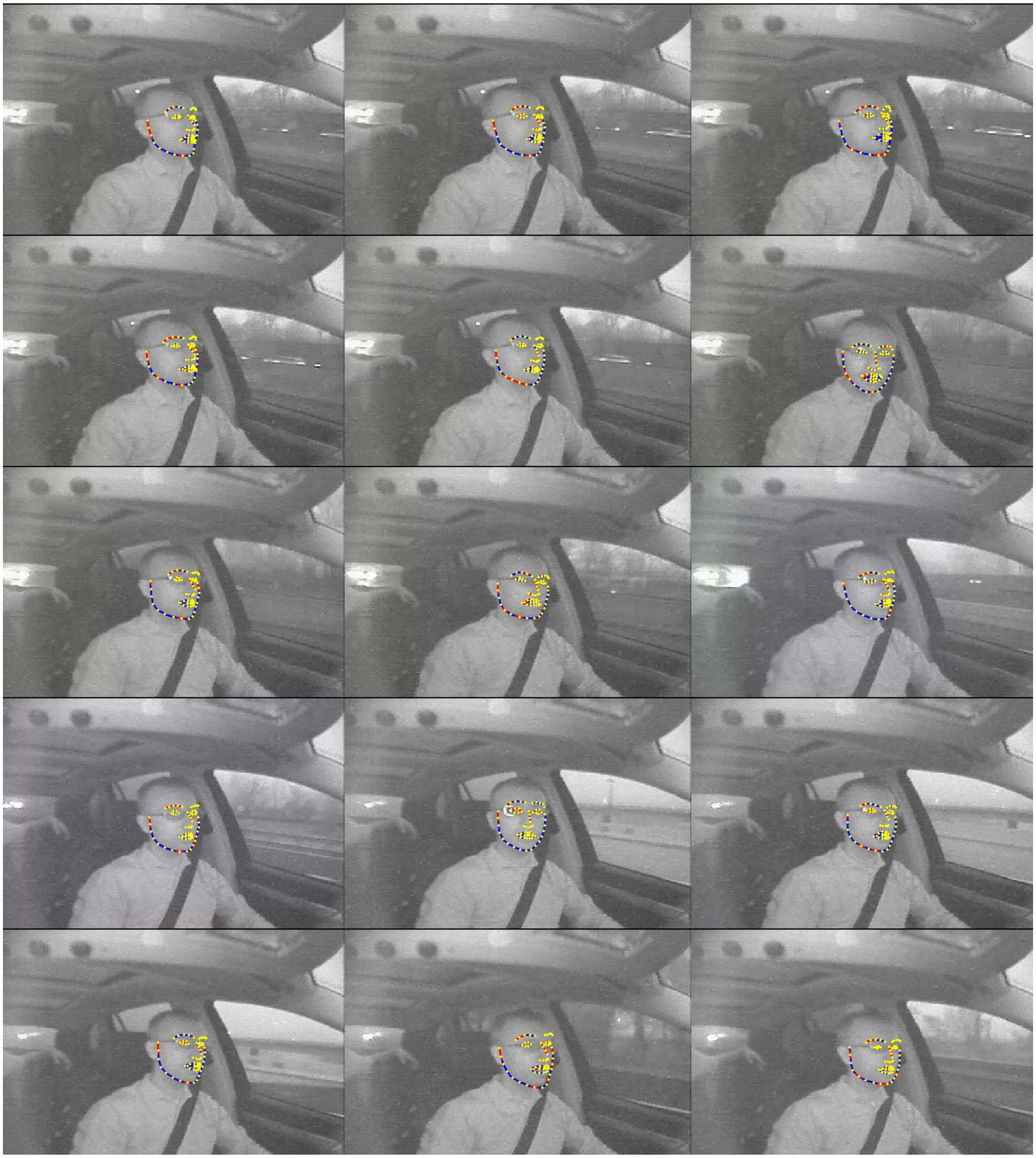


Figure 5.7: Qualitative landmark localization results produced by our approach on frames from a video that is similar to, though not part of, the videos in HPV dataset and can be found on the InSight data access website [19]. Zoom in to see details.

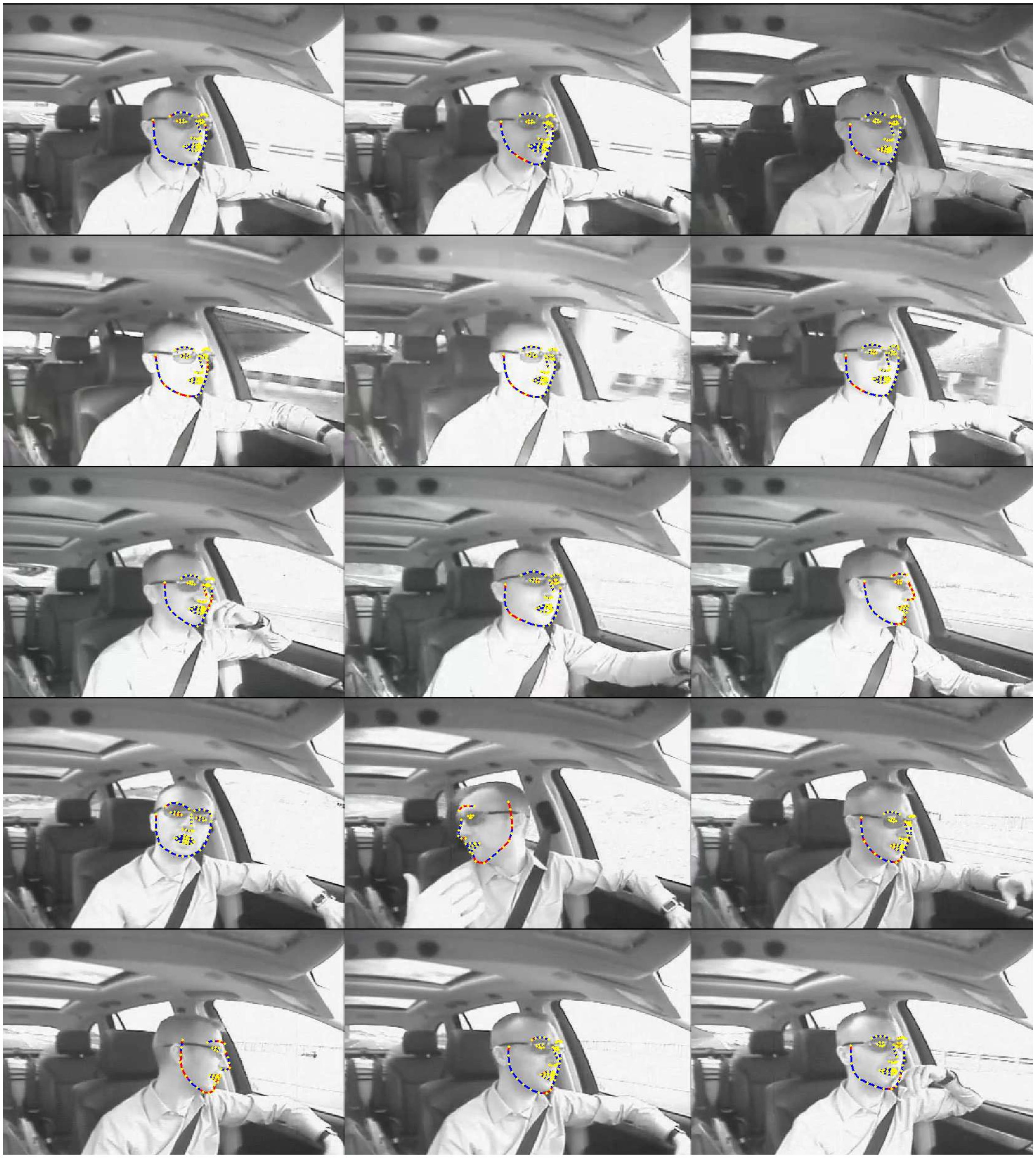


Figure 5.8: Qualitative landmark localization results produced by our approach on frames from a video that is similar to, though not part of, the videos in HPV dataset and can be found on the InSight data access website [19]. Zoom in to see details.

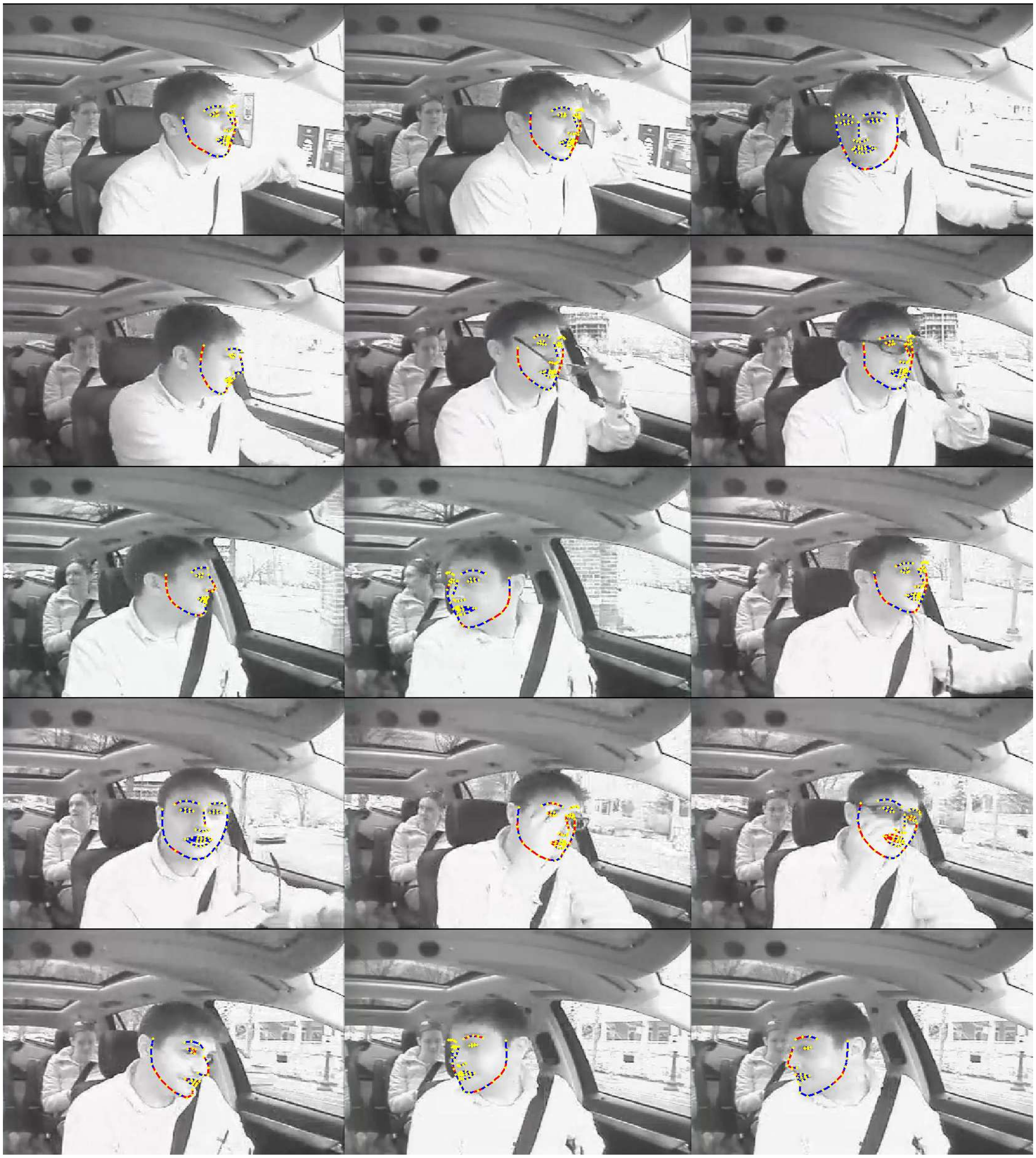


Figure 5.9: Qualitative landmark localization results produced by our approach on frames from a video that is similar to, though not part of, the videos in HPV dataset and can be found on the InSight data access website [19]. Zoom in to see details.

(811 images from the training set partition), Helen (2000 and 330 images from the training set and testing set partitions, respectively), AFW (337 images), and ibug (135 images) datasets in addition to the previously used $6,495$ MPIE training images (see section 3.2.1), as was the case in section 4.2, in which our alignment approach was used to localize landmarks that served as input to a face recognition algorithm. We used the commercial Pittsburgh Pattern Recognition (PittPatt) face detection algorithm for face detection on all frames where this was required for initialization. In all, a total of $911,018$ frames were processed in an extremely large-scale experiment. Due to the challenging illumination conditions (low light conditions, excessive glare, or transition from daylight to night) in the videos, several frames could not be annotated as the face detection algorithm failed to find a face in them and results from previous frames could not be used for initialization. In addition to these frames, several frames did not contain a face as the subject in the videos had stepped out of the car or had turned his/her head to an extent that no facial features were visible.

It must be noted that manually annotated ground truth landmark locations and head pose estimates are available for only a small fraction of the total frames in the videos (approximately 7%) that occur consecutively for short periods during the videos (usually when an event of interest occurs). These ground truth coordinates were determined by video reviewers (reductionists) on the high resolution 720×480 full face view videos and transferred to the low-resolution SHRP2 face view videos using appropriate post-processing. These frames were annotated by two reductionists trained to follow a fixed protocol for consistency. Reliable guesses were used when a landmark was not clearly visible (due to an occlusion or challenging illumination conditions) and the landmark coordinates were recorded as missing when a best guess could not be made. Manual annotations (ground truths) for only a maximum of 7 such landmarks, shown in Figure 5.5, were available for all these frames.

Some results produced by our landmark localization technique on frames from videos that are not part of the HPV SHRP2 face view videos set but are quite similar in resolution and acquisition conditions are shown in Figures 5.6 - 5.9. As can be seen, our approach is quite robust to the

Table 5.1: Landmark localization performance of our approach and the baseline on frames from all static trial videos in the HPV SHRP2 face view video set. MFE and MNFE values were computed using the landmarks common to those localized by the approaches and the manually provided ground truths over all frames for which coordinates of landmarks from both approaches and the ground truth coordinates were available.

Time of Day	Total Frames	Frames Evaluated	Algorithm					
			Ours			Baseline		
			Frames Annotated	MFE	MNFE (%)	Frames Annotated	MFE	MNFE (%)
Day	$150,352$	$12,121$	$131,181$	2.93	9.67	$117,510$	3.45	11.46
Transition	$109,385$	$11,177$	$100,647$	2.99	10.44	$102,422$	2.86	9.95
Night	$86,959$	$8,471$	$72,927$	3.78	12.47	$74,353$	2.91	9.58

Table 5.2: Landmark localization performance of our approach and the baseline on frames from all dynamic trial videos in the HPV SHRP2 face view video set. MFE and MNFE values were computed using the landmarks common to those localized by the approaches and the manually provided ground truths over all frames for which coordinates of landmarks from both approaches and the ground truth coordinates were available.

Time of Day	Total Frames	Frames Evaluated	Algorithm					
			Ours			Baseline		
			Frames Annotated	MFE	MNFE (%)	Frames Annotated	MFE	MNFE (%)
Day	$264,735$	$7,018$	$236,866$	2.92	10.06	$225,587$	2.96	10.16
Transition	$137,175$	$3,474$	$120,547$	3.08	10.75	$123,837$	2.96	10.35
Night	$162,412$	$4,121$	$140,250$	4.25	14.00	$144,804$	3.37	11.24

changes in head pose and the presence of facial occlusions in these videos. In some cases, our approach also demonstrated a tolerance to pitch variation, which is creditable considering that it was not explicitly trained on many images with such variations. Such results could be explained by the accurate initialization provided from previous frames and would be harder to obtain on still images with no contextual information.

In addition to these qualitative results, we also provide quantitative results of the landmark localization accuracy obtained by our approach on a fraction of the video frames for which ground truth annotations were available. We also provide results obtained using the PittPatt face tracking and landmark localization software, that serve as a baseline. This algorithm was run in serial tracking mode on the videos in order to improve face detection and landmark localization accuracies by tracking the face in the frames over time. Results obtained using this approach [186] were made available to us as we were involved with carrying out research for the previously mentioned FHWA exploratory project. PittPatt localizes a maximum of 3 landmarks that are common to our approach and the ground truths (the centers of the eyes, obtained by taking the mean of the respective eye corner coordinates, and the tip of the nose) and does not always provide an output indicating their locations for each frame. This posed a standardization problem to us when reporting results and thus we report results obtained by averaging the Euclidean distances between the coordinates of the landmarks common to those localized by the two algorithms and the ground truths over frames where all three sets of coordinates were available (2 landmarks for profile faces and 3 for all other cases). Our results are grouped into categories based on the circumstances under which the videos were captured (static or dynamic trials and day, transition, and night conditions).

Tables 5.1 and 5.2 summarize all results of this experiment and report landmark localization error values for the two algorithms on a subset of frames over which both algorithms localized a set of common landmarks and for which ground truths were also available on the static and dynamic trial videos, respectively. The results in both tables are also organized based on the time of acquisition (day, transition, and night). As the tables show, the number of frames over which the errors were computed (frames evaluated) is much smaller than the total number of frames in the videos or the number of frames over which both algorithms successfully localized facial landmarks (frames annotated) due to the limited number of manually annotated frames (frames with ground truth coordinates for facial landmarks). The landmark localization error metrics used in the tables are the Mean Fitting Error (MFE) values computed as the Euclidean distance between the coordinates of the landmarks that were automatically localized and the ground truth coordinates and averaged over all landmarks and frames and the Mean Normalized Fitting Error (MNFE) of these fitting errors obtained by normalizing using the average eye center to mouth corner distance (the same normalization distance that was used in section 3.2.3 when reporting results obtained on the MPIE test set).

The landmark localization error values (both MFE and MNFE) are quite similar for both approaches and is a reflection of the fact that they are reported by averaging over errors in localizing only 2 or 3 landmarks and the low resolution of the video frames. A very small pixel error translates to a large MNFE value as the normalization distance is quite small (typically around 30 pixels). In addition, these values were computed only using frames where ground truth annotations were available, which generally corresponded to cases when the subject's face was not heavily occluded and during events of interest, where excessive pose variation may not have been manifested. Thus, the actual landmark localization values are less important than the trends that they help to establish regarding the challenges posed by the time of acquisition of the videos (day, transition, or night) and the nature of the trials (static or dynamic). The more relevant statistic is the number of frames that were successfully annotated using our approach as compared to the baseline. In the static and dynamic videos that were acquired during the day, which constitute a majority of the frames, our approach was able to localize landmarks in far more frames than the PittPatt tracker. This is a significant result as it demonstrates the robustness of our approach to the challenging illumination changes, sudden changes in head pose, and the presence of facial occlusions (especially in the static trial videos, when subjects wore baseball caps, sunglasses, *etc.* at different points during the videos) that are frequently encountered in these videos. It must be kept in mind that the PittPatt algorithm is only able to localize a sparse set of landmarks unlike our approach, and does not always provide outputs indicating the locations of all of them, even if a face is detected in a frame. It must also be noted that our approach used the PittPatt face detection algorithm for initialization and for detecting faces in frames where the alignment result from the previous frame was not suitable for initialization purposes on the following frame. Thus, the gains made by our approach on the number of annotated frames are primarily a result of its using the previous frame for initialization (thus not requiring a face detection result) and its tolerance to excessive pose variation and the presence of facial occlusions.

Table 5.3: Landmark localization performance summary of our approach and the baseline on all frames from the videos in the HPV SHRP2 face view video set. MFE and MNFE values were computed using the landmarks common to those localized by the approaches and the manually provided ground truths over all frames for which coordinates of landmarks from both approaches and the ground truth coordinates were available.

Total Frames	Frames Evaluated	Algorithm					
		Ours			Baseline		
		Frames Annotated	MFE	MNFE (%)	Frames Annotated	MFE	MNFE (%)
911,018	46,382	802,418	3.23	10.89	788,513	3.09	10.46

From Tables 5.1 and 5.2, it is clear that the videos acquired at night (under low light conditions) and under transitional lighting (between day and night) posed quite a challenge to both face detection and landmark localization algorithms with a lower percentage of the total frames annotated for these videos in both the static and dynamic trial cases compared to the videos acquired during the day. Regardless of this, as Table 5.3, that summarizes the same landmark localization and frame annotation results for the two approaches over all 911,018 processed frames, demonstrates, our approach was able to localize landmarks in a higher percentage of frames than the PittPatt tracker.

5.4 Head Pose Estimation

Monitoring driver head pose can be particularly useful as this information can be used to ascertain driver state and where his/her attention is directed. For example, looking down while driving (either due to drowsiness or distraction) could prove dangerous. For these reasons, research focused on developing automated algorithms capable of estimating the head pose of drivers was deemed to be quite important by the FHWA. In this section we provide details on the results that were obtained when the landmarks localized by our approach (see section 5.3.3 for details on these alignment results) were used to estimate the head pose of subjects in the various HPV SHPR2 face view videos. The head pose coordinate system used in the study is shown in Figure 5.10

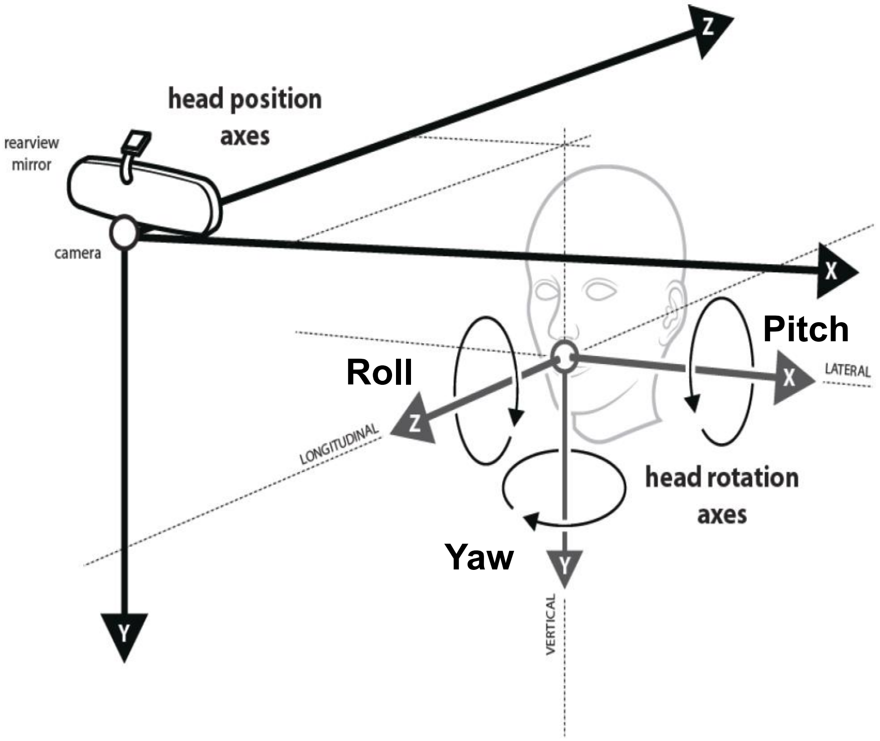


Figure 5.10: Head pose coordinate system used in the HPV dataset showing what yaw, pitch, and roll refer to. A positive yaw is indicated when the driver in the video turns to his/her right, a positive pitch is indicated when the driver looks down, and a positive roll is indicated when the driver's head tilts to his/her left. This image has been reproduced, with some minor changes, from a document providing an overview of the HPV data that was obtained after signing a data sharing agreement.

5.4.1 Results

As was the case with the facial landmark localization comparisons, ground truth head pose estimates are available only for a small fraction of the total frames in the videos (only for frames in which all 7 previously mentioned facial landmarks were visible and could be manually annotated). These ground truths estimates were obtained by using the 7 manually annotated landmark coordinates and facial feature measurements of participants in conjunction with landmark based pose estimation algorithms. The method outlined by Gee and Cipolla in [47] was used to determine the pitch, while the method outlined by Horprasert *et al.* in [48] was was used to determine yaw and roll. The former uses the corners of the eyes, the tip of the nose, and the corners of the mouth, in conjunction with predefined values for the ratios of distances between these points (based on a typical human face), in order to produce estimates for the pitch and yaw of the face, while the latter approach uses the coordinates of the corners of the eyes and the tip of the nose, along with anthropometric data and the coarse structure of the face, to determine facial orientation relative to the camera plane.

As described in section 5.3.3, facial landmark coordinates for most of the frames in the 41 SHRP2 face view videos in the HPV dataset were automatically determined using our approach and a baseline was determined using the PittPatt algorithm. The PittPatt algorithm also provides fairly precise head pose estimates for yaw and roll, with pitch always determined as $0°$. Our facial alignment algorithm can only provide accurate estimates of the roll angle (which can be trivially calculated as the angle between the corners of the eyes or by using two other landmarks and determining the difference between this angle and its typical value for a roll of $0°$) and coarse estimates of yaw, based on the index m (that corresponded to a particular range of yaw variation) of the model that resulted in the most confident shape output (see section 3.1.3). However, the landmarks localized by our approach can be used to determine more precise yaw and even pitch estimates using previously developed approaches. The first approach we used in such a fashion was the geometric based method described by Gee and Cipolla in [47]. The second approach we used was one that has been incorporated into the previously described DRMF [95] facial alignment algorithm by its authors and uses a 3D facial shape model that is aligned with the coordinates of 66 (all landmarks in Figure 3.2 (b) except for landmarks 61 and 65) 2D facial landmarks in order to compute pitch, yaw, and roll estimates. An open source implementation of code to carry out pose estimation using this algorithm is available [165]. We previously used this algorithm and code to obtain more precise yaw estimates from the landmark localization results produced by various alignment algorithms in order to provide this additional input to a face recognition algorithm in section 4.2. Thus, three head pose estimates obtained directly (coarse yaw estimates provided by our alignment algorithm with pitch determined as $0°$) or indirectly (using the previously described

Table 5.4: Performance of various facial landmarks (obtained using our alignment approach) based head pose estimation techniques and the baseline on frames from all static trial videos in the HPV SHRP2 face view video set. Mean Absolute Error (MAE) values for pitch, yaw, and roll, separated by commas and in this order in the table, were computed by considering all frames for which head pose estimates from both approaches and the ground truth coordinates were available.

Time of Day	Total Frames	Frames Evaluated	Algorithm			
			Ours (Coarse)	Ours + [165]	Ours + [47]	Baseline
			MAE (°)	MAE (°)	MAE (°)	MAE (°)
Day	150,352	14,572	9.04, 9.24, 2.64	9.54, 5.32, 4.13	10.72, 6.46, 2.64	9.04, 6.83, 2.93
Transition	109,385	10,319	13.03, 9.60, 3.45	12.38, 5.28, 5.91	12.85, 6.08, 3.45	13.03, 4.97, 2.14
Night	86,959	6,927	9.32, 11.68, 4.15	11.15, 6.49, 5.26	11.44, 7.26, 4.15	9.32, 5.80, 2.02

Table 5.5: Performance of various facial landmarks (obtained using our alignment approach) based head pose estimation techniques and the baseline on frames from all dynamic trial videos in the HPV SHRP2 face view video set. Mean Absolute Error (MAE) values for pitch, yaw, and roll, separated by commas and in this order in the table, were computed by considering all frames for which head pose estimates from both approaches and the ground truth coordinates were available.

Time of Day	Total Frames	Frames Evaluated	Algorithm			
			Ours (Coarse)	Ours + [165]	Ours + [47]	Baseline
			MAE (°)	MAE (°)	MAE (°)	MAE (°)
Day	264,735	7,021	13.10, 9.437, 2.35	12.52, 5.79, 4.03	14.00, 6.89, 2.35	13.10, 7.31, 2.71
Transition	137,175	3,545	16.89, 9.01, 3.21	13.83, 6.98, 4.25	15.64, 7.40, 3.21	16.89, 5.16, 2.70
Night	162,412	4,052	22.56, 11.81, 4.36	18.84, 8.27, 5.47	20.28, 10.14, 4.36	22.56, 6.99, 3.19

geometric and 3D based pose estimation techniques) using our alignment algorithm were evaluated by comparison against ground truth head pose estimates.

Tables 5.4, 5.5, and 5.6 summarize the results of our head pose estimation experiment and report the Mean Absolute Errors (MAE) in the pitch, yaw, and roll estimates obtained by the two algorithms when compared to the ground truth estimates for the same for a set of frames for which these ground truth estimates were available on the static, dynamic, and all video frames, respectively. As was the case with the results reported in section 5.3.3, the results in these tables are also organized based on the time of acquisition (day, transition, and night). As can be seen from the tables, the roll estimates obtained by our approach are very close to the roll estimates

Table 5.6: Performance of various facial landmarks (obtained using our alignment approach) based head pose estimation techniques and the baseline on all frames from the videos in the HPV SHRP2 face view video set. Mean Absolute Error (MAE) values for pitch, yaw, and roll, separated by commas and in this order in the table, were computed by considering all frames for which head pose estimates from both approaches and the ground truth coordinates were available.

Total Frames	Frames Evaluated	Algorithm			
		Ours (Coarse)	Ours + [165]	Ours + [47]	Baseline
		MAE (°)	MAE (°)	MAE (°)	MAE (°)
$911,018$	$46,436$	$12.36, 9.92, 3.20$	$12.00, 5.94, 4.80$	$13.01, 6.95, 3.20$	$12.36, 6.22, 2.59$

by the PittPatt algorithm, with the latter obtaining slightly closer estimates to the ground truth estimates. The coarse yaw estimates obtained using our alignment approach (based on the index m of the model that resulted in the most confident shape output) are not as accurate as the baseline estimates. This was to be expected, however, when the dense set of landmarks localized by our approach were used in combination with the landmark-based pose estimation algorithms in [47] and [165], more accurate (closer to the ground truth estimates) yaw estimates were obtained. The difficulty involved in pitch estimation is apparent as simply using a $0°$ estimate (a safe estimate for most frames in the videos) for pitch (as was provided by our approach (coarse estimate) and the PittPatt algorithm) often resulted similar or even lower MAE values than those obtained using the pose estimation algorithms. It must be noted though that the ground truth values themselves are slightly subjective as they too were obtained in an indirect manner using pose estimation based on a sparse set of ground truth landmark coordinates and not in a calibrated environment with a subject gazing at specified cues in order to measure head pose. Thus, the error values again are less important than the demonstration of the fact that our facial alignment algorithm could be useful in addressing with this problem. It must also be remembered that though our approach (with suitable initialization from the previous frame) does exhibit the capability to deal with faces that exhibit with more pitch variation than it was trained on (see Figure 5.6), this is still a difficult problem that needs to be investigated in future work. Thus, we report the errors in pitch estimates more for the sake of completeness and as a proof of concept and stress that the actual error values must be taken with a pinch of salt in this context.

5.5 Cell Phone Usage Detection

In this section we provide details on experiments that were carried out using the results from our previously obtained facial alignment results in order to automatically determine of whether the subjects in segments of various videos were using a cell phone or not.

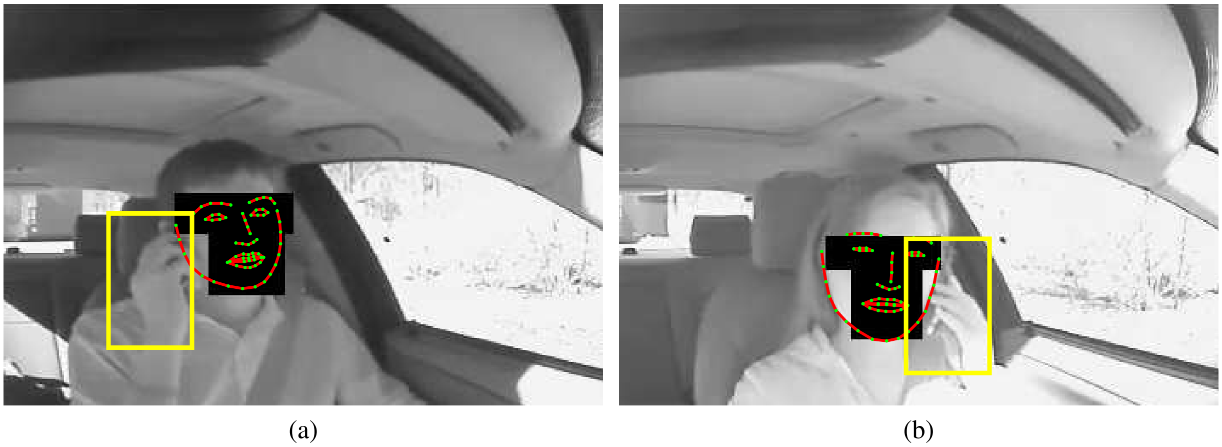
5.5.1 Related Work

Studies in a simulated driving environment under controlled settings have shown that impairment associated with using a cell phone while driving can be as profound as those associated with driving while drunk [187]. Braking reactions were delayed when drivers were conversing on a cell phone, leading to more traffic accidents [187], [188]. Therefore, it is becoming increasingly important to accurately detect cell phone usage by drivers, both from the safety and law enforcement points of view.

There has been a lot of recent work in the broad area of driver behavior monitoring and the specific problem of driver cell phone usage detection. Artan *et al.* [189] used data captured by a highway transportation imaging system, which was installed to manage High Occupancy Vehicle (HOV) and High Occupancy Tolling (HOT) lanes, for detecting cell phone usage by drivers. The cameras used were situated at an elevated position pointing towards the approaching traffic with Near Infrared (NIR) capability to tackle night vision. After the images were acquired, the authors adopted a series of computer vision and machine learning techniques for detection and classification. They first used a Deformable Part Model (DPM) [92] to localize the windshield region within the image and then used the TSMs algorithm [6] for simultaneous face detection, pose estimation, and landmark localization to locate the facial region and crop out a region of interest around the face to check for the presence of a cell phone. Finally, image descriptors extracted from the crops were aggregated to produce a vector representation which was classified using a Support Vector Machine (SVM) [117] classifier to determine if the driver was using a cell phone or not.

Zhang *et al.* [190] also studied a similar problem. In their work however, the camera acquiring the video footage was mounted above the dashboard of a car. They extracted features from the face, mouth, and hand regions and then passed them passed on to a Hidden Conditional Random Fields (HCRF) model for final cell phone usage classification. For face detection, they used a cascaded AdaBoost [115] classifier with Haar-like features [85]. For mouth detection, a simple color-based approach was found to be sufficient because the red component in the mouth region is stronger than the rest of facial region, and the blue component is weaker. Therefore, they operated in the YC_bC_r color space and measured the ratio of C_r/C_b as their cue for mouth region detection. For the detecting hand region, they incorporated both color and motion information.

There has also been some recent research on non-vision based approaches for detecting cell phone usage by drivers. Bo *et al.* [191] leveraged various sensors integrated in today's smartphones, such as accelerometers, gyroscopes, and magnetometer sensors, to distinguish between whether a phone was being used by a driver or a passenger. Yang *et al.* [192] harnessed a car's stereo system and Bluetooth network in an acoustic based approach to estimate the distance of a cell phone in use from the car's center and were thus able to determine whether the user was the driver or not. Breed *et al.* [193] monitored emissions from a cell phone by placing three directional antennas at various locations inside a car. A receiver was associated with each antenna and included an amplifier and a rectifier module that converted radio frequency signals to DC signals which were used to tell which antenna provided the strongest signal. A correlation could then be made for finding the most likely location of a cell phone being used by an occupant in the car.



(a)

(b)

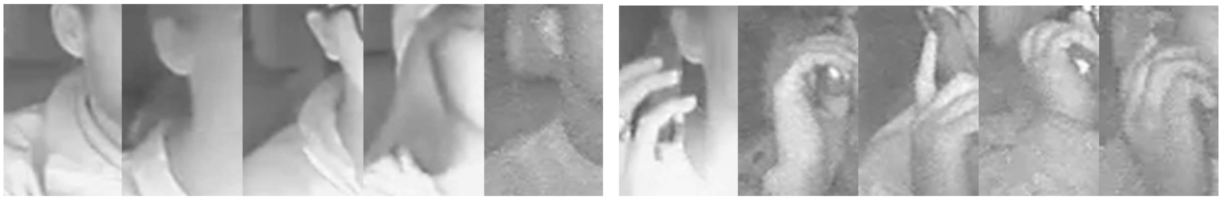
Figure 5.11: The process by which crops of the region of interest were generated to check for the presence of a cell phone being held in the (a) right hand of the subject and (b) left hand of the subject. The faces of the subjects have been covered with black patches as this information cannot be made public under the terms of a data sharing agreement.

5.5.2 Our Approach

This section provides details on our approach for automated cell phone usage detection in frames from the SHRP2 face view videos in the HPV dataset. For the purposes of this study, all training and testing data consisted of frames where the subject's absolute yaw did not exceed $45°$ (in the positive or negative directions). Details on these our training and testing stages follow.

Training Stage

In order to build classifier models for the automatic detection of a cell phone in a supervised setting, it is necessary to provide them with consistently labeled training data. Our training data for cases when a cell phone was not in use (negative class data) consisted of frames from video segments where the subject was either seated in a stationary car and performing tasks such as checking the side view mirrors, looking forward, looking at center console, *etc.*, or was driving and performing tasks such as signaling a lane change, checking the speed of the car, turning the radio on or off, looking forward, *etc.*. In similar fashion, we also used frames from video segments of the same



(a) (b)

Figure 5.12: Sample crops of the region of interest generated to train various classifiers for cases when (a) the subject did not have a cell phone in either hand and (b) the subject had a cell phone in his/her right hand. (a) and (b) appear in [20].

subjects where the subjects were using a cell phone (with one hand pressed close to one of their ears in order to hold it) in a stationary or moving car. In order to build more accurate models, frames where the subject used their right hand to hold the cell phone were manually separated from those in which in which the subject used their left hand to hold the phone. The same set of 68 facial landmarks that we have repeatedly used throughout this thesis were automatically localized in this training data using our facial alignment algorithm for video annotation (see section 5.3.3 for details on these alignment results).

The next step in our training stage involved the generation of crops of the region of interest for both the positive and negative class cases using the facial alignment results. We used 50×80 rectangular crops with landmark 18 (see Figure 3.2 (b)) as the top right corner of the crop region in order to generate positive and negative class crops for cases where subjects were holding (or not holding) a cell phone in their right hand. In similar fashion, 50×80 rectangular crops with landmark 23 (see Figure 3.2 (b)) as the top left corner of the crop region were generated for cases where subjects were holding (or not holding) a cell phone in their left hand. Use of such crops with reference provided by an interior facial landmark ensured more stability and less variance than crops that would be obtained using a facial landmark along the facial boundary as a reference point as these landmarks are usually localized with higher error and exhibit higher variance even in manually clicked ground truth data [4]. Figure 5.11 shows how these crops were generated. Sample crops generated for cases where a cell phone was not being held and cases when a cell

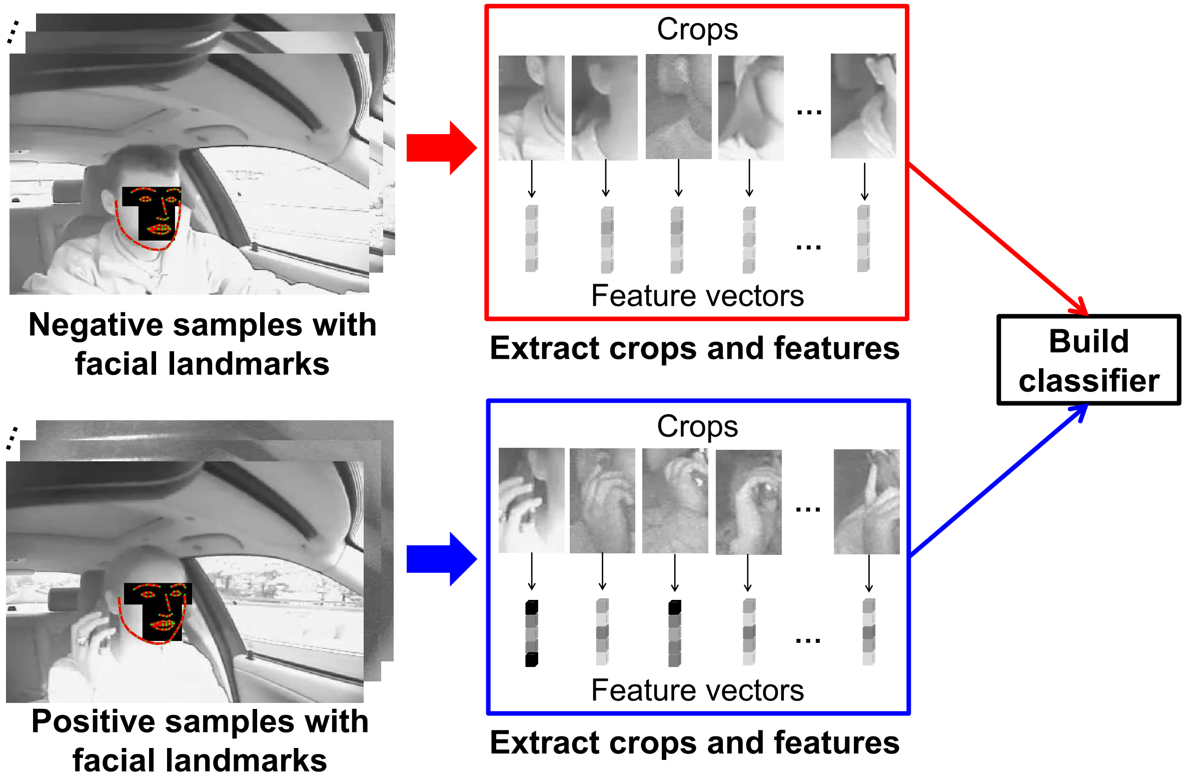

Negative samples with facial landmarks
Crops
...
Feature vectors
Extract crops and features
Build classifier
Positive samples with facial landmarks
Crops
...
Feature vectors
Extract crops and features

Figure 5.13: The process followed by us to train a classifier that could distinguish between cases when a cell phone was being held close to the right ear of a subject and cases when no cell phone was being held up to the right ear. A similar process was used to train another classifier (using the same corresponding algorithm) that could distinguish between cases when a cell phone was being held close to the left ear of a subject and cases when no cell phone was being held up to the left ear. The faces of the subjects have been covered with black patches as this information cannot be made public under the terms of a data sharing agreement.

phone was being held in the right hand are shown in Figure 5.12.

The final stage in the training process was the extraction of features from the positive class (holding a cell phone) and negative class (not holding a cell phone) cases and the building of classifiers using these features. We utilized two different feature representations. When we used raw pixels as features, the feature vectors were 4000 dimensional and were normalized to be unit norm vectors. We also utilized Histogram of Oriented Gradients (HOG) [93] feature descriptors that have been proven to be quite effective in object detection and recognition problems [92]. We utilized HOG descriptors generated with a spatial bin size of 10 and with 9 orientation bins

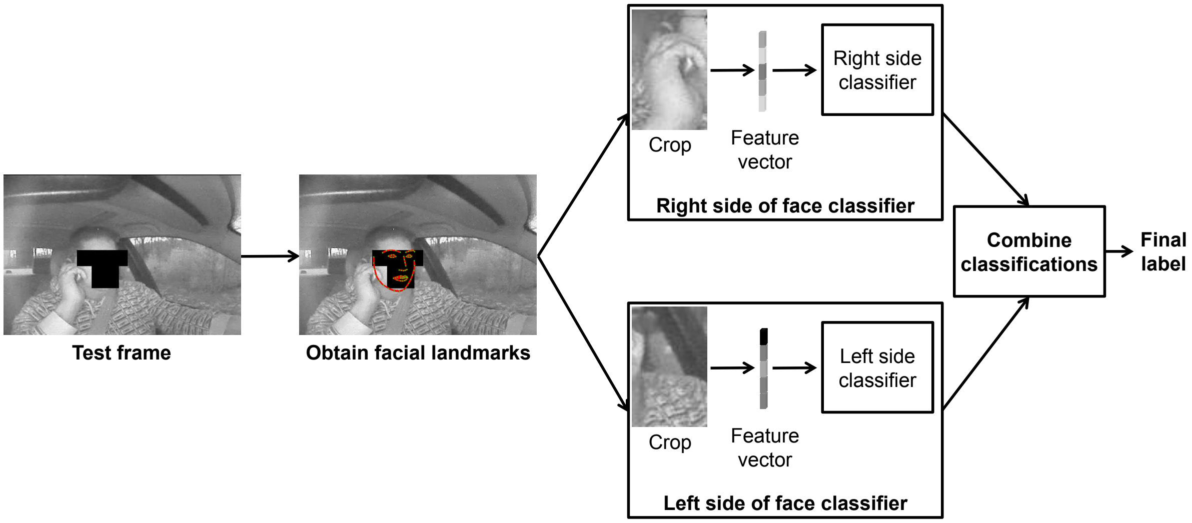


Figure 5.14: The process followed by us to determine if the subject in a test frame was using a cell phone (holding it close to his/her right/left ear) or not. The faces of the subjects have been covered with black patches as this information cannot be made public under the terms of a data sharing agreement.

resulting in a 1008 dimensional feature vector. We benchmarked the performance obtained using these two feature descriptors in conjunction with different classifiers, the first of which is the Real AdaBoost [114] framework of ensemble classifiers. We chose the Real AdaBoost classifier due to the minimal parameters that need to be determined to utilize it (only the number of boosting rounds or number of classifiers in the ensemble need to be specified) and its resistance to overfitting [115], [116]. The Real AdaBoost framework not only allows for the classification of a feature vector as positive or negative, but also returns a confidence score for the prediction. This allowed us to construct Receiver Operating Characteristic (ROC) curves to summarize performance.

The other classifiers we used were a Support Vector Machine (SVM) [117] with a Radial Basis Function (RBF) kernel and a random forest [118]. These classifiers can also be configured to return a value that can be interpreted as a confidence score of their class prediction (a probability value in the case of an SVM and the number of trees that vote for a class label in the case of the random forest). We built two different sets of classifiers to better deal with the problem of the cell phone being held in different hands. Figure 5.13 provides an overview of the training process.

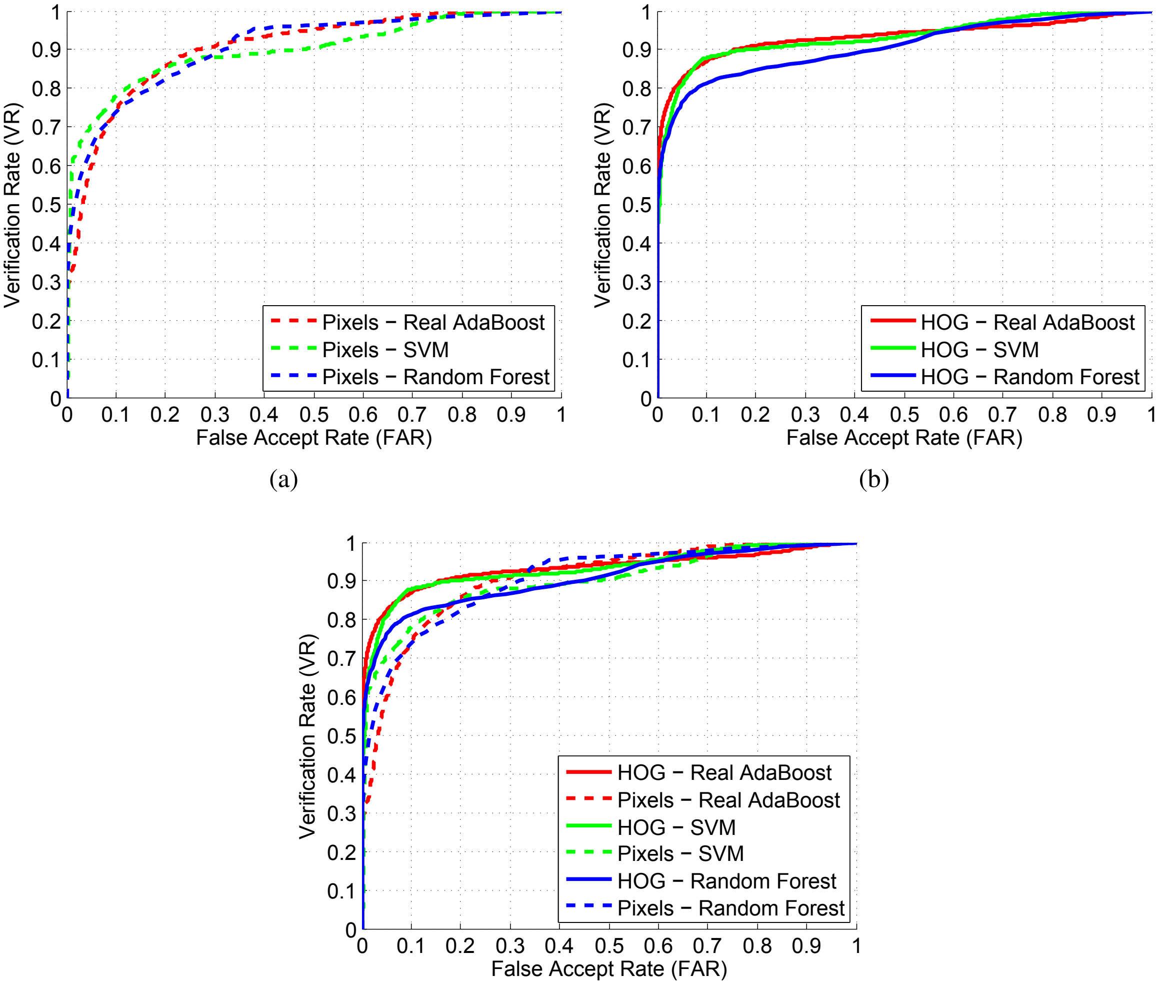
Testing Stage

During the test stage of our algorithm, a similar set of steps to those previously described in the training stage were used to extract region of interests in an input frame to determine if a cell phone was present in the extracted regions. Again, we utilized our facial alignment algorithm to localize facial landmarks and generate two crops on the right and left sides of the face in order to check for cell phone presence. Features extracted from these crops were classified using the appropriate (right or left side) side classifiers and the frame was labeled as not having a cell phone present only if both classifiers returned a negative result while in all other cases it was labeled as containing a cell phone. Figure 5.14 illustrates the sequence of steps followed during the test stage in order to determine if the subject in a test frame is using a cell phone or not.

5.5.3 Results

Our training data for cases when a cell phone was not in use (negative class data) consisted of $1,479$ frames obtained from 30 video segments of 11 subjects. We also used 489 frames obtained from 20 video segments of the same 11 subjects where the subjects were using a cell phone. Only one of the subjects (10 video segments and 137 frames) used his/her left hand to hold the cell phone while in all other cases the right hand was used to hold the cell phone. This was reflection of the skew in the data collected as only a few subjects used their left hand to hold a cell phone when requested to do so. This data was used to extract normalized pixel and HOG feature descriptors and build classifier models. Our Real AdaBoost ensemble was built using 100 weak decision trees of depth 2 and implemented using an open source toolbox [137]. We used 100 trees in our random forest classifier that was again implemented using open source code [194]. Finally, we used the LIBSVM library to build an SVM classifier [195], [196].

Our test data consisted of $8,824$ video frames of 30 subjects in which the subjects were driving a car or seated in a stationary one and not using a cell phone and a corresponding set of $2,503$



(a) (b) (c)

Figure 5.15: Receiver Operating Characteristic (ROC) curves obtained using three classifiers and (a) raw pixels as features, (b) HOG features, and (c) both raw pixels and HOG features.

frames in which the same subjects were using a cell phone. Thus, the total number of test frames was $11,327$, making our study more comprehensive than the one carried out in [189]. Only two subjects held a cell phone in their left hand in a total of 421 frames out of the $2,503$ frames in which a cell phone was being used. It must be noted that there was no overlap of subjects, and hence video frames, between the training and test data used in our study.

Figure 5.15 shows the ROCs obtained using the various classifiers and feature extraction techniques and Table 5.7 summarizes the key results obtained as part of our study. As can be seen,

Table 5.7: A summary of our cell phone detection results. The Verification Rates (VRs) at various False Accept Rates (FARs), Equal Error Rates (EER), Area Under the ROC Curve (AUC), and the classification accuracy rates obtained are listed for each feature extraction technique and classification algorithm combination. The best values for each evaluation metric are indicated in bold text.

Approach	VR @ 0.1% FAR (%)	VR @ 1% FAR (%)	VR @ 10% FAR (%)	EER	AUC	Accuracy (%)
Pixels – Real AdaBoost	15.90	39.19	74.47	0.171	0.905	79.69
HOG – Real AdaBoost	**38.87**	**70.83**	86.90	0.119	**0.931**	**91.23**
Pixels – SVM	0.40	55.85	77.87	0.168	0.898	75.90
HOG – SVM	33.56	61.57	**87.81**	**0.116**	0.930	78.12
Pixels – Random Forest	26.61	45.31	74.51	0.190	0.906	72.16
HOG – Random Forest	37.63	62.21	81.30	0.168	0.906	90.09

HOG features provided a more robust representation and resulted in higher classification accuracy rates, Area Under the Curve (AUC) values, and higher Verification Rates (VRs) at various False Accept Rates (FARs) for all three classifiers with the combination of AdaBoost and HOG features resulting in the highest classification accuracy of 91.20%. Thus, our results are promising and competitive with those obtained in similar studies carried out by Artan *et al.* [189] (highest classification accuracy of 86.19%) and Zhang *et al.* [190] (highest classification accuracy of 91.20%), although it must be noted that each study utilized different training and testing data. However, our study is far more thorough than the previously mentioned ones in that our tests are carried out over a much larger set of images and also in the choice of data used for evaluation, which was acquired using strict protocols by a government agency for a specific purpose. It is our hope that presenting our findings will be of use to the research community and further aid in the development of systems aimed at addressing this problem.

5.6 Concluding Remarks

The hazards associated with driver distraction have been studied in great detail over the past few years. This has motivated several research efforts aimed at developing algorithms and systems capable of automatically detecting events associated with dangerous driving, such as the use of cell phone by a driver. We have described a framework that extends our facial alignment algorithm to allow for accurate facial landmark localization in challenging low-resolution SHRP2 face view videos from the recently released HPV dataset that was acquired for of a study on naturalistic driving behavior. The data is relatively new and has been the focus of only recent studies, such as in [186], making our work quite relevant. The key element in our approach is the fact that our alignment algorithm provides feedback on the goodness of fit of each frame which allows for appropriate initialization for the next frame. Our alignment algorithm was evaluated and benchmarked against a commercial face detection and landmark localization algorithm (the PittPatt algorithm), that served as a baseline, on the HPV SHRP2 face view videos.

Our facial alignment also provided a foundation for the gaining of information, such as head pose estimates and region of interest determination for detection of cell phone usage, etc. Results obtained on these closely allied tasks have also been presented in this chapter. A paper that provides a preliminary version of our findings on cell phone usage detection has been published [20]. Future work in this area could involve using the landmark localization results to also determine if a driver's seat belt is in use or to aid in the determination of the state of mind of a driver (distracted, drowsy, *etc.*).

The facial landmark tracking algorithm we have presented in this chapter is one that made no assumptions about the video sequences and could be used for more general videos that also involve camera motion, scale changes (due to zooming in or out), or scene changes. More accurate results could potentially be obtained if more information were known about the videos and if valid prior assumptions could be made. For example, an on-line training mechanism to develop and update person specific models using accurately processed frames could also be investigated. Additionally, if the goal is only to annotate video sequences and not a system aimed at real-time video processing, then multiple passes of the sequences could be made and previously poorly fitted frames or frames where no landmarks could be localized could be corrected and higher accuracies obtained. Finally, superior tracking performance could also be obtained by more tightly coupling a face detection and tracking process with the landmark localization step.

Another area for future work could include researching the improvements and optimizations that could be made to the framework to better exploit the parallelization of intermediate steps using GPUs. A suitably developed system could be of even greater use in automatically annotating video data sets, as was the goal of our work, or for deployment in a real-world scenario to monitor drivers and aid in decreasing the number of car crashes due to distracted driving.

Chapter 6

Facial Alignment on Low-Resolution Images

“Let’s run this through video enhancement and bring up the ridge detail, okay?”

CSI: Miami - Spring Break (Season 1 *Episode* 21*)*, and a recurring theme on the show

So far our work has dealt with real-world images exhibiting pose, expression, and illumination variations, and varying levels of facial occlusion. While the joint presence of these factors poses a great problem to the face community, the challenge takes on an altogether different dimension when image resolution is very low.

There has been some prior work on facial alignment of frontal low-resolution facial images. Liu *et al.* [197] built a multi-resolution AAM at various scales of facial size and used the most appropriate model (with a model resolution slightly higher than the facial resolution) to fit low-resolution faces (of varying resolution) in a few video sequences. Dedeoğlu *et al.* [198] proposed a Resolution-Aware Formulation (RAF) that modified the original AAM fitting criterion in order to better fit low-resolution images and used their method to fit 180 frames of a video sequence. Qu *et al.* [199] extended a traditional CLM to a multi-resolution model consisting of a 4-level patch pyramid and also used various feature descriptors to construct the patch experts. They compared their approach (using various feature descriptors) against a baseline CLM approach on downsampled

35×35, 25×25, and 15×15 facial images from a few databases, such as the MPIE database, and demonstrated acceptable landmark localization accuracies on the low-resolution faces. Recently, Asthana *et al.* [200] (supplementary material) provided results on the performance of a few facial alignment algorithms on the AFW, LFPW, and Helen datasets when downsampling factors of 2, 4, 6, and 8 were used on the test images. However, since the resolutions of the images and individual faces in these datasets vary dramatically (images in the AFW and LFPW datasets of smaller size than those in the Helen dataset), it is difficult to isolate key results from their findings. Thus, in our work we adopt a more systematic approach and carry out a detailed study that also stress tests facial alignment algorithms by examining their performance on low-resolution images exhibiting yaw and expression variations and varying levels of facial occlusion.

In this chapter, we present an experiment-centric approach to understand the challenges that the resolution problem poses to facial alignment, especially when multiple degradations occur in a single image and no prior information is available regarding their presence. Such a situation routinely arises in law enforcement and a reliable automated facial landmark localization is a key pre-processing step that is required in such cases. We demonstrate how our algorithm can provide acceptable alignment results on such images by using resolution-specific texture models. While there is significant scope for carrying out future work aimed at dealing with such challenging images, it is our aim to take a step in this direction by focusing on some of the the most difficult scenarios and by providing insights gained from our experimental results that we hope may be of use in the future to designing better and truly all-purpose facial alignment algorithms.

6.1 Facial Alignment on Low-Resolution Images using a Single Resolution Model

As a preliminary step towards understanding the challenge posed by low-resolution images to the facial alignment process, we carried out an experiment in which our facial alignment approach

Table 6.1: Details on the test sets used in our experiment on facial alignment on low-resolution images using a single resolution model. The facial region refers to a square region around the convex hull of the ground truth facial landmark coordinates.

Test Set	Number of Images	Original Image Resolution (Pixels)	Facial Region Size (Pixels)	Yaw Variation	Expression Variation	Facial Occlusion
FERET	$1,800$	256×384	140×140	$-40°$ to $+40°$	No	No
AR	$1,340$	768×576	250×250	$-5°$ to $+5°$	No	Yes

and the RCPR algorithm [11], [12], which served as a baseline during benchmarking, were trained on an identical set of images and then tested on progressively downsampled images from various databases in order to provide an initial idea of accuracies that could be obtained when fitting low-resolution images using a single model that was not trained on such images. In addition to this, the experiment also provided some knowledge of the cross-resolution tolerance of such models and an idea of the limit to which such models could be utilized to produce acceptable landmark localization results, *i.e.*, what resolution level could cause excessively large errors to occur. We first provide details on the datasets used in this study in section 6.1.1 and then go on to describe the experiment carried out and the results obtained in section 6.1.2.

6.1.1 Test Sets Used

Details on the various test sets which were used in our experiments are provided below and summarized in Table 6.1.

(1) FERET: The Facial Recognition Technology (FERET) database [201], [202], [203], [204] consists of $14,051$ eight-bit grayscale images of human heads with views ranging from frontal to left and right profiles (yaw range from $-90°$ to $+90°$) and varying illuminations and expressions. A set of $1,800$ images of 200 subjects with neutral expressions and yaw variation from $-40°$ to $+40°$ with 79 manually annotated landmarks, using the annotation scheme shown in Figure 4.1, were available to us [18], [28], [29] and were used in our experiments.

(2) AR: The AR database [82], [83] contains over $4,000$ color images of 136 subjects (76 men

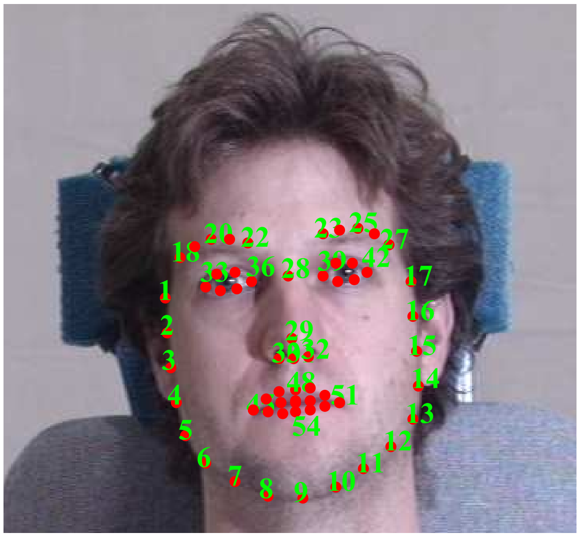


Figure 6.1: An image from the MPIE database showing the locations of the 64 landmarks that were used in our quantitative evaluations in this chapter.

and 60 women) collected over 2 sessions (separated by 14 days) under varying illumination conditions with the subjects showing varying expressions and sometimes wearing a scarf or sunglasses (facial occlusions). It must be noted that out of the 136 subjects, only 116 were acquired participated in both sessions, with 26 images acquired for these subjects and fewer images acquired for the remaining subjects. All these images were purely frontal ones and contained very minor yaw variation. In addition to these still images, a total of 30 sequences of images composed of 25 images each were acquired to test dynamic systems. These images contained pose variation as well and were not a part of our study and are generally not focused on when dealing with the AR database. A set of 1,340 images of 134 subjects (75 men and 59 women) exhibiting neutral expressions and wearing a scarf, sunglasses, or neither with 79 manually annotated landmarks, using the annotation scheme shown in Figure 4.1, were available to us [18] and were used in our experiments.

By testing the alignment algorithms on these downsampled versions of these images, we were simulating extremely challenging conditions as multiple variations/degradations were now present in the test images (yaw variation and low-resolution image artifacts for the FERET images and

Table 6.2: Performance of our approach and RCPR using a single resolution model on the FERET test set images with various downsampling factors.

Downsampling Factor	Facial Region Size (Pixels)	Algorithm					
		Ours			RCPR		
		MFE	MNFE (%)	Failure (%)	MFE	MNFE (%)	Failure (%)
1	140 × 140	6.24	7.13	5.6	9.59	11.13	24.4
2	70 × 70	3.12	7.14	5.8	4.64	10.71	23.4
4	35 × 35	1.64	7.51	9.2	2.31	10.61	22.9
8	18 × 18	1.02	9.39	27.9	1.48	13.59	46.4
16	9 × 9	1.16	21.90	73.8	1.31	23.77	99.9

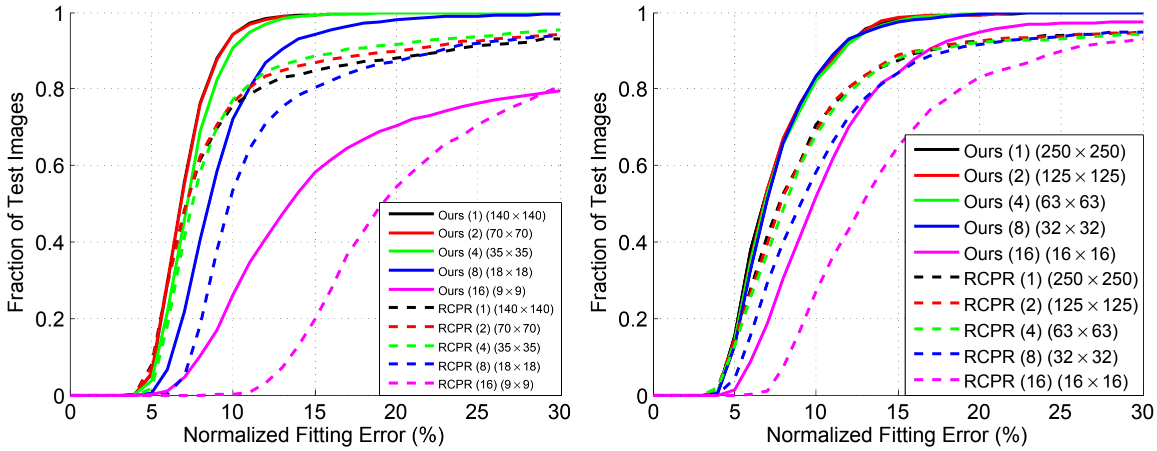
Table 6.3: Performance of our approach and RCPR using a single resolution model on the AR test set images with various downsampling factors.

Downsampling Factor	Facial Region Size (Pixels)	Algorithm					
		Ours			RCPR		
		MFE	MNFE (%)	Failure (%)	MFE	MNFE (%)	Failure (%)
1	250 × 250	11.87	7.48	16.9	16.81	10.59	29.5
2	125 × 125	5.97	7.52	16.9	8.46	10.63	31.1
4	63 × 63	3.02	7.62	17.9	4.30	10.81	31.9
8	32 × 32	1.51	7.62	16.7	2.29	11.53	41.7
16	16 × 16	1.11	11.21	48.1	1.53	15.40	73.1

occlusions and low-resolution image artifacts for the AR images).

6.1.2 Results

Our approach and the RCPR algorithm were trained on the same set of MPIE images previously used to train our models in section 3.2.1. However, since, the facial images in the AR and FERET test sets did not exhibit any expression variation, both RCPR and our approach were not trained on images exhibiting an open mouth expression (scream or surprise). Also, since the images in our test sets did not exhibit an absolute yaw in excess of 45°, both approaches were configured to always output a set of 68 landmarks that could be compared against ground truth coordinates of a set of 64 landmarks that could be obtained from both the 68 point and the 79 point annotation



(a)

(b)

Figure 6.2: Cumulative Error Distribution (CED) curves at various downsampling factors for our approach and RCPR using a single resolution model on the (a) FERET and (b) AR test sets. The downsampling factors and facial region sizes are indicated in brackets in the legends for (a) and (b).

schemes. The locations of these 64 landmarks that were used in our quantitative evaluations are shown in Figure 6.1.

The images in each of the test sets were progressively downsampled by a factor of 2 to synthesize low-resolution images that could now be tested on. The downsampling factors used were 1, 2, 4, 8, and 16. These low-resolution images were used as input to both alignment algorithms and landmark localization proceeded in a similar fashion to when high-resolution images were used as input (with changes made to a few of the search parameters in our approach to ensure the optimal results for each downsampling factor). This involved providing the appropriate initialization using crops that matched the training crops extracted from around the ground truth coordinates, and the subsequent resizing of the region of interest by the algorithms in order to extract appropriate features. For our alignment algorithm, this amounted to resizing of the low-resolution crop (obtained by growing a crop around the ground truth coordinates by a factor of 1.5) to a standard 100×100 region with all processing carried out on this resized crop before the final coordinates were scaled back to correspond to the original low-resolution image. For RCPR, the low-resolution crop (obtained by growing a crop around the ground truth coordinates by a factor of 1.15 to match the same process that was used at the training stage) was used as input to the implementation of the algorithm.

The progressively lower resolutions posed a severe challenge to both facial alignment approaches that were trained on higher resolution images. Tables 6.2 and 6.3 summarize the landmark localization error values that were obtained by both approaches on the FERET and AR test sets, respectively, using the same metrics (Mean Fitting Error (MFE), Mean Normalized Fitting Error (MNFE), and failure percentage) as in section 4.2, with normalized fitting errors computed by normalizing using the distance between the corners of the eyes in the ground truth images, the same normalization technique that was used in section 3.2.3 and section 4.2. Cumulative Error Distribution (CED) curves summarizing the performance of the approaches on the test sets can be found in Figure 6.2.

From tables 6.2 and 6.3 it can be seen that both approaches exhibited high MNFE values at extremely low resolutions. This was to be expected, however, both approaches also exhibited a tolerance to resolution effects until a certain downsampling factor was used that resulted in the texture models (constructed using MPIE images with a facial region, a square region around the convex hull of the ground truth facial landmark coordinates, approximately of size 160×160) being no longer able to model the texture signature manifested by the images. On the FERET test set, a significant increase in the fitting error values was observed for a downsampling factor of 8, while the corresponding increase occurred on the AR test set for a downsampling factor of 16. This is because the AR images have higher resolution than the FERET images to begin with. However, the fitting error values on the AR test set for most downsampling factors were higher than the corresponding fitting errors for the FERET test set due to the occlusions present in the AR images that are not present in the FERET images.

Our approach consistently provided more accurate results than RCPR on both test sets and for all downsampling factors. However, the results obtained when a downsampling factor of 16

was used on the FERET test set were quite poor. This was due to the difficulty in establishing the best fitting yaw specific model after the shape refinement step. We have already alluded to this problem in section 3.2.4 and draw attention to it again at this point because the problem is exacerbated when dealing with such low-resolution images. The local texture model classifiers used by our approach produce unreliable results at such resolutions leading to large landmark localization errors on certain images due to an error at the final stage of the alignment process. If this is corrected for by choosing the refined shape closest to the ground truth coordinates instead of relying on the metric based on the number of inliers, the fitting errors drop in all cases (see section 3.2.4 for details regarding this phenomenon). However, this problem can be alleviated to a certain extent by using resolution-specific texture models, as we demonstrate in sections 6.2 and 6.3.

6.2 Facial Alignment on Low-Resolution Images using Resolution-Specific Models

In this experiment, we aimed at determining how facial alignment on low-resolution images could be addressed by using resolution-specific texture models. The setup for this experiment was identical to that of the previous experiment except for this aspect. Our approach and the RCPR algorithm were trained on images (the same set of MPIE images, containing no faces with open mouth expressions, that were used for training in the previous experiment were used in this one as well) by progressively downsampling them by factors of 1, 2, 4, 8, and 16, resulting in facial region crops that were approximately of size 160×160, 80×80, 40×40, 20×20, and 10×10, respectively. These crops were resized to a standard 100×100 region for building the local texture models for our approach. Thus, resolution-specific texture models were constructed to match the texture in the test images that were generated using the same downsampling factors on the previously described FERET and AR test sets. This ensured that there was a slightly better match between the texture models constructed during the training stage and the texture in the synthesized low-resolution test

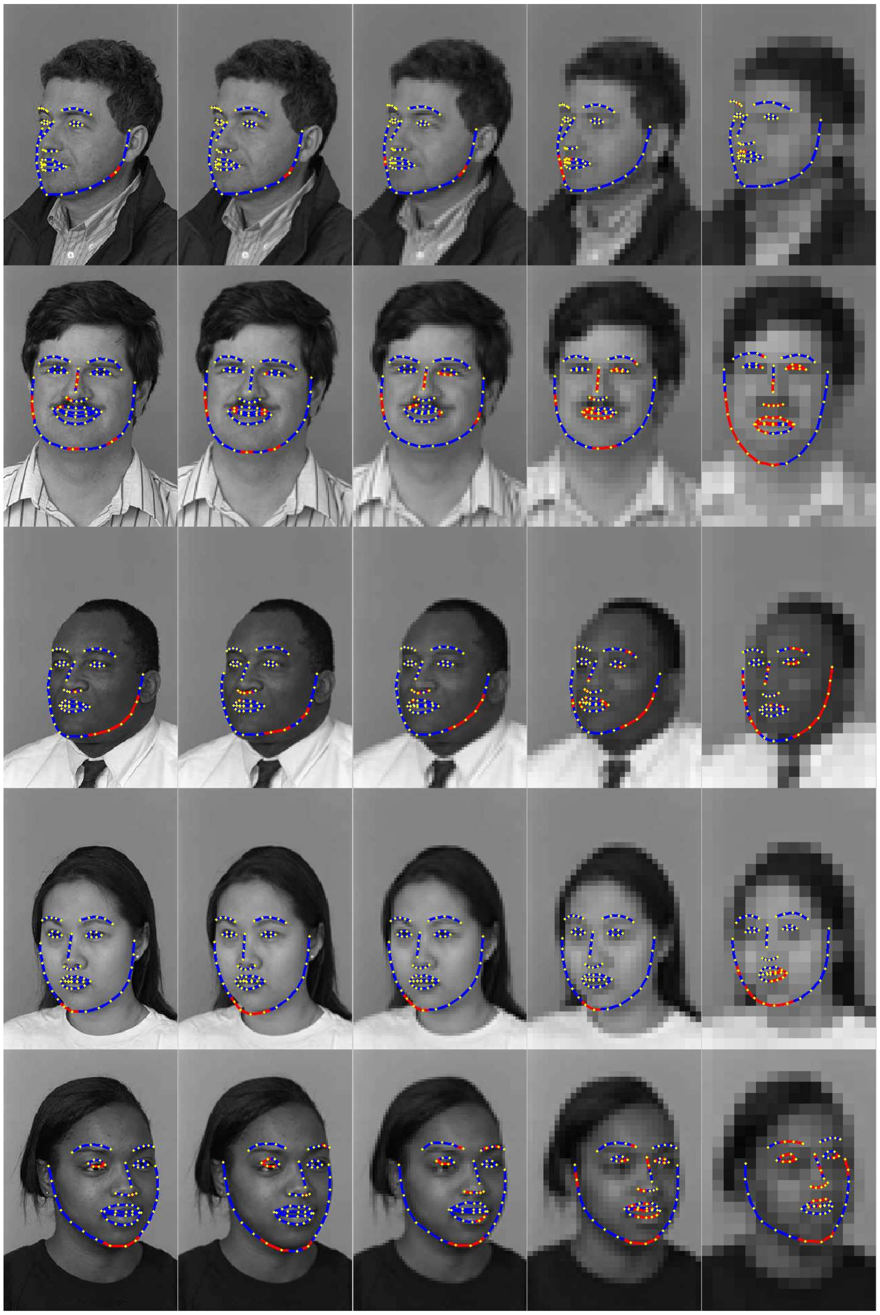


Figure 6.3: Qualitative landmark localization results produced by our approach using resolution-specific models on some images from the FERET test set. In each row the downsampling factors used (from left to right) are: 1, 2, 4, 8, and 16. In all facial images with landmarks overlaid on them, yellow dots are used to indicate the locations of facial landmarks, blue line segments indicate that the landmark at their center is accurately localized, and red line segments indicate that the landmark at their center is misaligned or potentially occluded. The same color scheme is maintained in all figures that show facial alignment results produced by our approach in this chapter.

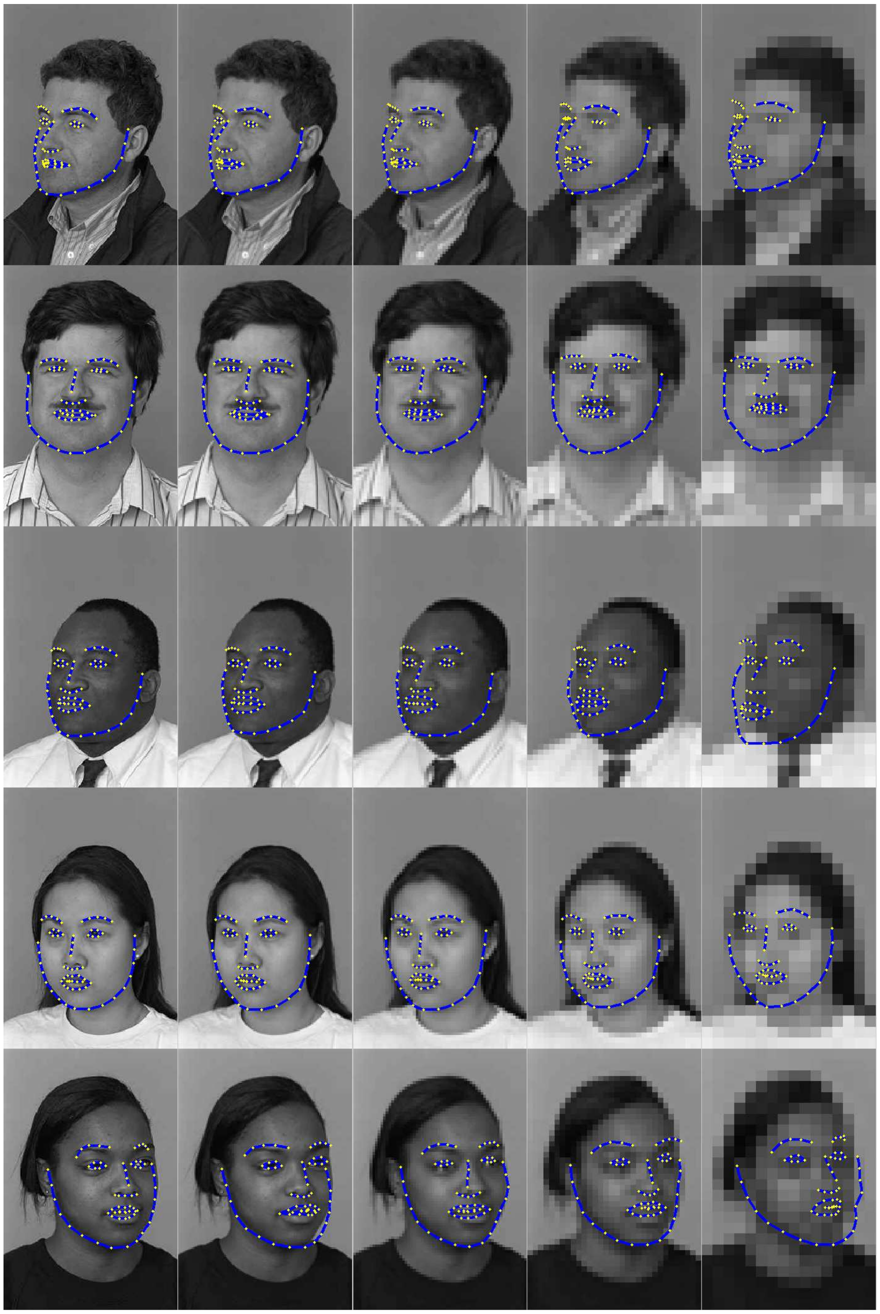


Figure 6.4: Qualitative landmark localization results produced by RCPR using resolution-specific models on some images from the FERET test set. In each row the downsampling factors used (from left to right) are: 1, 2, 4, 8, and 16. In all images with landmarks overlaid on them, yellow dots are used to indicate the locations of facial landmarks and blue line segments connect them (occlusion labels are not provided by RCPR in this case). The same color scheme is maintained in all figures that show facial alignment results produced by RCPR in this chapter.

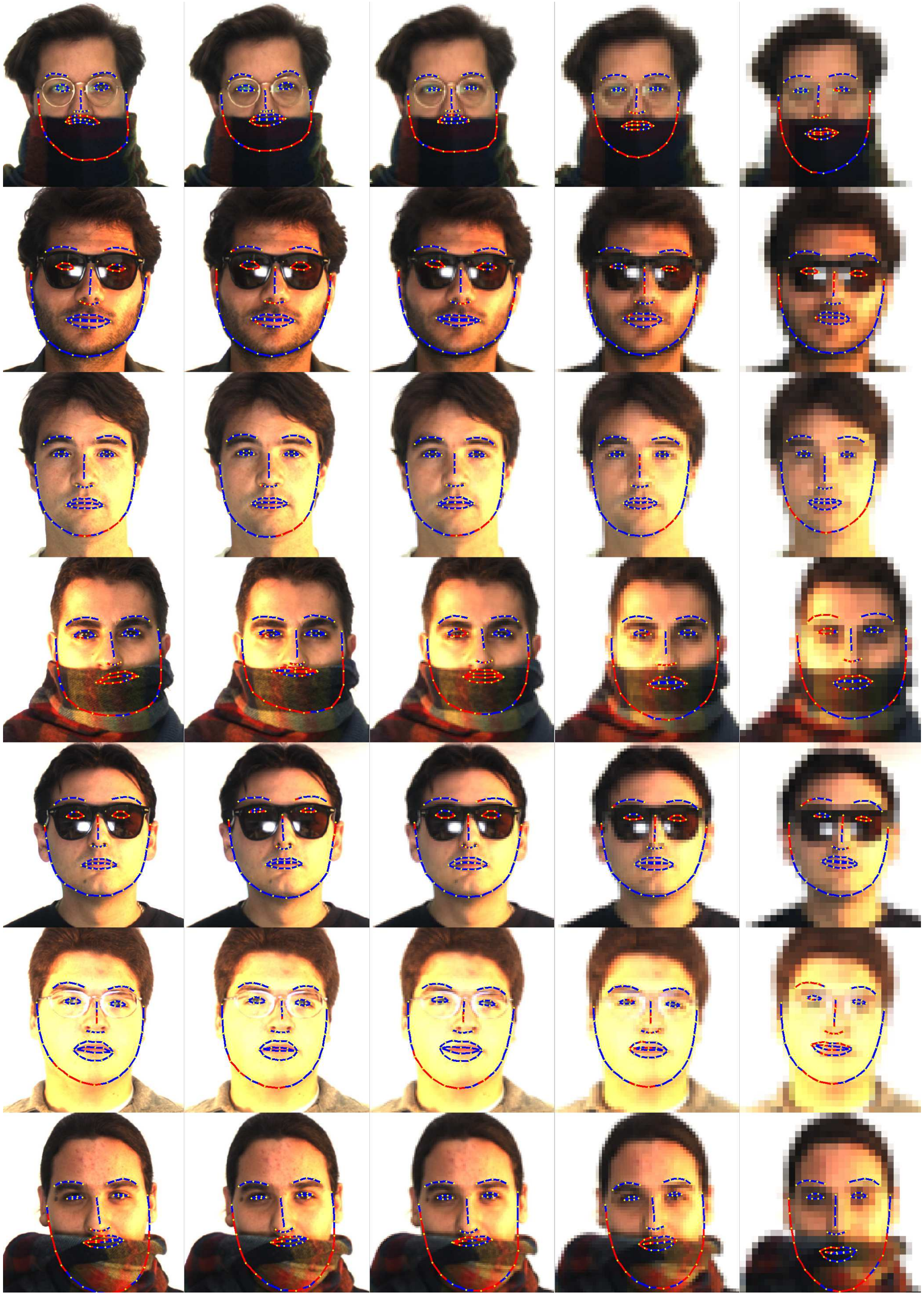


Figure 6.5: Qualitative landmark localization results produced by our approach using resolution-specific models on some images from the AR test set. In each row the downsampling factors used (from left to right) are: 1, 2, 4, 8, and 16.

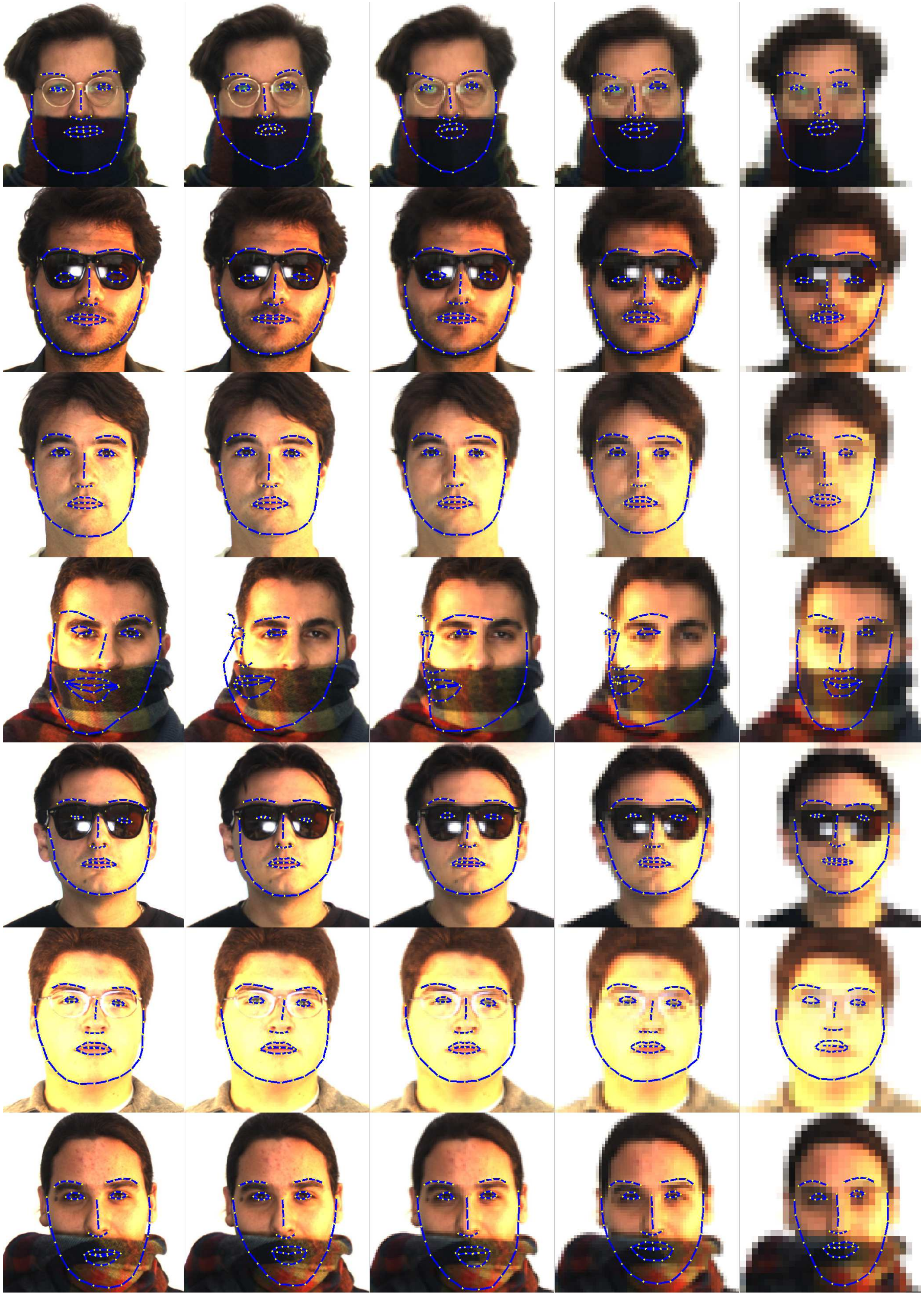


Figure 6.6: Qualitative landmark localization results produced by RCPR using resolution-specific models on some images from the AR test set. In each row the downsampling factors used (from left to right) are: 1, 2, 4, 8, and 16.

Table 6.4: Performance of our approach and RCPR using resolution-specific models on the FERET test set images with various downsampling factors.

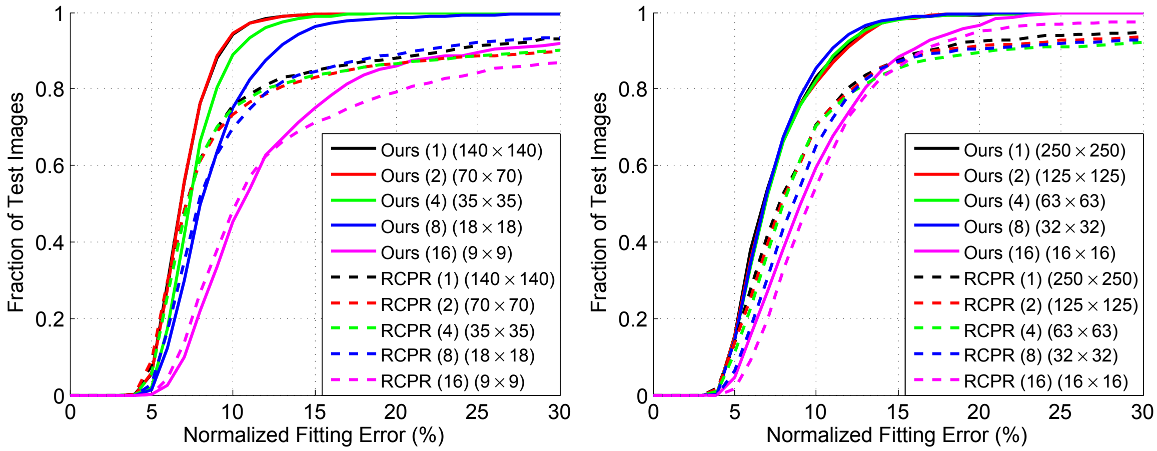
Downsampling Factor	Facial Region Size (Pixels)	Algorithm					
		Ours			RCPR		
		MFE	MNFE (%)	Failure (%)	MFE	MNFE (%)	Failure (%)
1	140 × 140	6.24	7.13	5.6	9.59	11.13	24.4
2	70 × 70	3.11	7.11	5.5	5.27	12.23	26.7
4	35 × 35	1.69	7.72	10.9	2.64	12.29	25.0
8	18 × 18	0.99	9.05	25.0	1.29	11.94	30.2
16	9 × 9	0.76	14.03	54.6	0.93	17.07	51.7

Table 6.5: Performance of our approach and RCPR using resolution-specific models on the AR test set images with various downsampling factors.

Downsampling Factor	Facial Region Size (Pixels)	Algorithm					
		Ours			RCPR		
		MFE	MNFE (%)	Failure (%)	MFE	MNFE (%)	Failure (%)
1	250 × 250	11.87	7.48	16.9	16.81	10.59	29.5
2	125 × 125	6.02	7.59	18.4	8.82	11.11	29.8
4	63 × 63	3.02	7.60	17.8	4.61	11.62	29.6
8	32 × 32	1.48	7.47	14.6	2.35	11.83	35.0
16	16 × 16	0.99	9.94	40.6	1.11	11.14	45.5

images generated using the various downsampling factors. Both approaches were again configured to always output a set of 68 landmarks that could be compared against ground truth coordinates of a set of 64 landmarks and the same initialization and cropping techniques were used at the training and testing stages for both approaches to provide consistent facial region crops that could be resized and processed by the alignment algorithms. Qualitative results produced by our approach and RCPR on some images from the FERET test set using the various downsampling factors are shown in Figures 6.3 and 6.4, respectively. Similarly, qualitative results produced by our approach and RCPR on the AR test set are shown in Figures 6.5 and 6.6, respectively.

Tables 6.4 and 6.5 summarize the performance of our approach and RCPR on the FERET and AR test sets, respectively, and Figure 6.7 shows the CED curves obtained for both approaches and



(a)

(b)

Figure 6.7: Cumulative Error Distribution (CED) curves at various downsampling factors for our approach and RCPR using resolution-specific models on the (a) FERET and (b) AR test sets. The downsampling factors and facial region sizes are indicated in brackets in the legends for (a) and (b).

the two test sets. As can be seen from the results, the landmark localization accuracies obtained for both approaches were higher for downsampling factors of 8 and 16 than those obtained using a single resolution based model. The accuracies obtained for a downsampling factor of 16 though still indicate that the extremely low-resolution of such images poses a great challenge to obtaining low MNFE values as the normalization distance is so small (in terms of pixels) that MNFE values of over 10% become common. However, it must be noted that the ground truth landmark coordinates at these resolutions were not actually obtained by annotating the low-resolution images themselves and that it is extremely hard, even for a human, to consistently annotate such low-resolution images with high accuracy and low variance.

A key point to note in this experiment was that the facial region sizes used during the training stage did not exactly match the sizes of the same regions on the test images. This mimicked real-world conditions when the nearest resolution-specific model would have to be used to annotate a test face of a certain size. The region of interest mismatch was not severe for the FEERT images but was higher for the AR images where the downsampling factors resulted in resolutions that fell approximately midway between one pre-trained resolution-specific model and the next. However, as we demonstrated in section 6.1.2, a certain amount of cross-resolution tolerance was exhibited by the texture models and this was borne out again in this experiment.

The determination of the optimal shape based on a particular yaw specific model after the shape refinement step was again a challenge for our approach on the FERET test set, especially for downsampling factors of 8 and 16. As we have repeatedly mentioned, this step can be the Achilles heal of our approach and is extremely difficult to deal with at low image resolutions. However, in a semi-automated scenario, such as those that routinely arise in law enforcement, where only a limited number of images need to be processed and manual intervention is possible, this could actually be an advantage as an operator could choose from a few alignment result possibilities and pick the most appropriate one, rather than just relying on a single alignment result produced automatically.

The key result to take note of from this experiment and the previous one that we carried out (using a single resolution model) is that reliable facial alignment is possible using our approach for faces approximately of size 16×16 or higher. For reliable facial alignment to be possible at lower resolutions, more assumptions and more accurate initialization would be required and it must also be kept in mind that automatic face detection at such resolutions is not a solved problem either.

6.3 Occlusion Tolerance of Facial Alignment Algorithms on Low-Resolution Images

To simulate the worst case scenario, we carried out an experiment on a set of images with four variations/degradations (pose, expression, facial occlusions, and low-resolution artifacts) simultaneously present. Our approach and RCPR were trained on three-fourths of the full set of MPIE training images (resolution-specific models were constructed for the various downsampling factors) and a test set of 800 images of unseen subjects was drawn from the remaining images. This

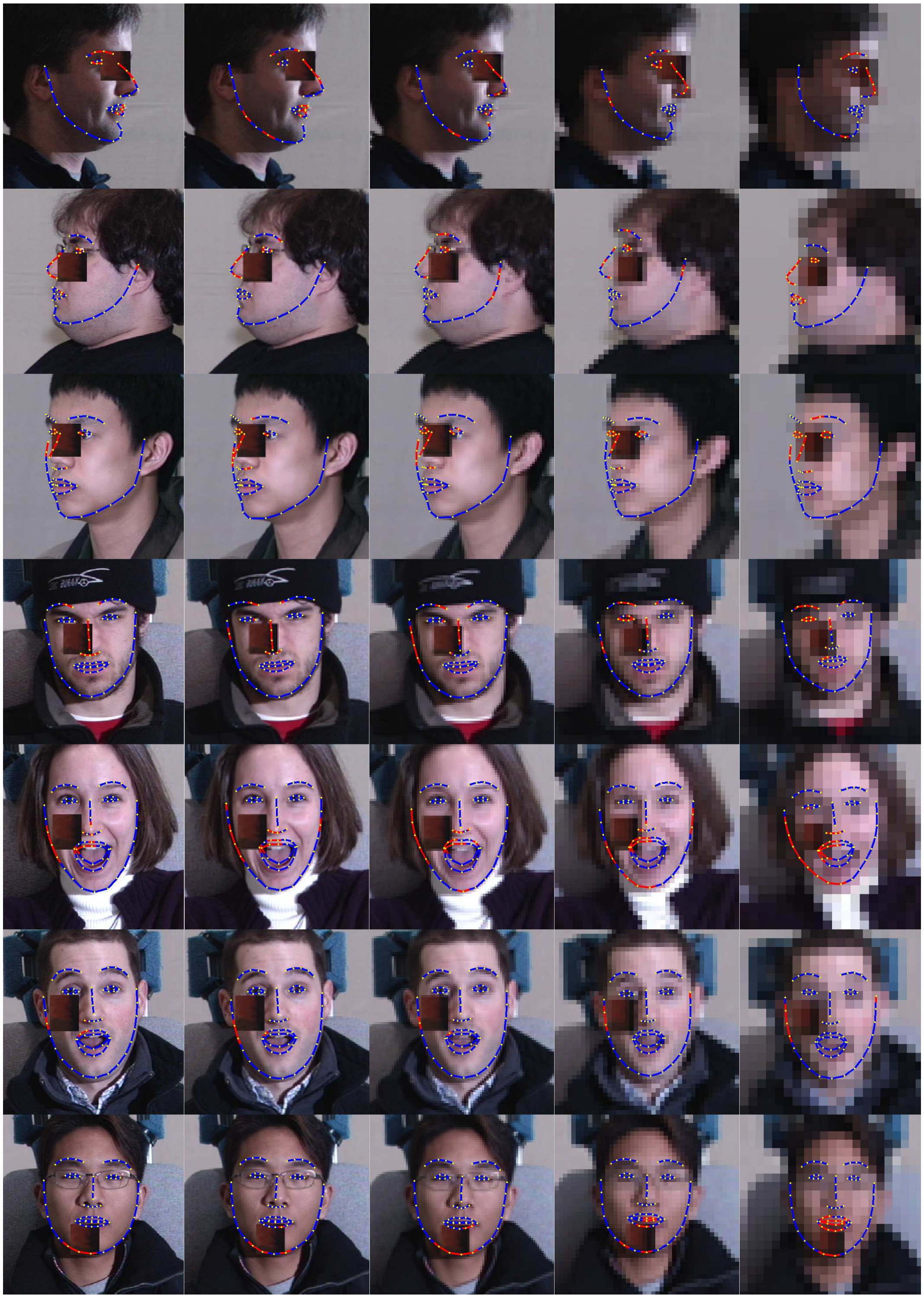


Figure 6.8: Qualitative landmark localization results produced by our approach using resolution-specific models on some images from the MPIE test set. In each row the downsampling factors used (from left to right) are: 1, 2, 4, 8, and 16.

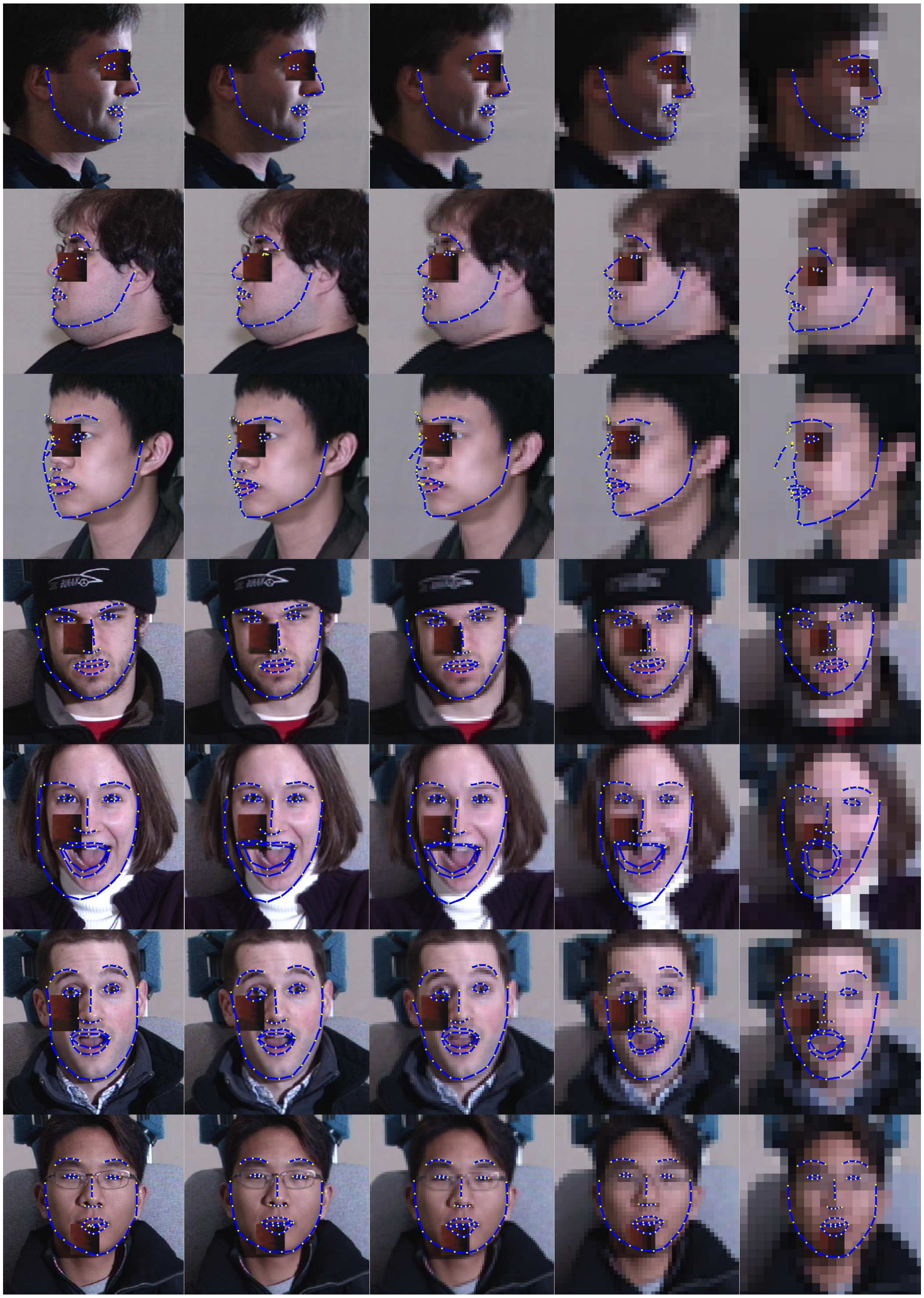


Figure 6.9: Qualitative landmark localization results produced by RCPR using resolution-specific models on some images from the MPIE test set. In each row the downsampling factors used (from left to right) are: 1, 2, 4, 8, and 16.

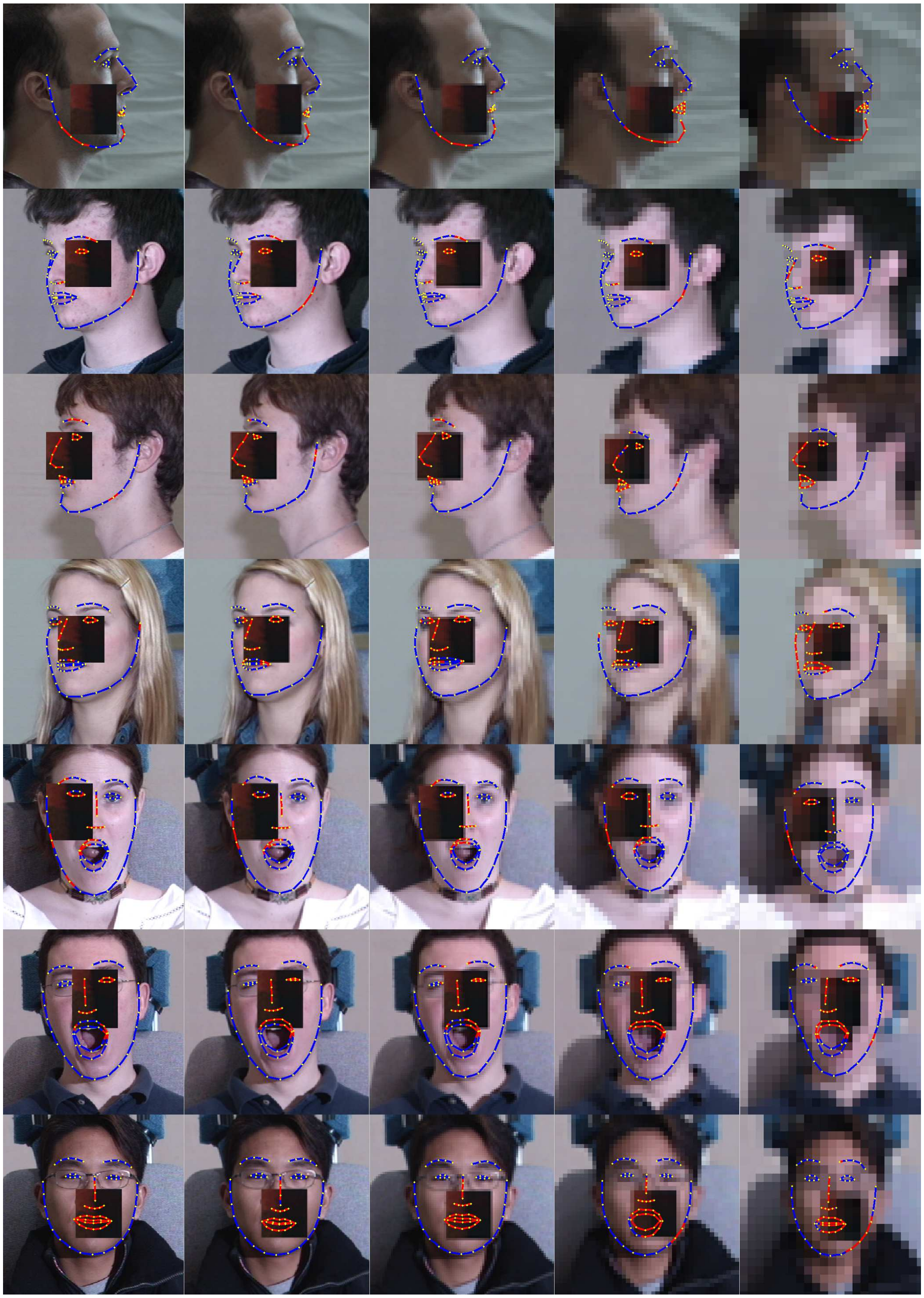


Figure 6.10: Qualitative landmark localization results produced by our approach using resolution-specific models on some images from the MPIE test set with an occlusion level of 25%. In each row the downsampling factors used (from left to right) are: 1, 2, 4, 8, and 16.

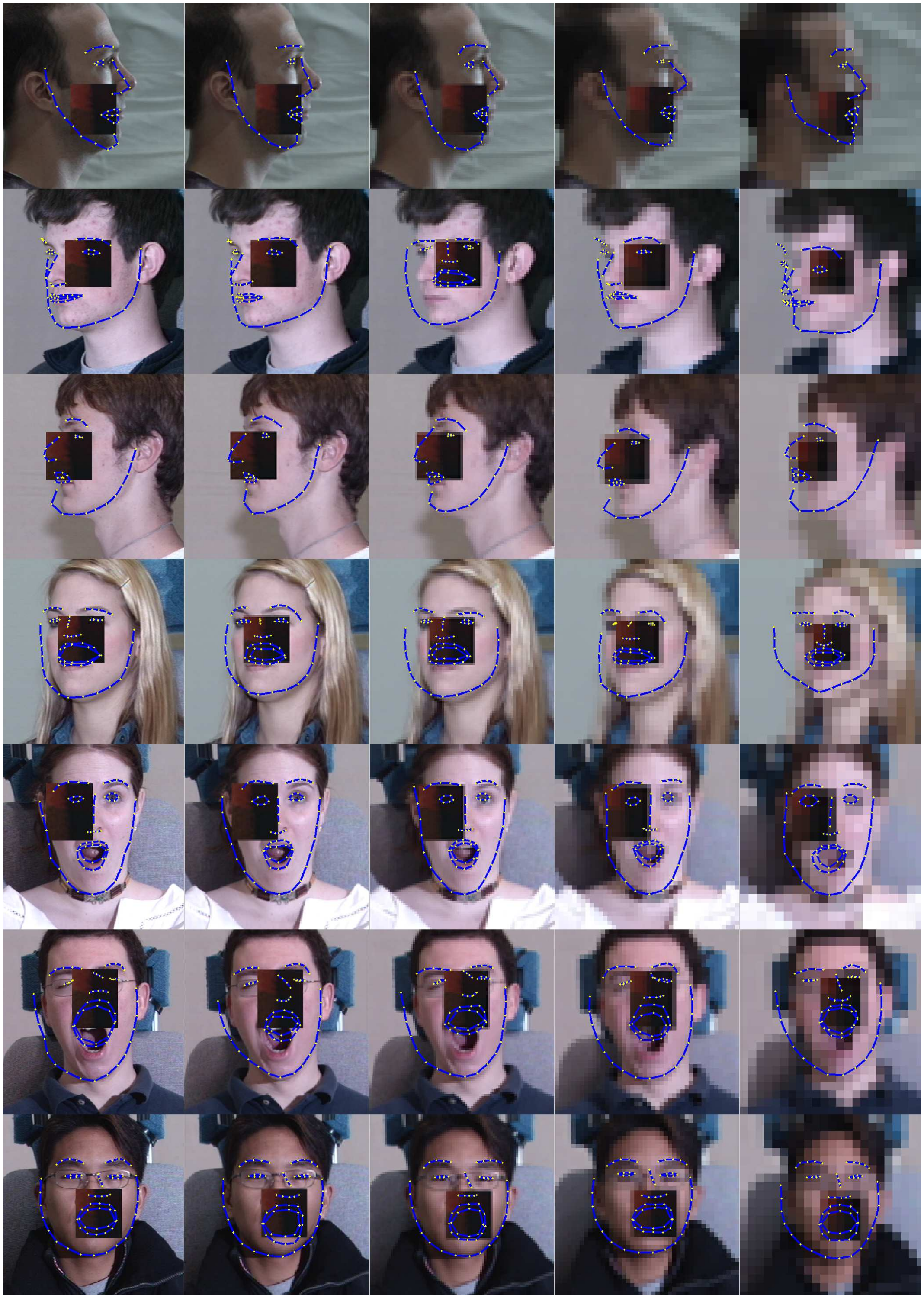


Figure 6.11: Qualitative landmark localization results produced by RCPR using resolution-specific models on some images from the MPIE test set with an occlusion level of 25%. In each row the downsampling factors used (from left to right) are: 1, 2, 4, 8, and 16.

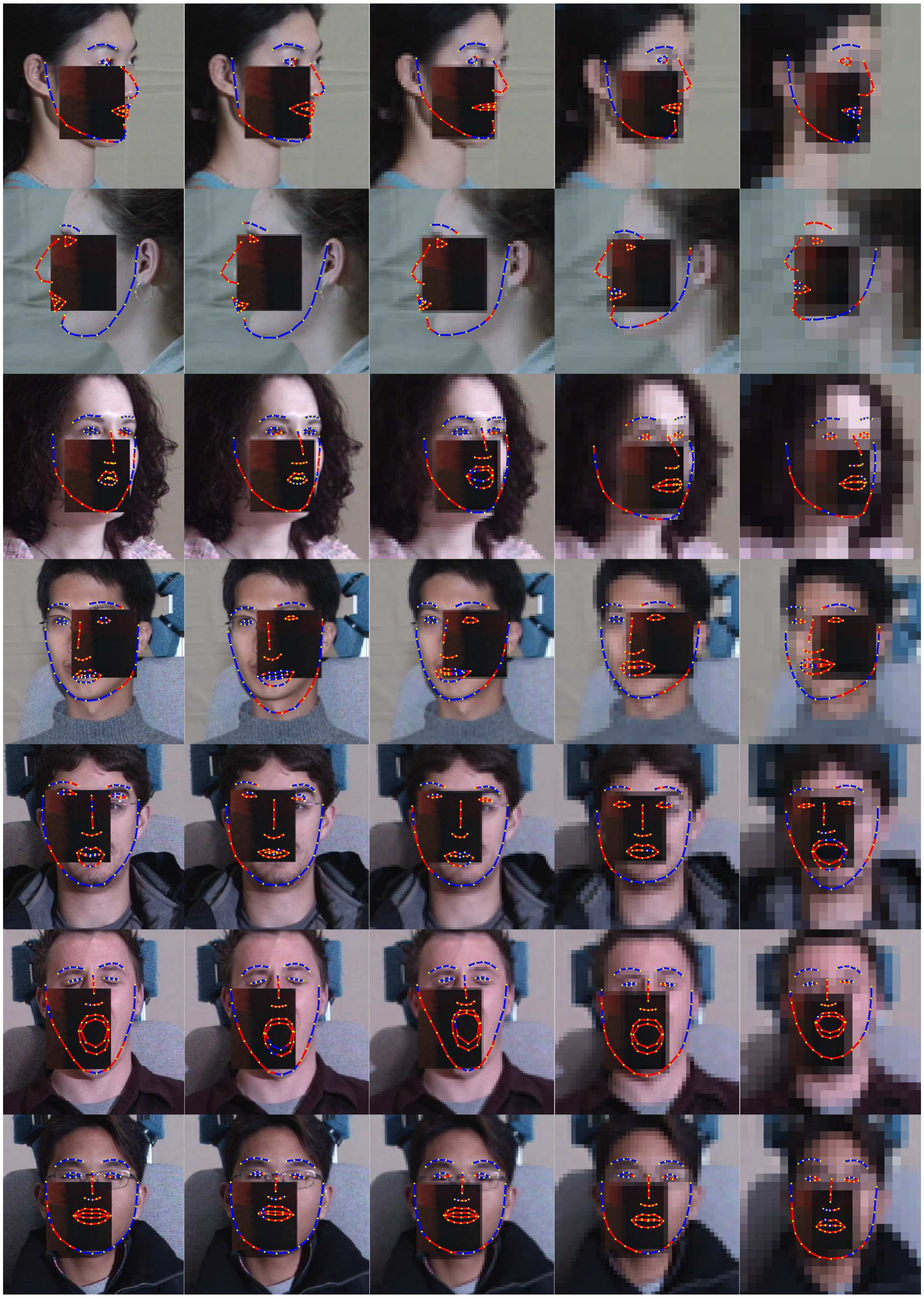


Figure 6.12: Qualitative landmark localization results produced by our approach using resolution-specific models on some images from the MPIE test set with an occlusion level of 50%. In each row the downsampling factors used (from left to right) are: 1, 2, 4, 8, and 16.

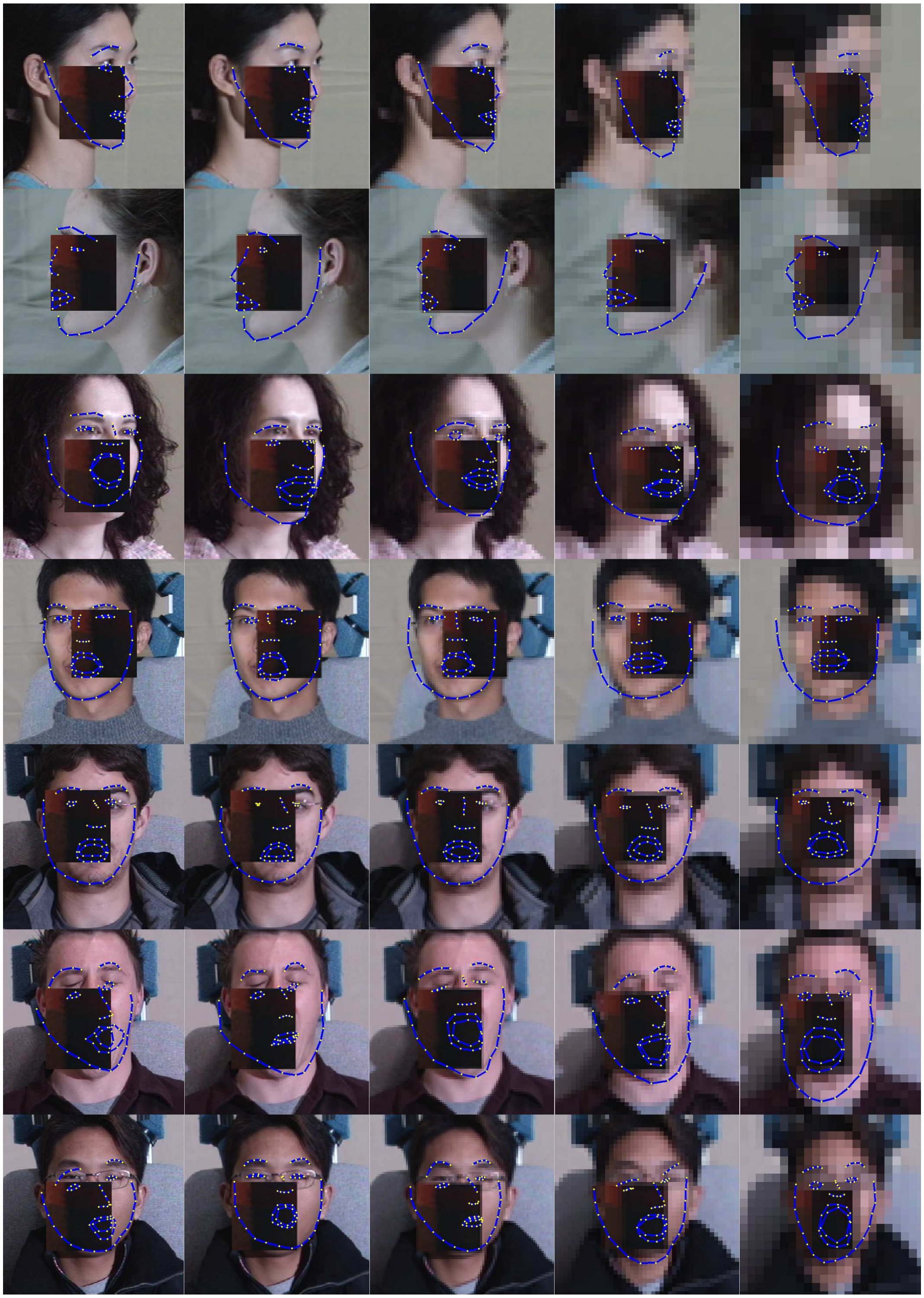
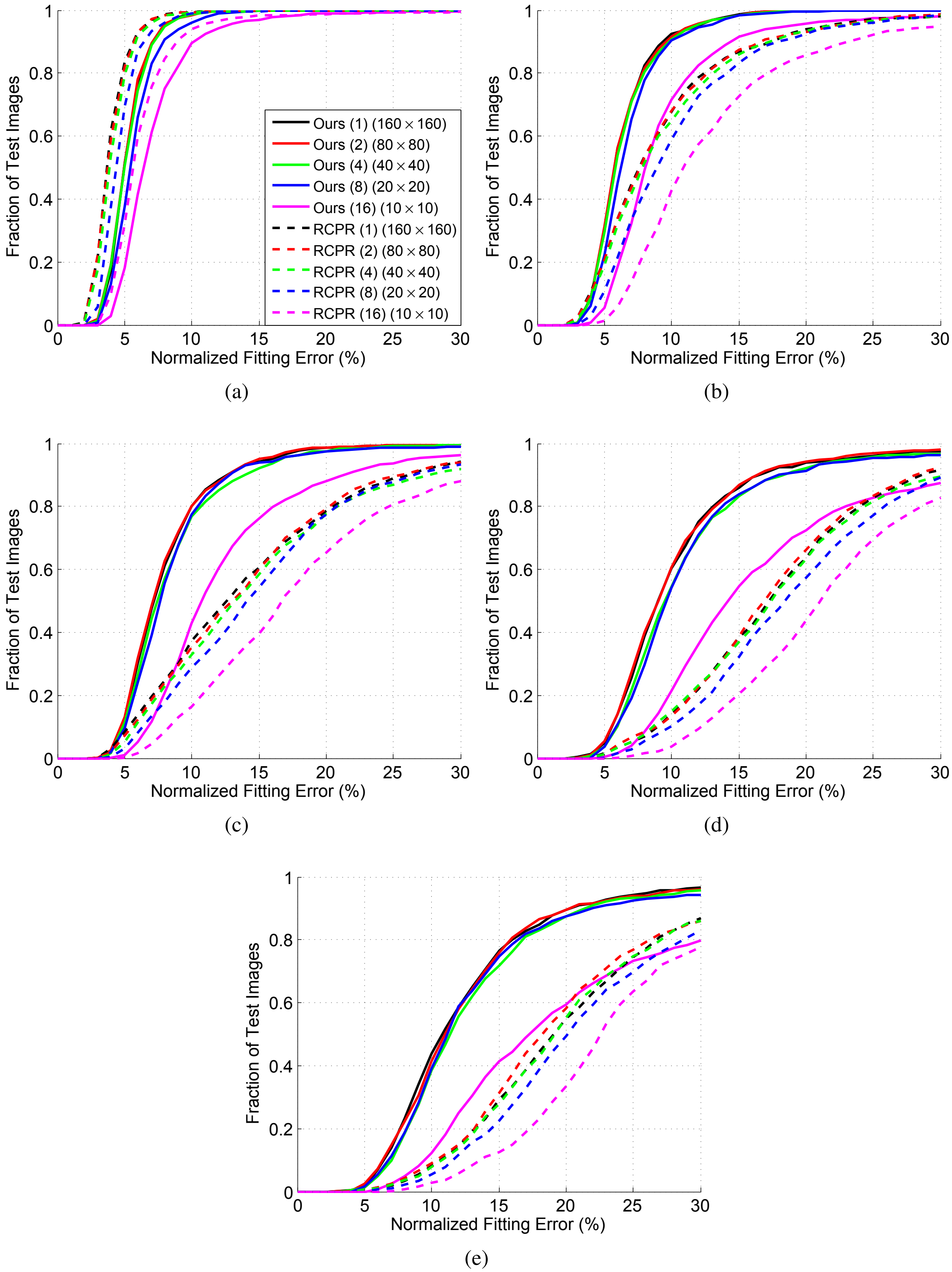


Figure 6.13: Qualitative landmark localization results produced by RCPR using resolution-specific models on some images from the MPIE test set with an occlusion level of 50%. In each row the downsampling factors used (from left to right) are: 1, 2, 4, 8, and 16.

set of images exhibited yaw variation from $-90°$ to $+90°$ as well as expression variation and were also downsampled to create a low-resolution test set. We also went one step further and artificially added occlusions to this set of images by cropping a patch of a scarf region from an AR database image and introducing this at a random location in the facial region of these MPIE images. The occlusion patches were also varied in size to create test sets that covered 10%, 25%, 40%, and 50% of the total facial area in each occluded test set version. This allowed us to evaluate our approach and RCPR on extremely challenging low-resolution images while also determining the effect that the varying occlusion level had on the performance both approaches at various resolutions.

Our approach and RCPR were tested on this set of low-resolution images for the various downsampling factors and occlusion levels using resolution-specific models and the same initialization process that was used in our previous experiments. Ground truth annotations for 39 or 68 landmarks were available for all 800 test set images enabling a straightforward comparison against the automatically localized landmark coordinates. It must be noted that RCPR is not capable of automatically determining whether to annotate a face with 39 or 68 landmarks, *i.e.*, it is not capable of automatic pose estimation for faces exhibiting an absolute yaw in excess of $45°$, unlike our approach. Thus, to ensure a fair comparison and a comparison over all test images, both our approach and RCPR were provided with information regarding the approximate pose range of the subject in each image, *i.e.*, left profile (39 landmarks visible), right profile (39 landmarks visible), and frontal (all 68 landmarks visible and yaw in the range from $-45°$ to $+45°$). Qualitative results produced by our approach on some images from the MPIE test set using the various downsampling factors and occlusion levels are shown in Figures 6.8, 6.10, and 6.12. Corresponding qualitative results produced by RCPR on the same images for the same occlusion levels are shown in Figures 6.9, 6.11, and 6.13.

The results obtained by our approach and RCPR for the various occlusion levels and resolution levels are summarized by Table 6.6 and Figure 6.14, which shows the CED curves obtained for both approaches for the various downsampling factors and synthesized occlusion levels. It



(a)

(b)

(c)

(d)

(e)

Figure 6.14: Cumulative Error Distribution (CED) curves at various downsampling factors for our approach and RCPR using resolution-specific models on the MPIE test set with (a) no added occlusions (0% occlusion level), (b) 10% occlusion level, (c) 25% occlusion level, (d) 40% occlusion level, and (e) 50% occlusion level. The legend in (a) is common to (b), (c), (d), and (e). The downsampling factors and facial region sizes are indicated in brackets in the legend for (a).

Table 6.6: Performance of our approach and RCPR using resolution-specific models on the MPIE test set images with various downsampling factors and artificial occlusion levels.

			Algorithm					
Occlusion	Downsampling	Facial Region	Ours			RCPR		
Level	Factor	Size (Pixels)	MFE	MNFE (%)	Failure (%)	MFE	MNFE (%)	Failure (%)
0%	1	160 × 160	4.05	5.22	1.3	3.07	3.96	0.6
	2	80 × 80	2.02	5.19	1.6	1.55	3.99	0.8
	4	40 × 40	1.02	5.25	1.4	0.81	4.15	0.6
	8	20 × 20	0.56	5.75	4.0	0.45	4.66	1.0
	16	10 × 10	0.36	7.31	10.4	0.31	6.36	6.1
10%	1	160 × 160	4.98	6.40	7.6	7.38	9.43	32.4
	2	80 × 80	2.49	6.41	8.5	3.70	9.44	32.0
	4	40 × 40	1.26	6.50	9.3	1.90	9.73	35.3
	8	20 × 20	0.66	6.85	9.6	1.02	10.50	41.2
	16	10 × 10	0.47	9.74	28.5	0.65	13.31	57.3
25%	1	160 × 160	6.37	8.16	19.9	11.47	14.62	62.7
	2	80 × 80	3.15	8.08	19.9	5.85	14.90	64.8
	4	40 × 40	1.67	8.59	23.3	3.04	15.48	67.1
	8	20 × 20	0.86	8.83	22.5	1.55	15.80	71.3
	16	10 × 10	0.64	13.08	57.1	0.94	19.32	83.7
40%	1	160 × 160	8.41	10.68	39.6	14.36	18.38	86.1
	2	80 × 80	4.13	10.56	39.4	7.13	18.22	86.5
	4	40 × 40	2.22	11.32	45.4	3.66	18.71	85.2
	8	20 × 20	1.16	11.91	45.9	1.95	19.89	89.9
	16	10 × 10	0.94	19.24	78.5	1.11	22.69	96.3
50%	1	160 × 160	9.92	12.69	56.1	16.13	20.57	92.0
	2	80 × 80	5.10	13.07	58.6	8.01	20.44	91.1
	4	40 × 40	2.67	13.69	61.5	4.07	20.79	92.2
	8	20 × 20	1.36	13.96	61.0	2.18	22.26	94.6
	16	10 × 10	1.14	23.06	87.9	1.20	24.52	97.1

must be noted that as has been our convention throughout this thesis when reporting results for test sets containing faces with an absolute yaw in excess of 45°, the normalization distance used when reporting the MNFE values was the average eye center to mouth corner distance. RCPR obtains higher MNFE accuracies than our approach on the un-occluded test set. However, as the occlusion level increases, our approach outperforms RCPR, which is very accurate when tested on images similar to those in its training set but does not exhibit an ability to generalize to dissimilar test images. As was expected, increased occlusion levels caused a drop in performance for both approaches and all resolutions, however, our approach was able to provide acceptable accuracies until an occlusion level of 40% was reached at most resolutions and downsampling factors. This is quite an important result and again serves to demonstrate the occlusion tolerance of our approach. It must also be kept in mind that the ground truths for the occluded images were actually obtained using the un-occluded images and that a larger variance could creep in if humans were asked to annotate the occluded images. This is an important fact as it means that a wider berth should be provided to alignment approaches when dealing with occluded images that are not synthetically occluded, such as those in the COFW dataset, because of this subjectivity. As was the case with the results obtained on the FERET and AR datasets using resolution-specific models, a downsampling factor of 16, resulting in faces of size 10×10, posed a great challenge to both approaches and caused a significant increase in MNFE values for all occlusion levels from those obtained using a downsampling factor of 8. However, as we have already pointed out, more assumptions regarding initialization could help in alleviating this problem.

6.4 Facial Alignment on Real-World Low-Resolution Images

In this section we describe results that were obtained when RCPR and our approach were tested on real-world low-resolution images that were not synthetically generated using downsampling.

6.4.1 Dataset Used

The dataset used in this experiment was a small in-house one consisting of 15 images of 7 subjects. The images were captured from approximately 325m away on the Carnegie Mellon University (CMU) campus using a Canon EOS60D camera mounted on a tripod and configured with a Canon 800mm telephoto lens with a 1.4x focal length extender. The cropped images with the face roughly centered were approximately of size 500×600 with a facial region of size 180×180 approximately. The images were manually annotated by us with the same previously described 79 landmarks. The

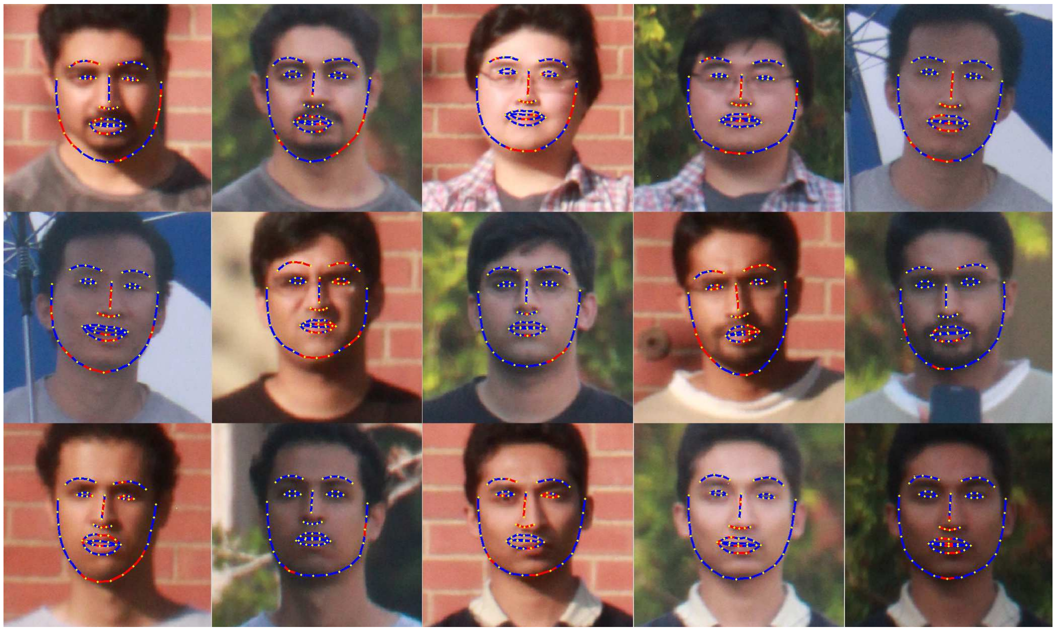


Figure 6.15: Qualitative landmark localization results produced by our approach (models trained on MPIE images using a downsampling factor of 8) on real-world low-resolution images.

images posed quite a challenge to the facial alignment process due to the imaging distance used, the shadows that were sometimes present, and the blurred nature of the images. Thus, the purpose of this experiment was to demonstrate that real-world low-resolution (blurred or out of focus) images could still be effectively dealt with by our facial alignment algorithm, albeit on a small dataset.

6.4.2 Results

RCPR and our alignment algorithm were run on the previously described dataset of real-world images using the same initialization process that was used in our previous experiments in this chapter and were both again configured to always output a set of 68 landmarks that could be compared against ground truth coordinates of a set of 64 landmarks. Downsampling of the images was not required in this case. Qualitative results produced by our approach and RCPR are shown in Figure 6.15 and Figure 6.16, respectively.

Table 6.7 summarizes the performance of our approach and RCPR on this test set using all 5

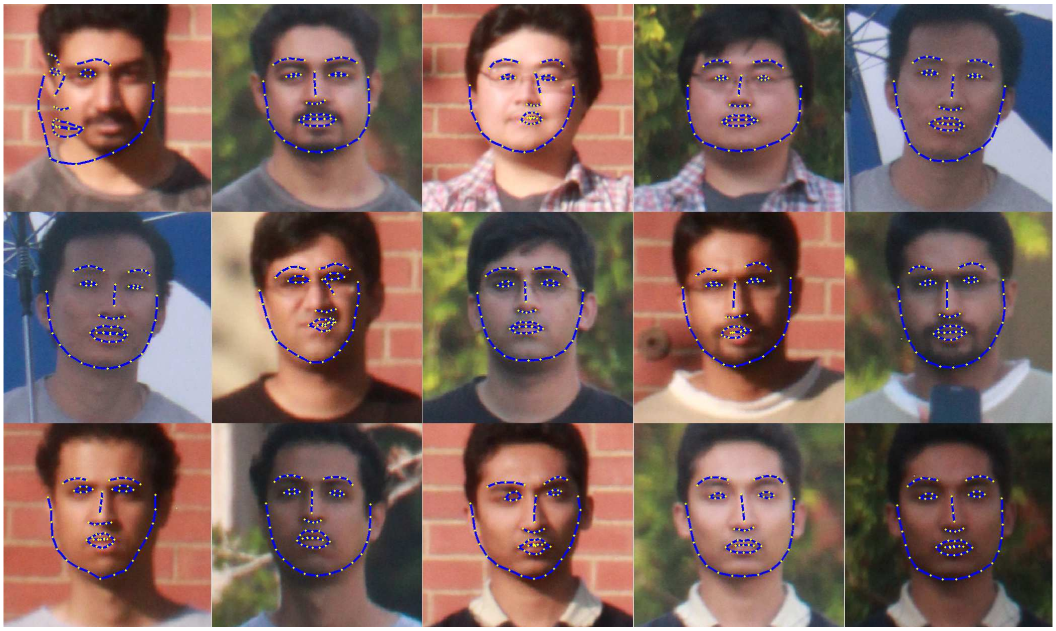


Figure 6.16: Qualitative landmark localization results produced by RCPR (models trained on MPIE images using a downsampling factor of 1) on real-world low-resolution images.

resolution-specific models. While in a practical scenario, it would only be the resolution-specific model that was the closest match to the facial size of the test image that would be used (in this case, the model trained on images with a facial region of size 160×160 on the MPIE images, corresponding to a downsampling factor of 1), we provide these values to again demonstrate the cross-resolution tolerance of the resolution-specific models. As can be seen from the results, the models trained using a downsampling factor of 16 were poorly suited for the task of landmark localization at a higher resolution, however, the fitting error values obtained using the other models are quite close to each other. It must be noted that though the pixel count in the facial region of the test images is misleading in this case due to blurred nature of the images. While the test set used in this experiment was a fairly small one, we carried it out more as a proof of concept and in order to demonstrate the ability of our approach to deal with low-resolution images acquired under real-world conditions.

Table 6.7: Performance of our approach and RCPR using resolution-specific models on the test set of real-world images.

Downsampling Factor in Models	Facial Region Size in Models (Pixels)	Algorithm					
		Ours			RCPR		
		MFE	MNFE (%)	Failure (%)	MFE	MNFE (%)	Failure (%)
1	160×160	9.64	7.71	13	12.17	9.85	13
2	80×80	9.27	7.44	20	15.50	12.62	13
4	40×40	9.42	7.58	13	13.42	11.16	20
8	20×20	8.65	6.97	7	16.33	13.78	20
16	10×10	11.59	9.27	27	14.41	11.56	60

6.5 Concluding Remarks

We explored the challenge posed to the facial alignment process by low-resolution images that also exhibited other variations and degradations, such as pose and expression variations and the presence of occlusions. Our approach was benchmarked against the RCPR algorithm on several test sets by using both single (all-purpose) resolution texture models as well as using resolution-specific texture models. In addition to this, we also stress tested both approaches on a set of images with four variations/degradations (pose, expression, facial occlusions, and low-resolution artifacts) simultaneously present and assessed their performance at various resolutions and occlusion levels. Finally, our approach was also tested on a set of real-world images acquired using a very large capture distance, thus resulting in image artifacts (due to blurring, *etc.*). The key observation that was made after carrying out our various experiments was that relatively reliable facial alignment is possible using our approach for a minimum facial size of approximately 16×16. More accurate initialization or assumptions regarding this could aid in lowering the landmark localization errors at lower resolutions than this. There is a lot of scope for future work in this particular area, and some of the work that could be carried out is discussed in further detail in section 7.1.7, in chapter 7.

Chapter 7

Conclusion

"I've started so I'll finish."

Magnus Magnusson on *Mastermind* and Siddhartha Basu on *Mastermind India*

The problem of designing an all-purpose facial alignment algorithm is not a trivial one. Face detection and alignment occur at the earliest stages in face recognition, expression analysis, or soft biometric system. Poor alignment results are extremely hard to recover from (as we have demonstrated in chapter 4) and thus the need for an algorithm that is able to reliably deal with real-world images and degrade gracefully, with the ability to provide performance feedback to allow for error handling, in the face of extremely challenging cases is crucial.

We have presented a framework for an alignment algorithm that is more accurate than several state-of-the-art approaches, is capable of seamlessly handling a range of yaw variation from $-90°$ to $+90°$, and provides performance feedback in the form of misalignment/occlusion labels despite not being trained on data with occlusion labels. Additionally, we cast the problem of shape regularization as an ℓ_1-regularized least squares problem and demonstrated an improvement in accuracy that was obtained as a result of using this shape modeling and regularization technique over the widely used PCA based shape modeling technique used in ASMs, AAMs, and CLMs. Our approach was thoroughly evaluated on several challenging real-world datasets and compared against several widely used state-of-the-art facial alignment algorithms (in chapter 3). We then provided context for our work on facial alignment by applying it in a real-world face recognition scenario (in chapter 4) and in a large-scale analysis of challenging naturalistic driving videos (in chapter 5). Finally, we extended our approach using resolution-specific models in order to handle the even more complex task of facial landmark localization on low-resolution images (chapter 6). By testing our approach on a large number of different datasets, we have accounted for the individual and joint presence of variations such as pose, illumination, expression, and occlusions, as well as degradation caused due to low-resolution artifacts. Thus, our study is an extremely thorough one that has truly demonstrated the applicability of our approach to real-world scenarios.

7.1 Future Work

While we have made an effort to deal with various aspects of the facial alignment problem, there are several more that could be investigated in the future and we now enumerate some of these possible research directions.

7.1.1 Speedups using Parallel Processing and GPUs

Our approach presently requires a larger amount of time to process an image. However, it is to be noted that this our implementation is currently purely MATLAB [135] based and is not heavily optimized for speed, which is something that we are in the process of addressing. Our approach also contains many steps that lend themselves to parallelization that we haven't taken advantage of yet. GPUs and GPU programming could be used to obtain massive speedups to our implementation and is an area of work that could be explored in the near future.

7.1.2 Better Occlusion Modeling

Presently our local texture classifiers are only trained on texture obtained from around correctly localized landmarks (positive samples) and from displaced landmarks (negative samples). Thus, we do not explicitly train on occluded texture, but group occlusions and misaligned landmarks under the same broad umbrella. This could be modified to treat the problem as a three-class one instead of a two-class one. Even though the texture space spanned by occlusions is extremely large and it is difficult to span this space even with large amounts of training data, it could still prove beneficial to treat this space as a distinct one. Our experiments on the COFW dataset in chapter 3 did show that our approach benefited from training on the training partition of this dataset and that it demonstrated higher landmark localizations on the test partition images after this training. While it is hard to determine if this is only due to the incorporation of the occlusion information at the training stage or due to the diverse shapes and non-occluded texture information that was brought to the table (as was the case when our approach demonstrated superior performance after being trained on images in the LFPW training set partition), it is certainly worthy of investigation. This investigation would also tie in with our previous point about merging face detection and alignment into a single step.

An additional area of improvement could involve better context aware occlusion modeling to make better inferences regarding the occlusion of various landmarks as a group rather than individually, as we presently do. Such information would be even more useful to face recognition algorithms, such as those described in [18], [28], and [29], as this could enable the exclusion of appropriate region from the facial matching process or for its suitable reconstruction (hallucination) prior to the matching stage. Such modeling could involve making assumptions about occlusion zones, as in [11] and [94], though there are some occlusions that may not conform to these assumptions.

7.1.3 Handling Pitch Variation

As we have mentioned, our present facial alignment algorithm does not explicitly account for the shape and textural variations manifested as a result of facial pitch. The problem of handling facial pitch is an extremely challenging one and to the best of our knowledge, no work in the field is able to robustly deal with excessive pitch variation (either individually or simultaneously along with yaw and roll variation). In fact, even the PittPatt face detection algorithm, that is able to robustly detect faces in several real-world images, does not account for pitch and often fails to detect faces exhibiting even slight pitch.

Unlike yaw, which can be modeled using facial symmetry assumptions, the shape and appearance changes to the face when a person looks down are completely different to those that occur when a person looks up. For example, when a person looks down even slightly, the eye region can disappear from view, which does not happen when a person looks up slightly. This asymmetry exacerbates the pitch problem. In addition, there is lack of training data to explicitly account for such variation and in the future more training sets with ground truth landmark locations, generated either synthetically by using 3D modeling techniques to generate these texture and shape views from frontal images, such as in [33], or by manual processing, will be needed to address this problem. Given such data, it would still require a significant amount of thought to suitably modify our approach, or any existing technique for that matter, in order to handle pitch variation. While separate yaw and pitch specific models could be built, this would significantly increase the computational complexity of any alignment algorithm and simultaneously demand more discriminative techniques in order to suitably pick the best fitting one when fitting an unseen image with no prior information available. However, as we have shown in Figure 5.6 in section 5.3.3, our approach does exhibit some ability to generalize to handling variations in pitch with suitable initialization (in this case using the previous frame in a video). Thus, it would also be of interest to determine if robust models could be built by incorporating training images that exhibit more pitch variation along with the the set of images with yaw variation used to build our current shape and texture models, *i.e.*, whether a smaller set of pose models with pitch and yaw variations could perform just as well as separate models trained by creating separate yaw and pitch models for various joint yaw and pitch ranges.

7.1.4 Better Feature Extraction Techniques and Classifiers

While the field of facial alignment has not experienced quite the surge in the application of deep learning methods and Convolutional Neural Networks (CNNs) as some other fields, such as face recognition [205], [206], there has been some recent work on using large training sets in conjunction with these techniques in order to improve landmark localization accuracies [106], [107], [207], [108]. However, while these methods achieve highly accurate landmark localization on challenging images, the restricted nature of the training data that is available is a problem that needs to be addressed in the future. Since training such facial alignment algorithms requires a massive number of images with manually annotated landmarks, it is hard to create or find such publicly available datasets. Many current approaches focus on using the the LFW database and the recently created Annotated Facial Landmarks in the Wild (AFLW) [208], [209] database. However, the images in these datasets have manual annotations for only a sparse set of landmarks (10 for the LFW database and 21 for the AFLW database) and there is thus still a shortage of large datasets with manual annotations for a dense set of landmarks, thus restricting many of these approaches to also localizing only the same sparse set of landmarks. This is likely to change in the future with more and more importance being placed on the task of facial alignment.

It could be possible to extend our framework to harness the power of deep learning techniques. However, a few architectural changes to our approach could be required, such as the avoidance of multiple pose models. Superior discriminative texture models could alleviate the burden placed on the shape regularization stage, which often plays a crucial role in many current approaches, including ours. Advances in the field of deep learning could also conceivably allow for a more robust tackling of the previously mentioned pitch problem.

7.1.5 Joint Face Detection and Alignment

Most present facial alignment approaches, including ours, require a bounding box that conforms to certain specifications as input for initialization purposes. While this is an acceptable requirement or assumption to make for images acquired under constrained conditions in which faces are detected with high reliability, such as those in databases, passport style photographs, *etc.*, it can be a serious failing when dealing with real-world images with occlusions or severely degraded images in a fully automated scenario, *i.e.*, without a human in the loop to correct any errors that occur at any stage of the pipeline. In such cases, such a pipeline could lead to a propagation of error from the face detection stage all the way to the face recognition stage, for example, with the facial alignment algorithm handicapped right from the start.

While progress has been made in the area of face detection in order to detect heavily occluded faces or faces that exhibit large pose variation [17], [210], [211], a pipeline involving facial alignment following this stage could be avoided. Recent efforts, such as those in [6], [16], [94], [212], and [213], have already begun to address this issue and carry out joint face detection and alignment, rather than have the latter follow the former in a sequential fashion. In similar fashion, we believe that it could be possible to extend our approach to function as a joint face detection and alignment algorithm. Our approach searches for an optimal shape initialization in its first two stages and uses simple scoring functions in order to do this. Additionally, we have observed that in cases when a spurious face detection result, *i.e.*, a region of an image that does not contain a face, is provided as input our approach, the resultant alignment has very few inliers with very low percentage of inliers. This fact could be used in a face detection scenario to reject non-faces. However, this would required the use of suitable optimizations and thresholds in order to scale the operation from one that operates on a limited bounding box region to one that operates on an entire image. The training stage of our approach would also need to be modified to incorporate negative classes (non-faces) into the setup, as in [6]. A massively parallelized GPU implementation of our alignment code would also be a crucial step to enable this transition to a simultaneous face detection and alignment algorithm.

7.1.6 Improvements to Landmark Tracking in Videos

We have already alluded to some of the possible improvements that could be made to our facial alignment approach for videos in section 5.6 and reiterate them at this point. The facial landmark alignment algorithm for video fitting that we have presented is one that made no assumptions about the video sequences and could be used for more videos that also involve camera motion, scale changes (due to zooming in or out), or scene changes. More precise models could be developed if certain assumptions about the input videos can be made or if more prior information is available. For example, an on-line training mechanism to develop and update person specific models using accurately processed frames could also be investigated. Since the face is a 3D object, it would also be interesting to explore whether superior landmark tracking results could be obtained using a single set of landmarks to represent a face, as is the case in [16], with visibility and occlusion labels used to denote missing landmarks in a frame. This could allow for smoother transitions between frames than our current technique that switches between using 39 or 68 landmarks based on the facial yaw of the subject in the video sequence. Superior tracking performance could also be obtained by more tightly coupling a face detection and tracking process with the landmark localization step. This could also allow for seamless tracking of multiple subjects in videos.

7.1.7 Improving Facial Alignment on Low-Resolution Images

There is a large scope for improving the performance of facial alignment algorithms when dealing with low-resolution images. While we have demonstrated that the building of resolution specific models is a crucial first step, this work could be advanced in several ways. For example, coupled dictionaries could be built to model relationships between the texture around landmarks in low-resolution and corresponding high-resolution images and used to validate the locations of landmarks in test images. In addition to this, improved texture models could be built using deep learning techniques using large training sets. Our work has also again focused on making very few assumptions regarding the nature of the low-resolution images and the initialization available. With low-resolution images, texture models are less reliable than the shape models and thus if accurate initialization can be guaranteed, shape models could be used to better effect to ensure more accurate localization of landmarks. Finally, facial alignment on low-resolution videos, in conjunction with improved landmark tracking in videos, could be extremely useful in performing superresolution of faces. Previous work has already been carried out on facial superresolution using single images [18], [28], under the assumption that a set of facial landmarks are available for the faces in question. This work could be extended to dealing with video-based superresolution, again under the assumption that accurate landmark localization has been carried out on these frames.

7.1.8 Landmark Localization in Objects

This thesis has focused on facial landmark localization. However, our approach is quite general in nature and can be easily modified to use a varying number of landmarks and also to localize landmarks on any rigid object that can be modeled using landmarks that lie along its contours. Objects, such as cars, that exhibit similar texture for a fixed landmark location and view and similar overall shape for a particular viewpoint can be modeled by our approach by treating each viewpoint as a different pose, as we did with faces (for which the different views corresponded to different yaw ranges).

To demonstrate this we carried out an experiment using images of cars, trucks, and other such similar vehicles in the MIT street scenes dataset [214], [215], which contains over 3500 images of street scenes collected from around Boston, MA using a DSC-F717 camera, for the purposes of object recognition and scene understanding. Landmarks for cars in the dataset are available and Boddeti *et al.* [216] have made annotated and Procrustes aligned data (with image crops of size 356×356 containing the car of interest approximately in the center with a typical size of 250×130)

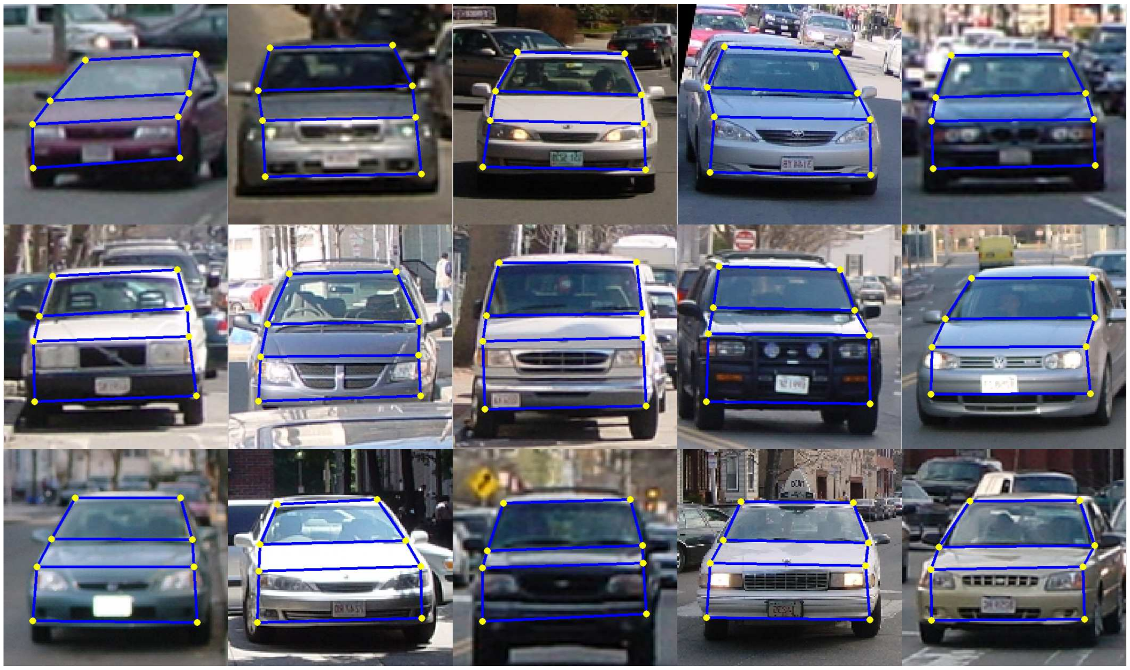


Figure 7.1: Qualitative landmark localization results produced by our approach on car images from our test set that was obtained from the MIT street scenes dataset. Alignment results for the frontal view are shown. In all images with landmarks overlaid on them, yellow dots are used to indicate the locations of facial landmarks and blue line segments connect them. The same color scheme is maintained in all figures in this section that show facial alignment results produced by our approach.

available for 3433 images of cars (varying in types, sizes, background, and presence of occlusions) in the dataset [217]. The images were manually categorized into 5 views with 932 frontal view, $1,400$ half-frontal view, 803 profile view, $1,230$ half-back view and $1,162$ back view images each with 8, 14, 10, 14 and 8 manually annotated landmarks, respectively.

Our alignment approach was trained using $3/4^{\text{th}}$ of the total number of images in each view and tested on the remaining images. Our facial alignment approach was only modified to use different seed landmarks and to use a varying number of landmarks to model the shape in each view. At the testing stage, a crop around the car of interest was generated using the available manual annotations and scaled to match the training crops. Sample alignment results obtained using our approach are shown in Figures 7.1 - 7.5. As can be seen, our approach generalizes quite well to localizing landmarks in such non-facial images with minor changes to our implementation. However, it

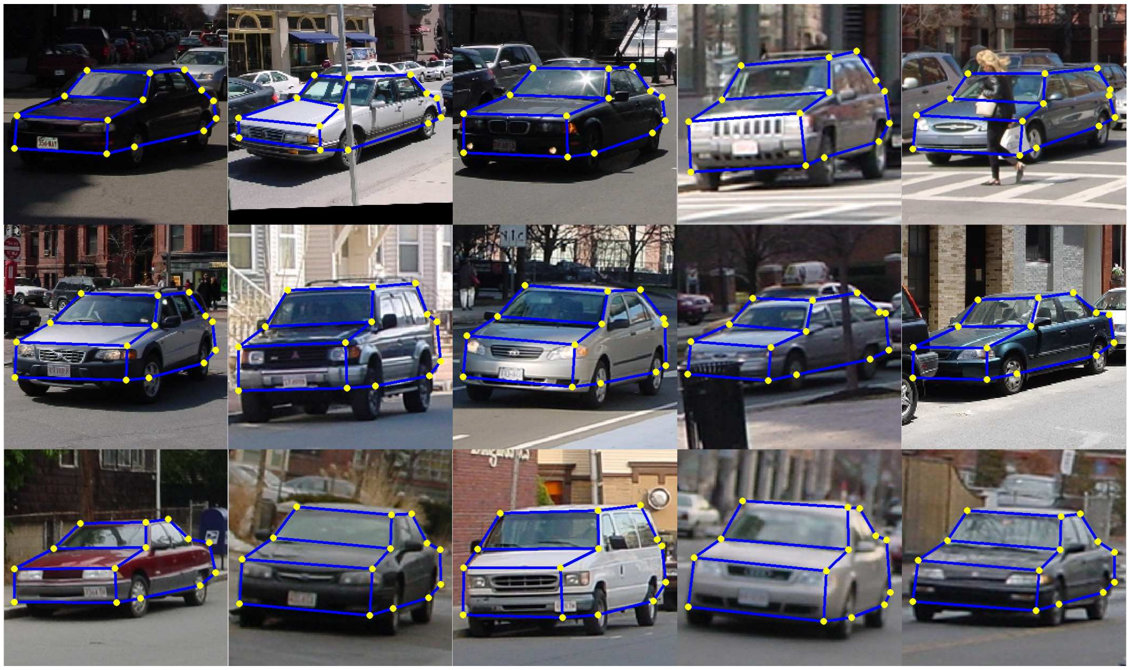


Figure 7.2: Qualitative landmark localization results produced by our approach on car images from our test set that was obtained from the MIT street scenes dataset. Alignment results for the half-frontal view are shown.

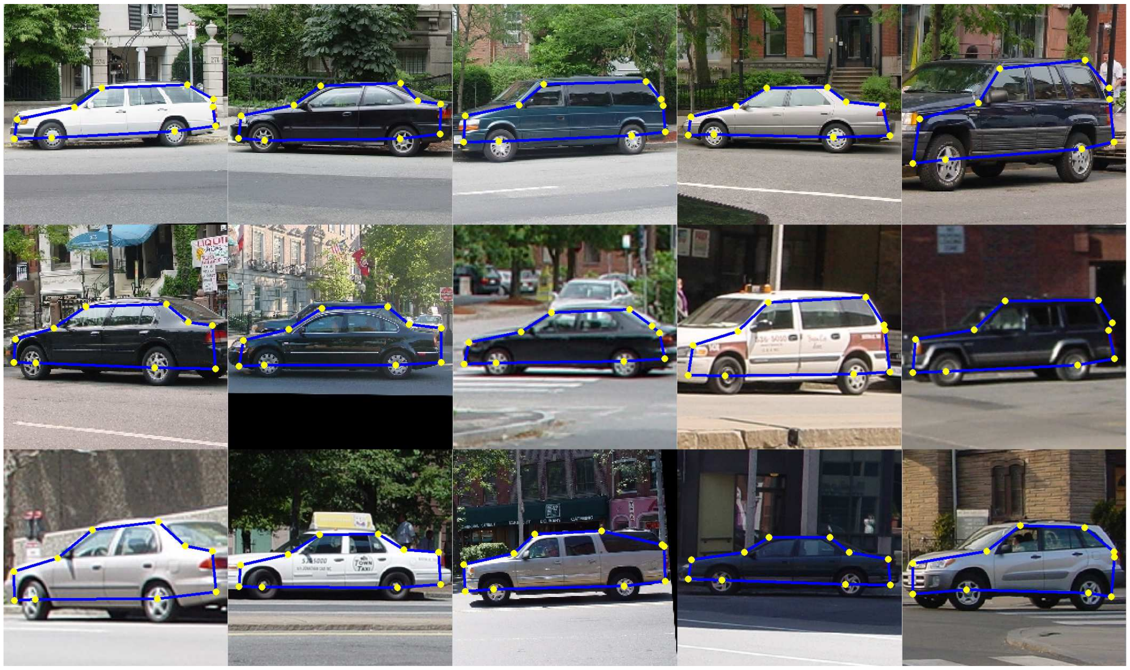


Figure 7.3: Qualitative landmark localization results produced by our approach on car images from our test set that was obtained from the MIT street scenes dataset. Alignment results for the profile view are shown.

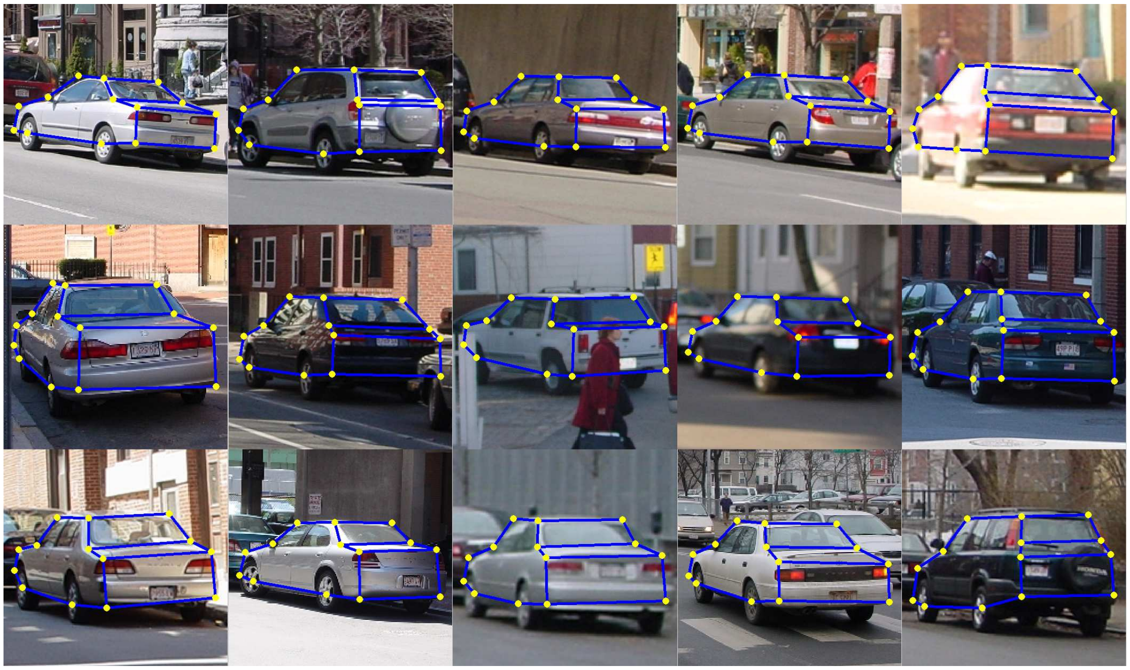


Figure 7.4: Qualitative landmark localization results produced by our approach on car images from our test set that was obtained from the MIT street scenes dataset. Alignment results for the half-back view are shown.

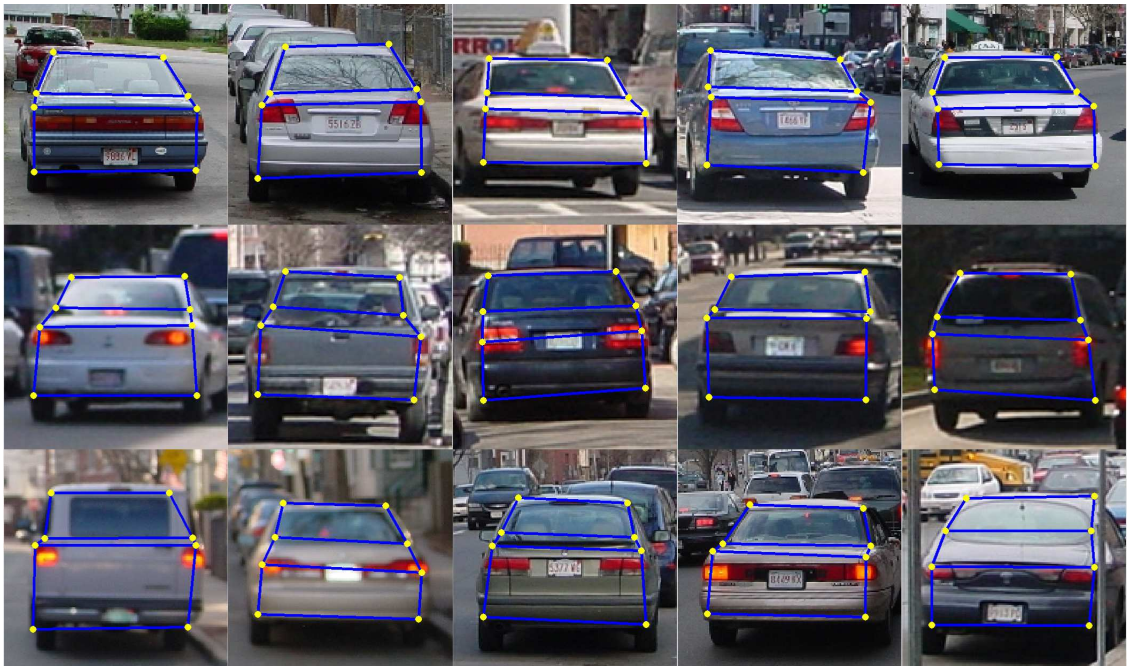


Figure 7.5: Qualitative landmark localization results produced by our approach on car images from our test set that was obtained from the MIT street scenes dataset. Alignment results for the back view are shown.

was observed that it was not trivial to automatically infer the best fitting model after the shape refinement stage due to the fact that all views had a similar number of sparse landmarks. Thus, the results in the figures show alignment results with the best model chosen according to the true view of the car in the image. The sparse number of landmarks in each view also posed a problem to the shape refinement stage that uses inliers (that were very few in number this experiment) landmarks only in order to determine a regularized shape. Finally, it must also be noted that the cars (cars is lose term for sedans, trucks, vans, *etc.*) in each view did exhibit more variation than typical human faces that posed a challenge at the test stage. Thus, we describe this experiment to demonstrate the capability of our approach and as a proof of concept. In such cases, with only a sparse set of landmarks to localize, better local texture models (appearance modeling) can lead to higher landmark localization accuracies and this stage assumes a more significant role than the shape regularization stage, as discussed in [216]. However, if more landmarks are available to model an object of interest, then our approach could deliver more accurate alignment results with better dense shape alignment to start with and our shape regularization technique to follow. Additionally, clustering the data to train models specifically for sedans, trucks, vans, *etc.* could also improve the alignment accuracy. Future work could include modifying our approach to align such objects of interest with investigations into modifications needed in cases where only a sparse set of landmarks are available to model the object. Research into a suitable joint object detection and landmark localization framework, in a similar fashion to joint face detection and landmark localization, would also be interesting to pursue.

Appendices

Appendix A

The Real AdaBoost Algorithm

Algorithms and equations in this appendix are reproduced for convenience from [114] with minor changes in notation.

Let $S = ((\mathbf{x}_1, y_1), \ldots, (\mathbf{x}_N, y_N))$ denote a set of N training examples where $\mathbf{x}_i \in \mathbb{R}^M$ is a set of feature vectors and $y_i \in \{-1, +1\}$ is a set of labels for the features vectors in a binary classification problem. Given these training samples, along with a set of weights w_i for each data sample over the indices of S, *i.e.*, over $\{1, \ldots, N\}$, the Real AdaBoost algorithm is an ensemble learning method that aims at combining a set of weak learners or classifiers $f_t(\mathbf{x})$ to form a stronger prediction rule. In the most general case, $f_t(\mathbf{x})$ has the form $f_t(\mathbf{x}) : \mathbb{R}^M \rightarrow \mathbb{R}$. Boosting uses a weak learner repeatedly over a set of rounds $t = 1, \ldots, T$. The weights of the training examples are updated after each round (iteration) based on which samples were correctly or incorrectly classified in that round. It is to be noted that the sign of $f_t(\mathbf{x})$ can be interpreted as the predicted label (-1 or $+1$) to be assigned to instance $\mathbf{x}$, and the magnitude of $f_t(\mathbf{x})$ ($|f(\mathbf{x})|$) as the confidence of the prediction. When decision trees are used as the weak learners, this form of Real AdaBoost coincides with one of the forms of the generalized AdaBoost algorithm, outlined by Schapire and Singer in [116]. The Real AdaBoost algorithm is summarized in Algorithm 3.

Algorithm 3 Overview of the Real AdaBoost algorithm.

```
Input: Training samples and labels $S = ((\mathbf{x}_1, y_1), \ldots, (\mathbf{x}_N, y_N))$ where $\mathbf{x}_i \in \mathbb{R}^M$ and $y_i \in \{-1, +1\}$ and initial weights for the samples $w_i = 1/N \;\; i = 1, \ldots, N$
Output: Ensemble classifier $h(\mathbf{x}) = \mathrm{sign}[\sum_{t=1}^{T} f_t(\mathbf{x})]$
for $t = 1, \ldots, T$ do
    Fit the classifier to obtain a class probability estimate $p_t(\mathbf{x}) = P_w(y = 1|\mathbf{x}) \in [0, 1]$ using weights $w_i$ on the training data
    Set $f_t(\mathbf{x}) \leftarrow \frac{1}{2}\log\frac{p_t(\mathbf{x})}{1-p_t(\mathbf{x})} \in \mathbb{R}$
    Set $w_i \leftarrow w_i\exp[-y_i f_t(\mathbf{x}_i)] \;\; i = 1, \ldots, N$ and re-normalize so that $\sum_{i=1}^{N}(w_i) = 1$
end for
Output the ensemble classifier $h(\mathbf{x}) = \mathrm{sign}[\sum_{t=1}^{T} f_t(\mathbf{x})]$
```

Appendix B

Solving the ℓ_1-Regularized Least Squares Problem (LSP)

Algorithms and equations in this appendix are reproduced for convenience from [134] with minor changes in notation. The reader is referred to [134] for more details on the problem and the proposed solution.

Kim *et al.* [134] proposed an approach to solve the ℓ_1-regularized Least Squares Problem (LSP), whose objective function is given in equation (B.1). In equation (B.1), $\mathbf{A} \in \mathbb{R}^{m \times n}$ is a data matrix, $\mathbf{y} \in \mathbb{R}^m$ is a vector of observations, $\mathbf{x} \in \mathbb{R}^n$ is a vector of unknowns, and $\lambda > 0$ is a regularization parameter.

$$\underset{\mathbf{x}}{\text{minimize}} \;\; \|\mathbf{A}\mathbf{x} - \mathbf{y}\|_2^2 + \lambda \|\mathbf{x}\|_1 \tag{B.1}$$

The objective function of the ℓ_1-regularized LSP is convex, but not differentiable and the following first-order optimality conditions, that are the necessary and sufficient conditions for $\mathbf{x}$ to be optimal, are obtained using subdifferential calculus:

$$(2\mathbf{A}^{\mathrm{T}}(\mathbf{Ax}-\mathbf{y}))_i \in \begin{cases} \{+\boldsymbol{\lambda}_i\} & \text{if } \mathbf{x}_i > 0 \\ \{-\boldsymbol{\lambda}_i\} & \text{if } \mathbf{x}_i < 0, \quad i = 1,\ldots,n \\ [-\boldsymbol{\lambda}_i, +\boldsymbol{\lambda}_i] & \text{if } \mathbf{x}_i = 0 \end{cases} \tag{B.2}$$

The condition that an all zero vector becomes an optimal solution is that $(2\mathbf{A}^{\mathrm{T}}\mathbf{y})_i \in [-\lambda, +\lambda] \;\; i = 1,\ldots,n$, *i.e.*, $\lambda \geq \lambda_{\max} = \|2\mathbf{A}^{\mathrm{T}}\mathbf{y}\|_\infty$.

A Lagrange dual of the problem in equation (B.1) can be obtained by introducing a variable $\mathbf{z} \in \mathbb{R}^m$ and an equality constraint $\mathbf{z} = \mathbf{Ax} - \mathbf{y}$, to formulate the equivalent problem:

$$\text{minimize } \; \mathbf{z}^{\mathrm{T}}\mathbf{z} + \lambda\|\mathbf{x}\|_1 \tag{B.3}$$

$$\text{subject to } \; \mathbf{z} = \mathbf{Ax} - \mathbf{y} \tag{B.4}$$

Associating dual variables $\boldsymbol{\nu}_i \in \mathbb{R}, \;\; i = 1,\ldots,m$ with the equality constraints $\mathbf{z}_i = (\mathbf{Ax}-\mathbf{y})_i$ results in the Lagrangian $L(\mathbf{x},\mathbf{z},\boldsymbol{\nu})$, given by equation (B.5).

$$L(\mathbf{x},\mathbf{z},\boldsymbol{\nu}) = \mathbf{z}^{\mathrm{T}}\mathbf{z} + \lambda\|\mathbf{x}\|_1 + \boldsymbol{\nu}^{\mathrm{T}}(\mathbf{Ax}-\mathbf{y}-\mathbf{z}) \tag{B.5}$$

The dual function is given by equation (B.6).

$$\inf_{\mathbf{x},\mathbf{z}} \; L(\mathbf{x},\mathbf{z},\boldsymbol{\nu}) = \begin{cases} -(1/4)\boldsymbol{\nu}^{\mathrm{T}}\boldsymbol{\nu} - \boldsymbol{\nu}^{\mathrm{T}}\mathbf{y}, \;\; |(\mathbf{A}^{\mathrm{T}}\boldsymbol{\nu})_i| \leq \boldsymbol{\lambda}_i, \;\; i = 1,\ldots,m \\ -\infty, \;\; \text{otherwise} \end{cases} \tag{B.6}$$

The Lagrangian dual of equation (B.4) is:

$$\text{maximize } \; G(\boldsymbol{\nu}) \tag{B.7}$$

$$\text{subject to } \; |(\mathbf{A}^{\mathrm{T}}\boldsymbol{\nu})_i| \leq \boldsymbol{\lambda}_i, \;\; i = 1, \ldots, m \tag{B.8}$$

where the dual objective $G(\boldsymbol{\nu}) = -(1/4)\boldsymbol{\nu}^{\mathrm{T}}\boldsymbol{\nu} - \boldsymbol{\nu}^{\mathrm{T}}\mathbf{y}$.

The dual problem in equation (B.8) is a convex optimization one with variable $\boldsymbol{\nu} \in \mathbb{R}^m$ that is dual feasible if it satisfies the constraints of equation (B.8). Any dual feasible point gives a lower bound on the optimal value p^* of the primal problem in equation (B.1), *i.e.*, $G(\boldsymbol{\nu}) \leq p^*$. This is called weak duality. However, it is to be noted that in this case, the optimal values of the primal and dual are equal due to strong duality.

An important property of the ℓ_1-regularized LSP is that it is easy to derive a bound on the suboptimality of an arbitrary $\mathbf{x}$ by constructing a dual feasible point, given by equation (B.10).

$$\boldsymbol{\nu} = 2s(\mathbf{A}\mathbf{x} - \mathbf{y}) \tag{B.9}$$

$$s = \min\{\lambda / |2(\mathbf{A}^{\mathrm{T}}\mathbf{A}\mathbf{x})_i - 2\mathbf{y}_i| \;\; i = 1, \ldots, m\} \tag{B.10}$$

The point $\boldsymbol{\nu}$ is dual feasible, so $G(\boldsymbol{\nu})$ is a lower bound on p^*, the optimal value of the primal version of the ℓ_1-regularized LSP in equation (B.1).

The difference between the primal objective value of $\mathbf{x}$ and the lower bound $G(\boldsymbol{\nu})$ is the duality gap η, that is given by equation (B.11).

$$\eta = \|\mathbf{A}\mathbf{x} - \mathbf{y}\|_2^2 + \lambda\|\mathbf{x}\|_1 - G(\boldsymbol{\nu}) \tag{B.11}$$

The duality gap is always nonnegative by weak duality. Strong duality holds at an optimal point, and the duality gap is zero.

The objective function in equation (B.1) is convex but not differentiable and can be transformed into a convex Quadratic Problem (QP) with linear inequality constraints, as shown in equation (B.12).

$$\begin{aligned} &\underset{\mathbf{x}}{\text{minimize}} \;\; \|\mathbf{A}\mathbf{x}-\mathbf{y}\|_2^2 + \lambda \sum_{i=1}^{n} \mathbf{u}_i \\ &\text{subject to} - \mathbf{u}_i \leq \mathbf{x}_i \leq \mathbf{u}_i, \quad i = 1, \ldots, n \end{aligned} \tag{B.12}$$

In equation (B.12), $\mathbf{x} \in \mathbb{R}^n$ and $\mathbf{u} \in \mathbb{R}^n$. This QP is solved using an interior-point method. The logarithmic barrier for the bound constraints $-\mathbf{u}_i \leq \mathbf{x}_i \leq \mathbf{u}_i$ in equation (B.12) is given by:

$$\Phi(\mathbf{x}, \mathbf{u}) = -\sum_{i=1}^{n} \log(\mathbf{u}_i + \mathbf{x}_i) - \sum_{i=1}^{n} \log(\mathbf{u}_i - \mathbf{x}_i) \tag{B.13}$$

defined over $\mathbf{dom}\Phi = \{(\mathbf{x}, \mathbf{u}) \in \mathrm{R}^n \times \mathrm{R}^n, \;\; |\mathbf{x}_i| < \mathbf{u}_i, \;\; i = 1, \ldots, n\}$. The central path consists of the unique minimizer $(\mathbf{x}^*(t), \mathbf{u}^*(t))$ of the convex function:

$$\phi_t(\mathbf{x}, \mathbf{u}) = t\|\mathbf{A}\mathbf{x}-\mathbf{y}\|_2^2 + t\sum_{i=1}^{n} \lambda \mathbf{u}_i + \Phi(\mathbf{x}, \mathbf{u}) \tag{B.14}$$

as the parameter t varies from 0 to ∞. In the primal barrier method, Newton's method is used to minimize ϕ_t, *i.e.*, the search direction is computed as the exact solution to the Newton system:

$$\mathbf{H} \begin{bmatrix} \Delta\mathbf{x} \\ \Delta\mathbf{u} \end{bmatrix} = -\mathbf{g} \tag{B.15}$$

where $\mathbf{H} = \nabla^2 \phi_t(\mathbf{x}, \mathbf{u}) \in \mathbb{R}^{2n \times 2n}$ is the Hessian and $\mathbf{g} = \nabla \phi_t(\mathbf{x}, \mathbf{u}) \in \mathbb{R}^{2n}$ is the gradient at the current iterate $(\mathbf{x}, \mathbf{u})$. For a large scale ℓ_1-regularized LSP, solving the Newton system is not computationally feasible. In the method proposed in [134], the search direction is computed using as an approximate solution to the Newton system in equation (B.15) using Preconditioned Conjugate Gradients (PCG). The overall method is referred to as a truncated Newton method.

Algorithm 4 The truncated Newton interior-point method for solving ℓ_1-regularized LSPs.

```
**Input**: relative tolerance $\epsilon_{\text{rel}} > 0$ and regularization parameter $\lambda > 0$
**Initialize**: $t = 1/\lambda$, $\mathbf{x} = [0\,0\dots0]^{\mathrm{T}} \in \mathbb{R}^n$, $\mathbf{u} = [1\,1\dots1]^{\mathrm{T}} \in \mathbb{R}^n$
**Repeat**:
Compute the search direction $(\Delta\mathbf{x}, \Delta\mathbf{u})$ as an approximate solution to the Newton system in equation (B.15)
Compute the step size $s$ by backtracking line search
Update the iterate by $(\mathbf{x}, \mathbf{u}) = (\mathbf{x}, \mathbf{u}) + s(\Delta\mathbf{x}, \Delta\mathbf{u})$
Construct a dual feasible point point $\boldsymbol{\nu}$ using equation (B.10)
Evaluate the duality gap $\eta$ using equation (B.11)
**Quit** if $\eta/G(\boldsymbol{\nu}) \le \epsilon_{\text{rel}}$
Update $t$
```

Truncated Newton methods have been applied to interior-point methods in prior work [218], [219], [220], [221]. In the primal barrier method, the parameter t is held constant until ϕ_t is approximately minimized For faster convergence, t is updated in each iteration based on the duality gap that is computed using the dual feasible point constructed using equation (B.10). The final algorithm proposed in [134] is summarized in Algorithm 4.

The stopping criterion used in Algorithm 4 is the duality gap divided by the dual objective value. By weak duality, the ratio is an upper bound on the relative suboptimality:

$$\frac{f(\mathbf{x}) - p^*}{p^*} \le \frac{\eta}{G(\boldsymbol{\nu})} \tag{B.16}$$

where p^* has already been previously defined as the optimal value of the ℓ_1-regularized LSP and $f(\mathbf{x})$ is the primal objective computed with the point $\mathbf{x}$. The method solves the problem to guaranteed relative accuracy (or tolerance) $\epsilon_{\text{rel}} > 0$. The update rule used for t is given by equation (B.17).

$$t = \begin{cases} \max\{\mu\min\{2n/\eta, t\}, t\}, & s \ge s_{\text{min}} \\ t, & s < s_{\text{min}} \end{cases} \tag{B.17}$$

In equation (B.17), $\mu > 1$ and $s_{\text{min}} \in (0, 1]$ are parameters to be chosen. $\mu = 2$ and $s_{\text{min}} = 0.5$

were the values chosen in [134] as these values were found to provide acceptable performance on most problems. The proposed update rule was found to be quite robust and worked well when combined with the PCG algorithm. The reader is referred to [134] for further details on the PCG algorithm and the overall solution to the ℓ_1-regularized LSP.

Bibliography

[1] R. Gross, I. Matthews, J. Cohn, T. Kanade, and S. Baker, “Multi-PIE,” in *Proceedings of the IEEE International Conference on Face and Gesture Recognition (FG)*, Sep. 2008, pp. 1–8. (document), 1.2, 2.1, 3.1.1, 3.2.1

[2] ——, “Multi-PIE,” *Image and Vision Computing*, vol. 28, no. 5, pp. 807–813, May 2010. (document), 1.2, 2.1, 3.1.1, 3.2.1

[3] ——, “The CMU Multi-PIE Face Database,” http://www.multipie.org/. (document), 1.2, 2.1, 3.1.1, 3.2.1

[4] P. N. Belhumeur, D. W. Jacobs, D. J. Kriegman, and N. Kumar, “Localizing Parts of Faces Using a Consensus of Exemplars,” in *Proceedings of the IEEE International Conference on Computer Vision and Pattern Recognition (CVPR)*, Jun. 2011, pp. 545–552. (document), 1.2, 2.1, 2.2, 3.1, 3.1.3, 3.2.2, 5.5.2

[5] ——, “Labeled Face Parts in the Wild (LFPW) Dataset,” http://homes.cs.washington.edu/~neeraj/databases/lfpw/, 2011. (document), 1.2, 2.1, 3.1, 3.2.2

[6] X. Zhu and D. Ramanan, “Face Detection, Pose Estimation, and Landmark Localization in the Wild,” in *Proceedings of the IEEE International Conference on Computer Vision and Pattern Recognition (CVPR)*, Jun. 2012, pp. 2879–2886. (document), 1.1, 1.2, 2.2, 2.3, 2.5, 3.1.1, 3.2.2, 3.2.3, 4, 4.2, 5.5.1, 7.1.5

[7] ——, “Face Detection, Pose Estimation, and Landmark Localization in the Wild,” http:

//www.ics.uci.edu/~xzhu/face/. (document), 1.2, 3.2.2, 3.2.3

[8] C. Sagonas, G. Tzimiropoulos, S. Zafeiriou, and M. Pantic, “300 faces in-the-Wild Challenge: The first facial landmark localization Challenge,” in *Proceedings of the IEEE International Conference on Computer Vision Workshops (ICCVW)*, Dec. 2013, pp. 397–403. (document), 1.2, 2.3, 3.2.2, 3.11

[9] “300 Faces in-the-Wild Challenge (300-W), ICCV 2013,” http://ibug.doc.ic.ac.uk/resources/300-W/. (document), 1.2, 2.3, 3.2.2, 3.2.3, 3.11

[10] “Intelligent Behaviour Understanding Group (ibug) Dataset,” http://ibug.doc.ic.ac.uk/download/annotations/ibug.zip/. (document), 1.2, 3.2.2

[11] X. P. Burgos-Artizzu, P. Perona, and P. Dollar, “Robust face landmark estimation under occlusion,” in *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, Dec. 2013, pp. 1513–1520. (document), 1.2, 1.2, 2.3, 2.5, 3.2.2, 3.2.3, 4.2, 6.1, 7.1.2

[12] ——, “Robust face landmark estimation under occlusion,” http://www.vision.caltech.edu/xpburgos/ICCV13/. (document), 1.2, 2.3, 3.2.2, 3.2.3, 6.1

[13] K. Seshadri and M. Savvides, “Towards a Unified Framework for Pose, Expression, and Occlusion Tolerant Automatic Facial Alignment,” *To appear in IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, 2016. (document), 1.2, 1.2, 3, 3.1, 3.2, 3.4, 3.6, 3.7, 3.14, 3.17, 3.18, 3.6, 3.7, 3.3, 4

[14] P. J. Phillips, P. J. Flynn, J. R. Beveridge, W. T. Scrugs, A. J. O’Toole, D. Bolme, K. W. Bowyer, B. A. Draper, G. H. Givens, Y. M. Lui, H. Sahibzada, J. A. Scallan III, and S. Weimer, “Overview of the Multiple Biometrics Grand Challenge,” in *Proceedings of the IAPR/IEEE International Conference on Biometrics (ICB)*, Jun. 2009, pp. 705–714. (document), 2.1, 2.2

[15] National Institute of Standards and Technology (NIST), “Multiple Biometric Grand Challenge,” http://www.nist.gov/itl/iad/ig/mbgc.cfm, 2008. (document), 2.1, 2.2

[16] X. Yu, J. Huang, S. Zhang, W. Yan, and D. N. Metaxas, “Pose-free Facial Landmark Fitting via Optimized Part Mixtures and Cascaded Deformable Shape Model,” in *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, Dec. 2013, pp. 1944–1951. (document), 1.1, 2.2, 2.4, 2.5, 3.2.3, 4, 4.2, 5.3.1, 7.1.5, 7.1.6

[17] M. Mathias, R. Benenson, M. Pedersoli1, and L. V. Gool, “Face detection without Bells and Whistles,” in *Proceedings of the European Conference on Computer Vision (ECCV), Part IV - Volume 8692 of the series Lecture Notes in Computer Science (LNCS)*, Sep. 2014, pp. 720–735. (document), 3.2.6, 3.11, 3.20, 7.1.5

[18] U. Prabhu, “Face Recognition and Recovery under Simultaneous Real-World Degradations,” Ph.D. dissertation, Carnegie Mellon University, 2015. (document), 1, 1.1, 1.2, 4.1, 4.1, 4.1.1, 4.2, 4.1, 4.1.1, 4.1.2, 4.1.3, 4.6, 4.3, 4.2, 4.2, 6.1.1, 7.1.2, 7.1.7

[19] “InSight Data Access Website Strategic Highway Research Program (SHRP2) Naturalistic Driving Study,” https://insight.shrp2nds.us/. (document), 5.2, 5.4, 5.2, 5.5, 5.6, 5.7, 5.8, 5.9

[20] K. Seshadri, F. Juefei-Xu, D. K. Pal, and M. Savvides, “Driver Cell Phone Usage Detection on Strategic Highway Research Program (SHRP2) Face View Videos,” in *Proceedings of the International Workshop on Computer Vision in Vehicle Technology (CVVT) in conjunction with the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Jun. 2015, pp. 35–43. (document), 1.2, 5.12, 5.6

[21] G. J. Edwards, T. F. Cootes, and C. J. Taylor, “Face Recognition Using Active Appearance Models,” in *Proceedings of the European Conference on Computer Vision (ECCV), Volume II - Volume 1407 of the series Lecture Notes in Computer Science (LNCS)*, Jun. 1998, pp. 581–595. 1

[22] N. Faggian, A. Paplinski, and T. J. Chin, “Face Recognition From Video Using Active Appearance Model Segmentation,” in *Proceedings of the International Conference on Pattern Recognition (ICPR) - Volume 01*, Aug. 2006, pp. 287–290. 1

[23] J. Heo, “3D Generic Elastic Models for 2D Pose Synthesis and Face Recognition,” Ph.D. dissertation, Carnegie Mellon University, 2010. 1

[24] L. Teijeiro-Mosquera, J. L. Alba-Castro, and D. González-Jiménez, “Face recognition across pose with automatic estimation of pose parameters through AAM-based landmarking,” in *Proceedings of the International Conference on Pattern Recognition (ICPR)*, Aug. 2010, pp. 1339–1342. 1

[25] R. Abiantun, U. Prabhu, K. Seshadri, J. Heo, and M. Savvides, “An Analysis of Facial Shape and Texture for Recognition: A large scale Evaluation on FRGC ver2.0,” in *Proceedings of the IEEE Workshop on Applications of Computer Vision (WACV)*, Jan. 2011, pp. 212–219. 1, 4.1.1

[26] U. Prabhu, J. Heo, and M. Savvides, “Unconstrained Pose-Invariant Face Recognition Using 3D Generic Elastic Models,” *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, vol. 33, no. 10, pp. 1952–1961, Oct. 2011. 1

[27] T. H. N. Le, K. Luu, K. Seshadri, and M. Savvides, “A Facial Aging Approach to Identification of Identical Twins,” in *Proceedings of the IEEE International Conference on Biometrics: Theory, Applications and Systems (BTAS)*, Sep. 2012, pp. 1–8. 1, 4.1.1

[28] R. A. Antoun, “Pose-Tolerant Face Recognition,” Ph.D. dissertation, Carnegie Mellon University, 2013. 1, 1.1, 1.2, 6.1.1, 7.1.2, 7.1.7

[29] R. Abiantun, U. Prabhu, and M. Savvides, “Sparse Feature Extraction for Pose-Tolerant Face Recognition,” *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, vol. 36, no. 10, pp. 2061–2073, Oct. 2014. 1, 1.1, 1.2, 6.1.1, 7.1.2

[30] T. H. N. Le, K. Seshadri, K. Luu, and M. Savvides, “Facial Aging and Asymmetry Decomposition Based Approaches to Identification of Twins,” *Pattern Recognition*, vol. 48, no. 12, pp. 3843–3856, Dec. 2015. 1, 4.1.1

[31] J. Zhu, S. C. H. Hoi, E. Yau, and M. R. Lyu, “Automatic 3D Face Modeling Using 2D Active

Appearance Models," in *Proceedings of the Pacific Conference on Computer Graphics and Applications*, Oct. 2005, pp. 12–14. 1

[32] S. W. Park, J. Heo, and M. Savvides, "3D Face Reconstruction from a Single 2D Face Image," in *IEEE International Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, Jun. 2008, pp. 1–8. 1

[33] S. Y. Baek, B. Y. Kim, and K. Lee, "3D Face Model Reconstruction From Single 2D Frontal Image," in *Proceedings of the International Conference on Virtual Reality Continuum and its Applications in Industry (VRCAI)*, Dec. 2009, pp. 95–101. 1, 7.1.3

[34] J. Heo and M. Savvides, "3-D Generic Elastic Models for Fast and Texture Preserving 2-D Novel Pose Synthesis," *IEEE Transactions on Information Forensics and Security (TIFS)*, vol. 7, no. 2, pp. 563–576, Apr. 2012. 1

[35] ——, "Gender and Ethnicity Specific Generic Elastic Models from a Single 2D Image for Novel 2D Pose Face Synthesis and Recognition," *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, vol. 34, no. 12, pp. 2341–2350, Dec. 2012. 1

[36] O. Rudovic, I. Patras, and M. Pantic, "Regression-based Multi-View Facial Expression Recognition," in *Proceedings of the International Conference on Pattern Recognition (ICPR)*, Aug. 2010, pp. 4121–4124. 1

[37] M. A. Nicolaou, H. Gunes, and M. Pantic, "Output-associative RVM regression for dimensional and continuous emotion prediction," *Image and Vision Computing*, vol. 30, no. 3, pp. 186–196, Mar. 2012. 1

[38] H. C. Choi and S. Y. Oh, "Real-time Recognition of Facial Expression Using Active Appearance Model with Second Order Minimization and Neural Network," in *Proceedings of the 2006 IEEE International Conference on Systems, Man, and Cybernetics*, Oct. 2006, pp. 1559–1564. 1

[39] A. B. Ashraf, S. Lucey, J. F. Cohn, T. Chen, Z. Ambadar, K. M. Prkachin, and P. E. Solomon,

"The Painful Face - Pain Expression Recognition Using Active Appearance Models," *Image and Vision Computing*, vol. 27, no. 12, pp. 1788–1796, Nov. 2009. 1

[40] J. Merkow, B. Jou, and M. Savvides, "An Exploration of Gender Identification Using Only the Periocular Region," in *Proceedings of the IEEE International Conference on Biometrics: Theory, Applications and Systems (BTAS)*, Sep. 2010, pp. 1–5. 1

[41] S. Y. D. Hu, B. Jou, A. Jaech, and M. Savvides, "Fusion of Region-Based Representations for Gender Identification," in *Proceedings of the IEEE International Joint Conference on Biometrics (IJCB)*, Oct. 2011, pp. 1–7. 1

[42] K. Luu, K. Ricanek, T. D. Bui, and C. Y. Suen, "Age Estimation using Active Appearance Models and Support Vector Machine Regression," in *Proceedings of the IEEE International Conference on Biometrics: Theory, Applications and Systems (BTAS)*, Sep. 2009, pp. 1–8. 1

[43] K. Luu, K. Seshadri, M. Savvides, T. D. Bui, and C. Y. Suen, "Contourlet Appearance Model for Facial Age Estimation," in *Proceedings of the IEEE International Joint Conference on Biometrics (IJCB)*, Oct. 2011, pp. 1–8. 1, 4.1.1

[44] K. Luu, T. H. N. Le, K. Seshadri, and M. Savvides, "FaceCut - A Robust Approach for Facial Feature Segmentation," in *Proceedings of the IEEE International Conference on Image Processing (ICIP)*, Sep. 2011, pp. 1841–1844. 1, 4.1.1

[45] T. H. N. Le, K. Luu, K. Seshadri, and M. Savvides, "Beard and Mustache Segmentation using Sparse Classifiers on Self-Quotient Images," in *Proceedings of the IEEE International Conference on Image Processing (ICIP)*, Sep. 2011, pp. 165–168. 1, 4.1.1

[46] T. H. N. Le, K. Luu, and M. Savvides, "SparCLeS: Dynamic ℓ_1 Sparse Classifiers With Level Sets for Robust Beard/Moustache Detection and Segmentation," *IEEE Transactions on Image Processing (TIP)*, vol. 22, no. 8, pp. 3097–3107, Aug. 2013. 1

[47] A. H. Gee and R. Cipolla, "Determining the Gaze of Faces in Images," *Image and Vision*

Computing, vol. 12, no. 10, pp. 639–647, Dec. 1994. 1, 5.4.1, 5.4, 5.5, 5.6, 5.4.1

[48] T. Horprasert, Y. Yacoob, and L. S. Davis, “Computing 3-D Head Orientation from a Monocular Image Sequence,” in *Proceedings of the IEEE International Conference on Face and Gesture Recognition (FG)*, Oct. 1996, pp. 242–247. 1, 5.4.1

[49] S. Zhao and Y. Gao, “Automated Face Pose Estimation Using Elastic Energy Models,” in *Proceedings of the International Conference on Pattern Recognition (ICPR) - Volume 04*, Aug. 2006, pp. 618–621. 1

[50] L. Morency, J. Whitehill, and J. Movellan, “Monocular Head Pose Estimation Using Generalized Adaptive View-based Appearance Model,” *Image and Vision Computing*, vol. 28, no. 5, pp. 754–761, May 2010. 1

[51] F. Juefei-Xu, M. Cha, J. Heyman, S. Venogopalan, R. Abiantun, and M. Savvides, “Robust Local Binary Pattern Feature Sets for Periocular Biometric Identification,” in *Proceedings of the IEEE International Conference on Biometrics: Theory, Applications and Systems (BTAS)*, Sep. 2010, pp. 1–8. 1.1

[52] F. Juefei-Xu, K. Luu, M. Savvides, T. Bui, and C. Y. Suen, “Investigating Age Invariant Face Recognition Based on Periocular Biometrics,” in *Proceedings of the IEEE International Joint Conference on Biometrics (IJCB)*, Oct. 2011, pp. 1–7. 1.1

[53] F. Juefei-Xu and M. Savvides, “Subspace-Based Discrete Transform Encoded Local Binary Patterns Representations for Robust Periocular Matching on NIST’s Face Recognition Grand Challenge,” *IEEE Transactions on Image Processing (TIP)*, vol. 23, no. 8, pp. 3409–3505, Aug. 2014. 1.1

[54] F. Juefei-Xu, K. Luu, and M. Savvides, “Spartans: Single-Sample Periocular-Based Alignment-Robust Recognition Technique Applied to Non-Frontal Scenarios,” *IEEE Transactions on Image Processing (TIP)*, vol. 24, no. 12, pp. 4780–4785, Dec. 2015. 1.1, 4.3, 4.2

[55] K. Seshadri and M. Savvides, "Robust Modified Active Shape Model for Automatic Facial Landmark Annotation of Frontal Faces," in *Proceedings of the IEEE International Conference on Biometrics: Theory, Applications and Systems (BTAS)*, Sep. 2009, pp. 319–326. 1.2, 2.1, 3.3, 4.1.1, 5.3.1

[56] ——, "An Analysis of the Sensitivity of Active Shape Models to Initialization When Applied to Automatic Facial Landmarking," *IEEE Transactions on Information Forensics and Security (TIFS)*, vol. 7, no. 4, pp. 1255–1269, Aug. 2012. 1.2, 2.1, 3.1.2, 3.3, 4.1.1

[57] T. F. Cootes, G. J. Edwards, and C. J. Taylor, "Active Appearance Models," *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, vol. 23, no. 6, pp. 681–685, Jun. 2001. 1.2, 2.1

[58] T. F. Cootes, C. J. Taylor, D. H. Cooper, and J. Graham, "Active Shape Models - Their Training and Application," *CVIU*, vol. 61, no. 1, pp. 38–59, Jan. 1995. 1.2, 2.1

[59] G. B. Huang, M. Ramesh, T. Berg, and E. Learned-Miller, "Labeled Faces in the Wild: A Database for Studying Face Recognition in Unconstrained Environments," University of Massachusetts, Amherst, Tech. Rep. 07-49, Oct. 2007. 1.2, 2.2, 3.2.3, 4.2

[60] "Labeled Faces in the Wild," http://vis-www.cs.umass.edu/lfw/index.html. 1.2, 2.2, 3.2.3, 4.2

[61] "U.S. Department of Transportation Federal Highway Administration (FHWA)," https://www.fhwa.dot.gov/. 1.2, 5

[62] "The Exploratory Advanced Research Program Automated Video Feature Extraction Workshop Summary Report," http://www.fhwa.dot.gov/advancedresearch/pubs/13037/001.cfm, Dec. 2012. 1.2, 5.1

[63] O. Celiktutan, S. Ulukaya, and B. Sankur, "A comparative study of face landmarking techniques," *EURASIP Journal on Image and Video Processing*, vol. 2013, Mar. 2013. 2

[64] I. Matthews and S. Baker, “Active Appearance Models Revisited,” *IJCV*, vol. 60, no. 2, pp. 135–164, Nov. 2004. 2.1, 2.1

[65] T. F. Cootes, C. J. Taylor, and A. Lanitis, “Active Shape Models : Evaluation of a Multi-Resolution Method for Improving Image Search,” in *Proceedings of the British Machine Vision Conference (BMVC)*, Sep. 1994, pp. 32.1–32.10. 2.1

[66] T. F. Cootes and C. J. Taylor, “Statistical Models of Appearance for Computer Vision,” http://www.face-rec.org/algorithms/AAM/app_models.pdf, Mar. 2004. 2.1, 2.1, 3.1.3

[67] D. Cristinacce and T. Cootes, “Feature Detection and Tracking with Constrained Local Models,” in *Proceedings of the British Machine Vision Conference (BMVC)*, Sep. 2006, pp. 929–938. 2.1

[68] ——, “Automatic feature localisation with constrained local models,” *Pattern Recognition*, vol. 41, no. 10, pp. 3054–3067, Oct. 2008. 2.1

[69] J. M. Saragih, S. Lucey, and J. F. Cohn, “Deformable Model Fitting by Regularized Landmark Mean-Shift,” *IJCV*, vol. 91, no. 2, pp. 200–215, Jan. 2011. 2.1, 2.2

[70] S. Milborrow and F. Nicolls, “Locating Facial Features with an Extended Active Shape Model,” in *Proceedings of the European Conference on Computer Vision (ECCV), Part IV - Volume 5305 of the series Lecture Notes in Computer Science (LNCS)*, Oct. 2008, pp. 504–513. 2.1

[71] T. F. Cootes, G. Edwards, and C. J. Taylor, “Comparing Active Shape Models with Active Appearance Models,” in *Proceedings of the British Machine Vision Conference (BMVC)*, Sep. 1999, pp. 18.1–18.10. 2.1

[72] R. Gross, I. Matthews, and S. Baker, “Generic vs. Person Specific Active Appearance Models,” *Image and Vision Computing*, vol. 23, no. 11, pp. 1080–1093, Nov. 2005. 2.1

[73] X. Liu, “Discriminative Face Alignment,” *IEEE Transactions on Pattern Analysis and Ma-*

chine Intelligence (TPAMI), vol. 31, no. 11, pp. 1941–1954, Nov. 2009. 2.1

[74] G. Tzimiropoulos, J. A. Medina, S. Zafeiriou, and M. Pantic, “Generic Active Appearance Models Revisited,” in *Proceedings of the Asian Conference on Computer Vision (ACCV) - Volume Part III*, Nov. 2012, pp. 650–663. 2.1

[75] G. Tzimiropoulos and M. Pantic, “Optimization problems for fast AAM fitting in-the-wild,” in *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, Dec. 2013, pp. 593–600. 2.1, 3.2.3, 4.2

[76] T. F. Cootes, G. V. Wheeler, K. N. Walker, and C. J. Taylor, “View-Based Active Appearance Models,” in *Proceedings of the IEEE International Conference on Face and Gesture Recognition (FG)*, Mar. 2000, pp. 227–232. 2.1

[77] S. Romdhani, S. Gong, and A. Psarrou, “A Multi-View Nonlinear Active Shape Model Using Kernel PCA,” in *Proceedings of the British Machine Vision Conference (BMVC)*, Sep. 1999, pp. 48.1–48.10. 2.1

[78] L. Zhang and H. Ai, “Multi-View Active Shape Model with Robust Parameter Estimation,” in *Proceedings of the ICPR - Volume 4*, Aug. 2006, pp. 469–472. 2.1

[79] S. Milborrow, T. E. Bishop, and F. Nicolls, “Multiview Active Shape Models with SIFT Descriptors for the 300-W Face Landmark Challenge,” in *Proceedings of the IEEE International Conference on Computer Vision Workshops (ICCVW)*, Dec. 2013, pp. 378–385. 2.1

[80] Y. Zhou, L. Gu, and Hong-Jiang Zhang, “Bayesian Tangent Shape Model: Estimating Shape and Pose Parameters via Bayesian Inference,” in *Proceedings of the IEEE International Conference on Computer Vision and Pattern Recognition (CVPR)*, Jun. 2003, pp. 109–116. 2.1

[81] L. Gu and T. Kanade, “A Generative Shape Regularization Model for Robust Face Alignment,” in *Proceedings of the European Conference on Computer Vision (ECCV), Part I -*

Volume 5302 of the series Lecture Notes in Computer Science (LNCS), Oct. 2008, pp. 413–426. 2.1

[82] A. M. Martinez and R. Benavente, “The AR Face Database,” The Computer Vision Center (CVC) at the Universitat Autonoma de Barcelona, Tech. Rep. CVC #24, Jun. 1998. 2.1, 6.1.1

[83] A. M. Martinez, “AR Face Database,” http://www2.ece.ohio-state.edu/~aleix/ARdatabase.html. 2.1, 6.1.1

[84] M. Everingham, J. Sivic, and A. Zisserman, ““Hello! My name is... Buffy” – Automatic Naming of Characters in TV Video,” in *Proceedings of the British Machine Vision Conference (BMVC)*, Sep. 2006, pp. 889–908. 2.2, 2.5

[85] P. Viola and M. Jones, “Rapid Object Detection using a Boosted Cascade of Simple Features,” in *Proceedings of the IEEE International Conference on Computer Vision and Pattern Recognition (CVPR) - Volume 1*, Jun. 2001, pp. 511–518. 2.2, 5.5.1

[86] M. Dantone, J. Gall, G. Fanelli, and L. V. Gool, “Real-time Facial Feature Detection using Conditional Regression Forests,” in *Proceedings of the IEEE International Conference on Computer Vision and Pattern Recognition (CVPR)*, Jun. 2012, pp. 2578–2585. 2.2, 2.5, 3.2.3, 4.2

[87] F. Zhou, J. Brandt, and Z. Lin, “Exemplar-based Graph Matching for Robust Facial Landmark Localization,” in *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, Dec. 2013, pp. 1025–1032. 2.2

[88] B. M. Smith, J. Brandt, Z. Lin, and L. Zhang, “Nonparametric Context Modeling of Local Appearance for Pose and Expression-Robust Facial Landmark Localization,” in *Proceedings of the IEEE International Conference on Computer Vision and Pattern Recognition (CVPR)*, Jun. 2014, pp. 1741–1748. 2.2, 3.1.3

[89] M. C. Roh, T. Oguri, and T. Kanade, “Face Alignment Robust to Occlusion,” in *Proceedings*

of the IEEE International Conference on Face and Gesture Recognition (FG) Workshops, Mar. 2011, pp. 239–244. 2.2

[90] M. A. Fischler and R. C. Bolles, "Random Sample Consensus: A Paradigm for Model Fitting with Applications to Image Analysis and Automated Cartography," *Communications of the ACM*, vol. 24, no. 6, pp. 381–395, Jun. 1981. 2.2

[91] X. Yu, F. Yang, J. Huang, and D. N. Metaxas, "Explicit Occlusion Detection based Deformable Fitting for Facial Landmark Localization," in *Proceedings of the IEEE International Conference on Face and Gesture Recognition (FG) Workshops*, Apr. 2013, pp. 1–6. 2.2

[92] P. F. Felzenszwalb, R. B. Girshick, D. McAllester, and D. Ramanan, "Object Detection with Discriminatively Trained Part Based Models," *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, vol. 32, no. 9, pp. 1627–1645, Sep. 2010. 2.2, 5.5.1, 5.5.2

[93] N. Dalal and B. Triggs, "Histograms of Oriented Gradients for Human Detection," in *Proceedings of the IEEE International Conference on Computer Vision and Pattern Recognition (CVPR) - Volume 1*, Jun. 2005, pp. 886–893. 2.2, 3.1.1, 5.5.2

[94] G. Ghiasi and C. C. Fowlkes, "Occlusion Coherence: Localizing Occluded Faces with a Hierarchical Deformable Part Model," in *Proceedings of the IEEE International Conference on Computer Vision and Pattern Recognition (CVPR)*, Jun. 2014, pp. 1899–1906. 2.2, 7.1.2, 7.1.5

[95] A. Asthana, S. Zafeiriou, S. Cheng, and M. Pantic, "Robust Discriminative Response Map Fitting with Constrained Local Models," in *Proceedings of the IEEE International Conference on Computer Vision and Pattern Recognition (CVPR)*, Jun. 2013, pp. 3444–3451. 2.2, 3.2.3, 4.2, 5.4.1

[96] X. Xiong and F. De la Torre, "Supervised Descent Method and its Application to Face Alignment," in *Proceedings of the IEEE International Conference on Computer Vision and*

Pattern Recognition (CVPR), Jun. 2013, pp. 532–539. 2.2, 3.2.3, 5.3.1

[97] D. G. Lowe, "Distinctive Image Features from Scale-Invariant Keypoints," *International Journal of Computer Vision (IJCV)*, vol. 60, no. 2, pp. 91–110, Nov. 2004. 2.2

[98] X. Cao, Y. Wei, F. Wen, and J. Sun, "Face Alignment by Explicit Shape Regression," in *Proceedings of the IEEE International Conference on Computer Vision and Pattern Recognition (CVPR)*, Jun. 2012, pp. 2887–2894. 2.3

[99] J. Yan, Z. Lei, D. Yi, and S. Z. Li, "Learn to Combine Multiple Hypotheses for Accurate Face Alignment," in *Proceedings of the IEEE International Conference on Computer Vision Workshops (ICCVW)*, Dec. 2013, pp. 392–396. 2.3, 3.1.1

[100] S. Ren, X. Cao, Y. Wei, and J. Sun, "Face Alignment at 3000 FPS via Regressing Local Binary Features," in *Proceedings of the IEEE International Conference on Computer Vision and Pattern Recognition (CVPR)*, Jun. 2014, pp. 1685–1692. 2.3

[101] V. Kazemi and J. Sullivan, "One Millisecond Face Alignment with an Ensemble of Regression Trees," in *Proceedings of the IEEE International Conference on Computer Vision and Pattern Recognition (CVPR)*, Jun. 2014, pp. 1867–1874. 2.3

[102] O. Jesorsky, K. Kirchberg, and R. Frischholz, "Robust Face Detection Using the Hausdorff Distance," in *Proceedings of the International Conference on Audio- and Video-Based Biometric Person Authentication (AVBPA)*, Jun. 2001, pp. 90–95. 2.3

[103] "The BioID Face Database," https://www.bioid.com/About/BioID-Face-Database. 2.3

[104] V. Le, J. Brandt, Z. Lin, L. Bourdev, and T. S. Huang, "Interactive Facial Feature Localization," in *Proceedings of the European Conference on Computer Vision (ECCV), Part III - Volume 7574 of the series Lecture Notes in Computer Science (LNCS)*, Oct. 2012, pp. 679–692. 2.3, 3.2.2

[105] ——, "Helen dataset," http://www.ifp.illinois.edu/~vuongle2/helen/, 2012. 2.3, 3.2.2

[106] Y. Sun, X. Wang, and X. Tang, "Deep Convolutional Network Cascade for Facial Point Detection," in *Proceedings of the IEEE International Conference on Computer Vision and Pattern Recognition (CVPR)*, Jun. 2013, pp. 3476–3483. 2.4, 2.5, 7.1.4

[107] E. Zhou, H. Fan, Z. Cao, Y. Jiang, and Q. Yin, "Extensive Facial Landmark Localization with Coarse-to-fine Convolutional Network Cascade," in *Proceedings of the IEEE International Conference on Computer Vision Workshops (ICCVW)*, Dec. 2013, pp. 386–391. 2.4, 7.1.4

[108] Z. Zhang, P. Luo, C. C. Loy, and X. Tang, "Facial Landmark Detection by Deep Multi-task Learning," in *Proceedings of the European Conference on Computer Vision (ECCV), Part VI - Volume 8694 of the series Lecture Notes in Computer Science (LNCS)*, Sep. 2014, pp. 94–108. 2.4, 7.1.4

[109] K. Pearson, "On Lines and Planes of Closest Fit to Systems of Points in Space," *Philosophical Magazine*, vol. 2, no. 6, pp. 559–572, 1901. 3.1.1

[110] H. Hotelling, "Analysis of a Complex of Statistical Variables Into Principal Components," *Journal of Educational Psychology*, vol. 24, no. 6, pp. 417–441, 498–520, Oct. 1933. 3.1.1

[111] ——, "Relations Between Two Sets of Variates," *Biometrika*, vol. 28, no. 3/4, pp. 321–377, Dec. 1936. 3.1.1

[112] B. Tamersoy, C. Hu, and J. K. Agarwal, "Nonparametric Facial Feature Localization," in *Proceedings of the IEEE International Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, Jun. 2013, pp. 838–845. 3.1.1

[113] X. Zhao, S. Shan, X. Chai, and X. Chen, "Cascaded Shape Space Pruning for Robust Facial Landmark Detection," in *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, Dec. 2013, pp. 1033–1040. 3.1.1, 3.1.1

[114] J. Friedman, T. Hastie, and R. Tibshirani, "Additive Logistic Regression: a Statistical View of Boosting," *Annals of Statistics*, vol. 28, no. 2, pp. 337–407, Dec. 2000. 3.1.1, 5.5.2, A

[115] Y. Freund and R. E. Schapire, “A Short Introduction to Boosting,” *Journal of Japanese Society for Artificial Intelligence*, vol. 14, no. 5, pp. 771–780, Sep. 1999. 3.1.1, 5.5.1, 5.5.2

[116] R. E. Schapire and Y. Singer, “Improved Boosting Algorithms Using Confidence-rated Predictions,” *Machine Learning*, vol. 37, no. 3, pp. 297–336, Dec. 1999. 3.1.1, 5.5.2, A

[117] C. Cortes and V. Vapnik, “Support-Vector Networks,” *Machine Learning*, vol. 20, no. 3, pp. 273–297, Sep. 1995. 3.1.1, 5.5.1, 5.5.2

[118] L. Breiman, “Random Forests,” *Machine Learning*, vol. 45, no. 1, pp. 5–32, Oct. 2001. 3.1.1, 5.5.2

[119] M. Özuysal, P. Fua, and V. Lepetit, “Fast Keypoint Recognition in Ten Lines of Code,” in *Proceedings of the IEEE International Conference on Computer Vision and Pattern Recognition (CVPR)*, Jun. 2007, pp. 1–8. 3.1.1

[120] X. Zhao, X. Chai, Z. Niu, C. Heng, and S. Shan, “Context modeling for facial landmark detection based on Non-Adjacent Rectangle (NAR) Haar-like feature,” *Image and Vision Computing*, vol. 30, no. 3, pp. 136–146, Mar. 2012. 3.1.1

[121] L. Zhang, H. Ai, S. Xin, C. Huang, S. Tsukiji, and S. Lao, “Robust Face Alignment Based on Local Texture Classifiers,” in *Proceedings of the IEEE International Conference on Image Processing (ICIP) - Volume 2*, Sep. 2005, pp. 354–357. 3.1.1

[122] L. Zhang, H. Ai, and S. Lao, “Robust Face Alignment Based on Hierarchical Classifier Network,” in *Proceedings of the Human Computer Interaction Workshop in conjunction with the European Conference on Computer Vision (ECCV) - Volume 3979 of the series Lecture Notes in Computer Science (LNCS)*, May 2006, pp. 1–11. 3.1.1

[123] J. C. Gower, “Generalized Procrustes Analysis,” *Psychometrika*, vol. 40, no. 1, pp. 33–51, Mar. 1975. 3.1.2

[124] C. Goodall, “Procrustes Methods in the Statistical Analysis of Shape,” *Journal of the Royal*

Statistical Society. Series B (Methodological), vol. 53, no. 2, pp. 285–339, 1991. 3.1.2

[125] E. J. Candes, J. K. Romberg, and T. Tao, "Stable Signal Recovery from Incomplete and Inaccurate Measurements," *Communications on Pure and Applied Mathematics*, vol. 59, no. 8, pp. 1207–1223, Aug. 2006. 3.1.3

[126] R. J. Tibshirani, "Regression Shrinkage and Selection via the Lasso," *Journal of the Royal Statistical Society. Series B (Methodological)*, vol. 58, no. 1, pp. 267–288, Jan. 1996. 3.1.3

[127] Y. Nesterov and A. Nemirovsky, *Interior-point Polynomial Algorithms in Convex Programming*, ser. Studies in Applied and Numerical Mathematics. Society for Industrial and Applied Mathematics (SIAM), 1994, vol. 13. 3.1.3

[128] S. J. Wright, *Primal-Dual Interior-Point Methods*. Society for Industrial and Applied Mathematics (SIAM), 1997. 3.1.3

[129] M. R. Osborne, B. Presnell, and B. A. Turlach, "A new approach to variable selection in least squares problems," *IMA Journal of Numerical Analysis*, vol. 20, no. 3, pp. 389–403, 2000. 3.1.3

[130] B. Efron, T. Hastie, I. Johsntone, and R. Tibshirani, "Least angle regression," *Annals of Statistics*, vol. 32, no. 2, pp. 407–499, 2004. 3.1.3

[131] T. Hastie, S. Rosset, R. Tibshirani, and J. Zhu, "The Entire Regularization Path for the Support Vector Machine," *Journal of Machine Learning Research*, vol. 5, pp. 1391–1415, Dec. 2004. 3.1.3

[132] N. Z. Shor, *Minimization Methods for Non-Differentiable Functions*, ser. Springer Series in Computational Mathematics. Springer-Verlag, 1985, vol. 3. 3.1.3

[133] B. Polyak, *Minimization Methods for Non-Differentiable Functions*, ser. Translations Series in Mathematics and Engineering. Optimization Software, 1987. 3.1.3

[134] Seung-Jean Kim, K. Koh, M. Lustig, S. Boyd, and D. Gorinevsky, "An Interior-Point

Method for Large-Scale ℓ_1-Regularized Least Squares," *IEEE Journal on Selected Topics in Signal Processing*, vol. 1, no. 4, pp. 606–617, Dec. 2007. 3.1.3, B, B, B

[135] "MATLAB," http://www.mathworks.com/products/matlab/. 3.1.3, 3.2.1, 3.2.3, 7.1.1

[136] Seung-Jean Kim, K. Koh, M. Lustig, S. Boyd, and D. Gorinevsky, "l1_ls: Simple Matlab Solver for l1-regularized Least Squares Problems," http://www.stanford.edu/~boyd/l1_ls/. 3.1.3

[137] Piotr Dollár, "Piotr's Computer Vision Matlab Toolbox (PMT)," http://vision.ucsd.edu/~pdollar/toolbox/doc/index.html. 3.2.1, 5.5.3

[138] C. Sagonas, G. Tzimiropoulos, S. Zafeiriou, and M. Pantic, "A semi-automatic methodology for facial landmark annotation," in *Proceedings of the IEEE International Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, Jun. 2013, pp. 896–903. 3.2.2, 3.11

[139] "Facial Point Annotations," http://ibug.doc.ic.ac.uk/resources/facial-point-annotations/. 3.2.2, 3.11

[140] G. Tzimiropoulos and M. Pantic, "Fitting AAMs In-The-Wild ICCV 2013," http://ibug.doc.ic.ac.uk/resources/fitting-aams-wild-iccv-2013/. 3.2.3

[141] X. Yu, J. Huang, S. Zhang, W. Yan, and D. N. Metaxas, "Pose-free Facial Landmark Fitting via Optimized Part Mixtures and Cascaded Deformable Shape Model," http://www.research.rutgers.edu/~xiangyu/face_align.html. 3.2.3

[142] A. Asthana, S. Zafeiriou, S. Cheng, and M. Pantic, "Discriminative Response Map Fitting DRMF 2013," http://ibug.doc.ic.ac.uk/resources/drmf-matlab-code-cvpr-2013/. 3.2.3

[143] X. Xiong and F. De la Torre, "Download intraFace (Matlab Functions)," http://www.humansensing.cs.cmu.edu/intraface/download_functions_matlab.html. 3.2.3, 4.2

[144] P. Viola and M. Jones, "Robust Real-Time Face Detection," *IJCV*, vol. 57, no. 2, pp. 137–

154, May 2004. 3.2.3, 4.2

[145] M. Mathias, R. Benenson, M. Pedersoli1, and L. V. Gool, “Face Detection Without Bells and Whistles,” http://markusmathias.bitbucket.org/2014_eccv_face_detection/. 3.2.6

[146] U. Prabhu, K. Seshadri, and M. Savvides, “Automatic Facial Landmark Tracking in Video Sequences using Kalman Filter Assisted Active Shape Models,” in *Trends and Topics in Computer Vision - Proceedings of the Workshop on Human Motion in Conjunction with the European Conference on Computer Vision (ECCV), Part I - Volume 6553 of the series Lecture Notes in Computer Science (LNCS)*, Sep. 2010, pp. 86–99. 4.1.1, 5.3.1

[147] F. L. Bookstein, “Principal Warps: Thin-Plate Splines and the Decomposition of Deformations,” *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, vol. 11, no. 6, pp. 567–585, Jun. 1989. 4.1.1

[148] G. Wahba, *Spline Models for Observational Data*, ser. CBMS-NSF Regional Conference Series in Applied Mathematics. Society for Industrial and Applied Mathematics (SIAM), 1990, vol. 59. 4.1.1

[149] C. T. Loop, “Smooth Subdivision Surfaces Based on Triangles,” Master’s thesis, University of Utah, Department of Mathematics, 1987. 4.1.1

[150] V. Blanz and T. Vetter, “A morphable model for the synthesis of 3D faces,” in *Proceedings of the Annual Conference on Computer Graphics and Interactive Techniques (SIGGRAPH)*, Aug. 1999, pp. 187–194. 4.1, 4.1.2

[151] S. Gupta, M. K. Markey, and A. C. Bovik, “Anthropometric 3D Face Recognition,” *International Journal of Computer Vision (IJCV)*, vol. 90, no. 3, pp. 331–349, Dec. 2010. 4.1, 4.1.2

[152] S. Gupta, K. R. Castleman, M. K. Markey, and A. C. Bovik, “Texas 3D Face Recognition Database,” in *Proceedings of the IEEE Southwest Symposium on Image Analysis Interpretation (SSIAI)*, May 2010, pp. 97–100. 4.1, 4.1.2

[153] ——, “Texas 3D Face Recognition Database (Texas 3DFRD),” http://live.ece.utexas.edu/research/texas3dfr/index.htm. 4.1, 4.1.2

[154] P. J. Phillips, P. J. Flynn, T. Scruggs, K. W. Bowyer, J. Chang, K. Hoffman, J. Marques, J. Min, and W. Worek, “Overview of the Face Recognition Grand Challenge,” in *Proceedings of the IEEE International Conference on Computer Vision and Pattern Recognition (CVPR)*, Jun. 2005, pp. 947–954. 4.1, 4.1.2

[155] National Institute of Standards and Technology (NIST), “Face Recognition Grand Challenge,” http://www.nist.gov/itl/iad/ig/frgc.cfm. 4.1, 4.1.2

[156] “Online Mugshots Database,” http://www.mugshots.com. 4.1, 4.1.2

[157] M. Aharon, M. Elad, and A. Bruckstein, “K-SVD: An Algorithm for Designing Overcomplete Dictionaries for Sparse Representation,” *IEEE Transactions on Signal Processing (TSP)*, vol. 54, no. 11, pp. 4311–4322, Nov. 2006. 4.1.2

[158] C. Xie, M. Savvides, and B. V. K. V. Kumar, “Redundant Class-Dependence Feature Analysis Based on Correlation Filters Using FRGC2.0 Data,” in *Proceedings of the IEEE International Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, Jun. 2005, pp. 153–158. 4.1.3

[159] K. Q. Weinberger, J. Blitzer, and L. K. Saul, “Distance Metric Learning for Large Margin Nearest Neighbor Classification,” in *Proceedings of the Conference on Neural Information Processing Systems (NIPS)*, Dec. 2005, pp. 1473–1480. 4.1.3

[160] K. Q. Weinberger and L. K. Saul, “Distance Metric Learning for Large Margin Nearest Neighbor Classification,” *Journal of Machine Learning Research (JMLR)*, vol. 10, pp. 207–244, Feb. 2009. 4.1.3

[161] G. B. Huang, V. Jain, and E. Learned-Miller, “Unsupervised Joint Alignment of Complex Images,” in *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, Oct. 2007, pp. 1–8. 4.2

[162] G. B. Huang, M. Mattar, H. Lee, and E. Learned-Miller, “Learning to Align from Scratch,” in *Proceedings of the Conference on Neural Information Processing Systems (NIPS)*, Dec. 2012. 4.2

[163] M. Dantone, “Facial Features - LFW,” http://www.dantone.me/datasets/facial-features-lfw/. 4.2

[164] M. Dantone, J. Gall, G. Fanelli, and L. V. Gool, “Facial Feature Detection,” http://www.dantone.me/projects-2/facial-feature-detection/. 4.2

[165] A. Asthana, “3D Head Pose Estimator for Matlab,” https://sites.google.com/site/akshayasthana/codes. 4.2, 5.4.1, 5.4, 5.5, 5.6, 5.4.1

[166] J. Ruiz del Solar, R. Verschae, and M. Correa, “Recognition of Faces in Unconstrained Environments: A Comparative Study,” *EURASIP Journal on Advances in Signal Processing (Recent Advances in Biometric Systems: A Signal Processing Perspective)*, vol. 2009, Jan. 2009. 4.3

[167] H. J. Seo and P. Milanfar, “Face Verification Using the LARK Representation,” *IEEE Transactions on Information Forensics and Security (TIFS)*, vol. 6, no. 4, pp. 1275–1286, Dec. 2011. 4.3

[168] G. Sharma, S. ul Hussain, and F. Jurie, “Local Higher-Order Statistics (LHS) for Texture Categorization and Facial Analysis,” in *Proceedings of the European Conference on Computer Vision (ECCV), Volume II - Volume 7578 of the series Lecture Notes in Computer Science (LNCS)*, Oct. 2012, pp. 1–12. 4.3

[169] S. R. Arashloo and J. Kittler, “Efficient Processing of MRFs for Unconstrained-Pose Face Recognition,” in *Proceedings of the IEEE International Conference on Biometrics: Theory, Applications and Systems (BTAS)*, Sep. 2013, pp. 1–8. 4.3, 4.2

[170] D. Yi, Z. Lei, and S. Z. Li, “Towards Pose Robust Face Recognition,” in *Proceedings of the IEEE International Conference on Computer Vision and Pattern Recognition (CVPR)*, Jun.

2013, pp. 3539–3545. 4.3, 4.2

[171] "Official US Government Website for Distracted Driving," http://www.distraction.gov/stats-research-laws/facts-and-statistics.html. 5.1

[172] "SHRP 2 Naturalistic Driving Study (SHRP 2 NDS)," http://www.shrp2nds.us/. 5.1

[173] "The second Strategic Highway Research Program (2006-2015)," http://www.trb.org/StrategicHighwayResearchProgram2SHRP2/Blank2.aspx. 5.1

[174] "Virginia Tech Transportation Institute," http://www.vtti.vt.edu/. 5.1

[175] K. Cambell, "The SHRP2 Naturalistic Driving Study," in *TR News*, vol. 282, Sep. 2012, pp. 30–35. 5.1

[176] "Analyzing Driver Behavior Using Data from the SHRP2 Naturalistic Driving Study," http://onlinepubs.trb.org/onlinepubs/shrp2/SHRP2_PB_S08_2013-05.pdf, May 2013. 5.1

[177] "Transportation Research Board of the National Academies of Science, The 2$^{\text{nd}}$ Strategic Highway Research Program Naturalistic Driving Study Dataset," Available from the SHRP2 NDS InSight Data Access Website: https://insight.shrp2nds.us, 2013. 5.1

[178] R. E. Kalman, "A New Approach to Linear Filtering and Prediction Problems," *Transactions of the ASME – Journal of Basic Engineering*, vol. 82, no. Series D, pp. 35–45, 1960. 5.3.1

[179] Y. Bar-Shalom, , X. R. Li, and T. Kirubarajan, *Estimation with Applications to Tracking and Navigation*. John Wiley & Sons, Inc., 2001. 5.3.1

[180] A. M. Baumberg and D. C. Hogg, "An Efficient Method for Contour Tracking using Active Shape Models," in *Proceedings of the IEEE Workshop on Motion of Nonrigid and Articulated Objects*, Nov. 1994, pp. 194–199. 5.3.1

[181] Adam Baumberg, "Hierarchical shape fitting using an iterated linear filter," *Image and Vision Computing*, vol. 16, no. 5, pp. 329–335, Apr. 1998. 5.3.1

[182] S. W. Lee, J. Kang, J. Shin, and J. Paik, "Hierarchical active shape model with motion

prediction for real-time tracking of non-rigid objects," *IET Computer Vision*, vol. 1, no. 1, pp. 17–24, Apr. 2007. 5.3.1

[183] Jörgen Ahlberg, "An Active Model for Facial Feature Tracking," *EURASIP Journal on Applied Signal Processing*, vol. 2002, no. 6, pp. 566–571, Jan. 2002. 5.3.1

[184] B. Pu, S. Liang, Y. Xie, Z. Yi, and Pheng-Ann Heng, "Video Facial Feature Tracking with Enhanced ASM and Predicted Meanshift," in *Proceedings of the 2^{nd} International Conference on Computer Modeling and Simulation*, vol. 2, Jan. 2010, pp. 151–155. 5.3.1

[185] D. Comaniciu, V. Ramesh, and P. Meer, "Real-Time Tracking of Non-Rigid Objects using Mean Shift," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Jun. 2000, pp. 142–149. 5.3.1

[186] J. Paone, D. Bolme, R. Ferrell, D. Aykac, and T. Karnowski, "Baseline Face Detection, Head Pose Estimation, and Coarse Direction Detection for Facial Data in the SHRP2 Naturalistic Driving Study," in *Proceedings of the 2015 IEEE Intelligent Vehicles Symposium (IV)*, Jun. 2015, pp. 174–179. 5.3.3, 5.6

[187] D. L. Strayer, F. A. Drews, and D. J. Crouch, "A Comparison of the Cell Phone Driver and the Drunk Driver," *Human factors: The Journal of the Human Factors and Ergonomics Society*, vol. 48, no. 2, pp. 381–391, 2006. 5.5.1

[188] D. L. Strayer and F. A. Drews, "Cell-Phone–Induced Driver Distraction," *Current Directions in Psychological Science*, vol. 16, no. 3, pp. 128–131, 2007. 5.5.1

[189] Y. Artan, O. Bulan, R. P. Loce, and P. Paul, "Driver Cell Phone Usage Detection from HOV/HOT NIR Images," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, Jun. 2014, pp. 225–230. 5.5.1, 5.5.3

[190] X. Zhang, N. Zheng, F. Wang, and Y. He, "Visual Recognition of Driver Hand-held Cell Phone Use Based on Hidden CRF," in *Proceedings of the IEEE International Conference on Vehicular Electronics and Safety (ICVES)*, Jul. 2011, pp. 248–251. 5.5.1, 5.5.3

[191] C. Bo, X. Jian, X. Li, X. Mao, Y. Wang, and F. Li, “You’re Driving and Texting: Detecting Drivers Using Personal Smart Phones by Leveraging Inertial Sensors,” in *Proceedings of the Annual International Conference on Mobile Computing and Networking (MobiCom)*, Sep. 2013, pp. 199–202. 5.5.1

[192] J. Yang, S. Sidhom, G. Chandrasekaran, T. Vu, H. Liu, N. Cecan, Y. Chen, M. Gruteser, and R. P. Martin, “Detecting Driver Phone Use Leveraging Car Speakers,” in *Proceedings of the Annual International Conference on Mobile Computing and Networking (MobiCom)*, Sep. 2011, pp. 97–108. 5.5.1

[193] D. S. Breed and W. E. Duvall, “In-vehicle driver cell phone detector,” http://www.google.com/patents/US8731530, May 2014, uS Patent 8731530 B1. 5.5.1

[194] A. Jaiantilal, “randomforest-matlab,” http://code.google.com/p/randomforest-matlab/. 5.5.3

[195] Chih-Chung Chang and Chih-Jen Lin, “LIBSVM: A Library for Support Vector Machines,” *ACM Transactions on Intelligent Systems and Technology (TIST)*, vol. 2, no. 3, pp. 27:1–27:27, Apr. 2011, software available at http://www.csie.ntu.edu.tw/~cjlin/libsvm/. 5.5.3

[196] ——, “LIBSVM – A Library for Support Vector Machines,” http://www.csie.ntu.edu.tw/~cjlin/libsvm/. 5.5.3

[197] X. Liu, P. H. Tu, and F. W. Wheeler, “Face Model Fitting on Low Resolution Images,” in *Proceedings of the British Machine Vision Conference (BMVC)*, Sep. 2006, pp. 1079–1088. 6

[198] G. Dedeoğlu, S. Baker, and T. Kanade, “Resolution-Aware Fitting of Active Appearance Models to Low Resolution Images,” in *Proceedings of the European Conference on Computer Vision (ECCV), Volume II - Volume 3952 of the series Lecture Notes in Computer Science (LNCS)*, May 2006, pp. 83–97. 6

[199] C. Qu, E. Monari, and T. Schuchert, “Resolution-Aware Constrained Local Model with Mixture of Local Experts,” in *Proceedings of the IEEE International Conference on Ad-*

vanced Video and Signal Based Surveillance (AVSS), Aug. 2013, pp. 454–459. 6

[200] A. Asthana, S. Zafeiriou, G. Tzimiropoulos, S. Cheng, and M. Pantic, “From Pixels to Response Maps: Discriminative Image Filtering for Face Alignment in the Wild,” *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, vol. 37, no. 6, pp. 1312–1320, Jun. 2015. 6

[201] P. J. Phillips, P. J. Rauss, and S. Z. Der, “FERET (Face Recognition Technology) Recognition Algorithm Development and Test Results,” Army Research Laboratory (ARL), Tech. Rep. ARL-TR-995, Oct. 1996. 6.1.1

[202] S. Rizvi, P. J. Phillips, and H. Moon, “The FERET Verification Testing Protocol for Face Recognition Algorithms,” National Institute of Standards and Technology (NIST), Tech. Rep. NISTIR 6281, Oct. 1998. 6.1.1

[203] P. J. Phillips, H. Moon, S. Rizvi, and P. J. Rauss, “The FERET Evaluation Methodology for Face-Recognition Algorithms,” *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, vol. 22, no. 10, pp. 1090–1104, Oct. 2000. 6.1.1

[204] National Institute of Standards and Technology (NIST), “The Facial Recognition Technology (FERET) Database,” http://www.itl.nist.gov/iad/humanid/feret/feret_master.html. 6.1.1

[205] Y. Taigman, M. Yang, M. Ranzato, and L. Wolf, “DeepFace: Closing the Gap to Human-Level Performance in Face Verification,” in *Proceedings of the IEEE International Conference on Computer Vision and Pattern Recognition (CVPR)*, Jun. 2014, pp. 1701–1708. 7.1.4

[206] F. Schroff, D. Kalenichenko, and J. Philbin, “FaceNet: A Unified Embedding for Face Recognition and Clustering,” in *Proceedings of the IEEE International Conference on Computer Vision and Pattern Recognition (CVPR)*, Jun. 2015, pp. 815–823. 7.1.4

[207] J. Zhang, S. Shan, M. Kan, and X. Chen, “Coarse-to-Fine Auto-encoder Networks (CFAN) for Real-time Face Alignment,” in *Proceedings of the European Conference on Computer*

Vision (ECCV), Part II - Volume 8690 of the series Lecture Notes in Computer Science (LNCS), Sep. 2014, pp. 1–16. 7.1.4

[208] M. Kostinger, P. Wohlhart, P. M. Roth, and H. Bischof, "Annotated Facial Landmarks in the Wild: A Large-scale, Real-world Database for Facial Landmark Localization," in *Proceedings of the Workshop on Benchmarking Facial Image Analysis Technologies (BeFIT) in conjunction with the IEEE ICCV*, Nov. 2011, pp. 2144–2151. 7.1.4

[209] ——, "Annotated Facial Landmarks in the Wild: A Large-scale, Real-world Database for Facial Landmark Localization," http://lrs.icg.tugraz.at/research/aflw/. 7.1.4

[210] H. Liy, Z. Linz, J. Brandtz, X. Shenz, and G. Huay, "Efficient Boosted Exemplar-based Face Detection," in *Proceedings of the IEEE International Conference on Computer Vision and Pattern Recognition (CVPR)*, Jun. 2014, pp. 1843–1850. 7.1.5

[211] H. Liy, Z. Linz, X. Shenz, J. Brandtz, and G. Huay, "A Convolutional Neural Network Cascade for Face Detection," in *Proceedings of the IEEE International Conference on Computer Vision and Pattern Recognition (CVPR)*, Jun. 2015, pp. 5325–5334. 7.1.5

[212] X. Shen, Z. Lin, J. Brandt, and Y. Wu, "Detecting and Aligning Faces by Image Retrieval," in *Proceedings of the IEEE International Conference on Computer Vision and Pattern Recognition (CVPR)*, Jun. 2013, pp. 3460–3467. 7.1.5

[213] D. Chen, S. Ren, Y. Wei, X. Cao, and J. Sun, "Joint Cascade Face Detection and Alignment," in *Proceedings of the European Conference on Computer Vision (ECCV), Part VI - Volume 8694 of the series Lecture Notes in Computer Science (LNCS)*, Sep. 2014, pp. 109–122. 7.1.5

[214] S. M. Bileschi, "StreetScenes: Towards Scene Understanding in Still Images," Ph.D. dissertation, Massachusetts Institute of Technology, 2006. 7.1.8

[215] ——, "CBCL StreetScenes Challenge Framework," http://cbcl.mit.edu/software-datasets/streetscenes/. 7.1.8

[216] V. N. Boddeti, T. Kanade, and B. V. K. V. Kumar, “Correlation Filters for Object Alignment,” in *Proceedings of the IEEE International Conference on Computer Vision and Pattern Recognition (CVPR)*, Jun. 2013, pp. 2291–2298. 7.1.8

[217] V. N. Boddeti, T. Kanade, B. V. K. V. Kumar, and Y. Li, “Object Alignment,” http://vishnu.boddeti.net/projects/alignment. 7.1.8

[218] L. Vandenberghe and S. Boyd, “A primal-dual potential reduction method for problems involving matrix inequalities,” *Mathematical Programming, Series B*, vol. 69, no. 1, pp. 205–236, Jul. 1995. B

[219] L. F. Portugal, M. G. C. Resende, G. Veiga, and J. J. Judice, “A Truncated Primal-Infeasible Dual-Feasible Network interior point method,” *Networks*, vol. 35, no. 2, pp. 91–108, Feb. 2000. B

[220] C. A. Johnson, J. Seidel, and A. Sofer, “Interior-Point Methodology for 3-D PET Reconstruction,” *IEEE Transactions on Medical Imaging*, vol. 19, no. 4, pp. 271–285, Apr. 2000. B

[221] K. Koh, Seung-Jean Kim, and S. Boyd, “An Interior-Point Method for Large-Scale ℓ_1-Regularized Logistic Regression,” *Journal of Machine Learning Research*, vol. 8, pp. 1519–1555, Jul. 2007. B