Exploiting Space-Time Statistics of Videos for Face "Hallucination"

Göksel Dedeoğlu

CMU-RI-TR-07-05

Submitted in partial fulfillment of the
requirements for the degree of
Doctor of Philosophy in Robotics

The Robotics Institute
Carnegie Mellon University
Pittsburgh, Pennsylvania 15213

April 2007

Thesis Committee
Takeo Kanade, Chair
Simon Baker
Jonas August
Henry W. Schneiderman
William T. Freeman (MIT)

Abstract

Face "Hallucination" aims to recover high quality, high-resolution images of human faces from low-resolution, blurred, and degraded images or video. This thesis presents person-specific solutions to this problem through careful exploitation of space (image) and space-time (video) models. The results demonstrate accurate restoration of facial details, with resolution enhancements upto a scaling factor of 16.

The algorithms proposed in this thesis follow the analysis-by-synthesis paradigm; they explain the observed (low-resolution) data by fitting a (high-resolution) model. In this context, the first contribution is the discovery of a scaling-induced bias that plagues most model-to-image (or image-to-image) fitting algorithms. It was found that models and observations should be treated *asymmetrically*, both to formulate an unbiased objective function and to derive an accurate optimization algorithm. This asymmetry is most relevant to Face Hallucination: when applied to the popular Active Appearance Model, it leads to a novel face tracking and reconstruction algorithm that is significantly more accurate than state-of-the-art methods. The analysis also reveals the inherent trade-off between computational efficiency and estimation accuracy in low-resolution regimes.

The second contribution is a statistical generative model of face videos. By treating a video as a composition of space-time patches, this model efficiently encodes the temporal dynamics of complex visual phenomena such as eye-blinks and the occlusion or appearance of teeth. The same representation is also used to define a data-driven prior on a three-dimensional Markov Random Field in space and time. Experimental results demonstrate that temporal representation and reasoning about facial expressions improves robustness by regularizing the Face Hallucination problem.

The final contribution is an approximate compensation scheme against illumination effects. It is observed that distinct illumination subspaces of a face (each coming from a different pose and expression) still exhibit similar variation with respect to illumination. This motivates augmenting the video model with a low-dimensional illumination subspace, whose parameters are estimated jointly with high-resolution face details. Successful Face Hallucinations beyond the lighting conditions of the training videos are reported.

Acknowledgments

I would like to thank my advisor Takeo Kanade for the many years of mentorship. His enthusiasm, persistence and never-ending energy in tackling research problems have been most inspiring. I am indebted to him for his careful guidance and feedback as I explored the various avenues of my thesis research. I could always count on his support, both academically and financially.

I am also grateful to Simon Baker for two memorable years of collaboration. His mastery in computer vision and patience in listening made our meetings both instructive and enjoyable. I would also like to thank Jonas August for his mentorship and collaboration. His rigor and thoroughness in formulating vision problems have been among the most influential in my graduate years. Many thanks to Henry Schneiderman and Bill Freeman for joining my thesis committee and providing insightful comments.

Special thanks to Iain Matthews for providing the AAM-toolbox. The visuals of my presentations have been greatly enhanced by the many tools he developed and kindly shared.

At the Robotics Institute, my interaction with Martial Hebert and the members of his "misc-reading" group has been extremely enriching. Thank you all for your critiques and feedback on the earliest incarnations of my thesis research. I have also enjoyed the friendship and assistance of many. In particular, I would like to thank Louise Ditmore, Tim Doebler, Sanjiv Kumar, Bilge Mutlu, Raju Patil, Hulya Yalcin and Chuck Rosenberg.

Finally, I don't think I would have survived the graduate school without the social and family support that I have enjoyed throughout the years. My wife, in particular, has been among the most patient during the doctoral "process". Thank you, Susan, for being there.

Table of Contents

List of Figures

Chapter 1

Introduction

As humans, we are competent in perceiving our environment. In a rainstorm, we manage to drive safely on cues as rough as the blur of oncoming headlights. As parents, we notice the subtlest shift of expression on our toddler's face. The signals around us may be extremely weak; yet an innate perceptual inference mechanism compensates for this: we *attune* to a situation, resolve any ambiguities and accurately extract the relevant bits of information.

Inspired by human visual acuity, this thesis aims to recover and interpret subtle signals in degraded images and videos. In particular, it proposes mathematical models and statistical inference algorithms to analyze extremely low-resolution videos of human faces.

1.1 The Human Face and its Perception

We critically rely on seeing each other's faces for communication purposes. We discriminate friend from foe by their facial appearance. We express our emotions and intentions through facial mimicry. In return, we instinctively search for clues on the faces of others. Our ability to accurately interpret the faces around us is an important social skill [Young, 1998; Bruce and Young, 2000].

While the human visual system is impressive in adapting to lighting conditions and resolving fine visual details [Hubel, 1988], it has biologically-imposed limits: our face perception deteriorates with increasing darkness and distance.

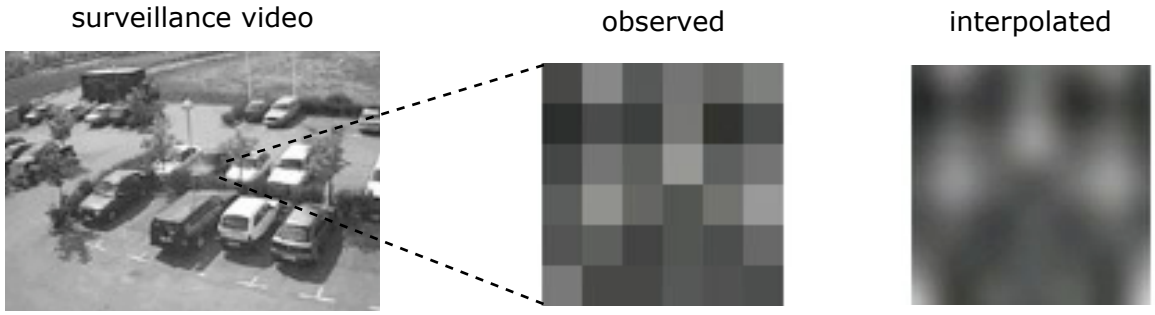

Figure 1.1: In this snapshot of a surveillance video, we may be interested in enhancing the subject's face (cropped, middle) for identification or expression analysis. Conventional methods such as bicubic interpolation yield overly smooth images (right).

1.1.1 The Resolution Problem

When we observe a scene through a camera, our perceptual capability largely depends on the characteristics of the sensing and display devices. For instance, the finite spatial resolution of images can impose limits to human interpretation. Fig. 1.1 shows a snapshot of a surveillance video, with a subject's face cropped and magnified for closer inspection. With as little image information, identification and expression analysis is a challenge to both humans and computers.

1.1.2 Image Degradation $\neq$ Face Degradation

For a closer look at the resolution problem, consider the face images in Fig. 1.2, downsampled progressively from left to right. The highest-resolution image (left) has substantial detail, revealing the smiling face and even the gaze direction of the subject. Moving to the right, the same image is shown under increasing blur and quantization. Facial features such as the eyes, nose, and mouth become more difficult to discern: their intensities blend with each other and with the surrounding skin.

What type of facial information is preserved through resolution degradation? It is instructive to inspect the lowest-resolution image (Fig. 1.2, rightmost) and to try to infer the properties of the face. To start with, human observers can reliably detect faces at this resolution [Bhatia et al., 1995; Torralba and Sinha, 2001]. The head pose is not ambiguous, the mouth looks closed and the subject seems to have his eyes open. As we move to the left and inspect higher-resolution versions of the same face, we gradually recover more details.

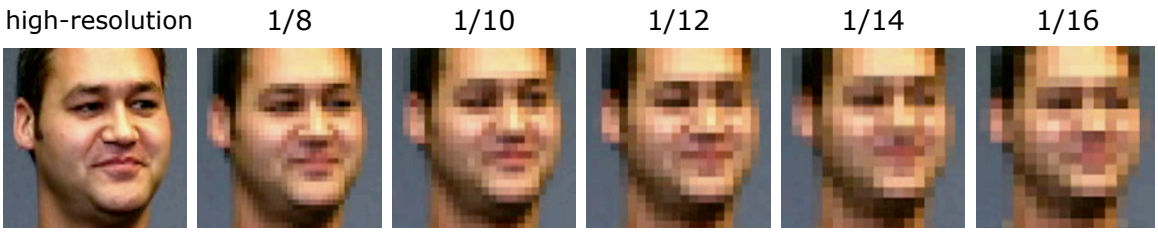

Figure 1.2: From left to right, a high-resolution face image is progressively downsampled. Under increasing blur and quantization, facial features such as eyes, noses, and mouths become more difficult to discern: their intensities blend with each other and with the surrounding skin. What type of facial information is preserved through this degradation?

The above example suggests that *image degradation* and perceived *face degradation* are related yet different phenomena: the former is a deterministic pixel-level intensity degradation, while the latter results from a (subjective) cognitive process. Although individual pixels may be blurred and minute details destroyed, as a whole, a low-resolution face image still conveys useful information to humans. Our resilience to poor resolution was first observed by [Harmon and Julesz, 1973] and has since been studied extensively [Bachmann, 1991; Samal, 1991; Costen et al., 1994; Sinha et al., 2005].

1.1.3 Temporal Signatures

Difficult viewing conditions reveal an interesting aspect of our visual acuity: in surveillance videos, people at a distance may appear very small, fitting perhaps in a 10-by-2 pixel window. When we look at static images of this quality, we rarely interpret such "vertical sticks" as people. Yet if we watch the same-size *video* of a *walking* person, we are immediately able to detect the human, and sometimes even identify the subject based on her silhouette and gait. This capability suggests that spatial and temporal aspects of our visual stream play a joint role in our interpretation of noisy, ambiguous images. Examples of vision algorithms that exploit whole-body motion cues include [Polana, 1994; Cutler and Davis, 2000; Efros et al., 2003] and those reported in [Shah and Jain, 1997].

Though not as periodic as walking silhouettes, faces too exhibit temporal patterns. Facial expressions are not instantaneous random events; they are produced through continuous muscle activation over time [Ekman and Friesen, 1978]. Under non-optimal conditions (*e.g.*, viewing blurred or negative images), motion has been shown to help human face perception [Bruce and Valentine, 1988; Pike et al., 1997; Lander et al., 2001]. The

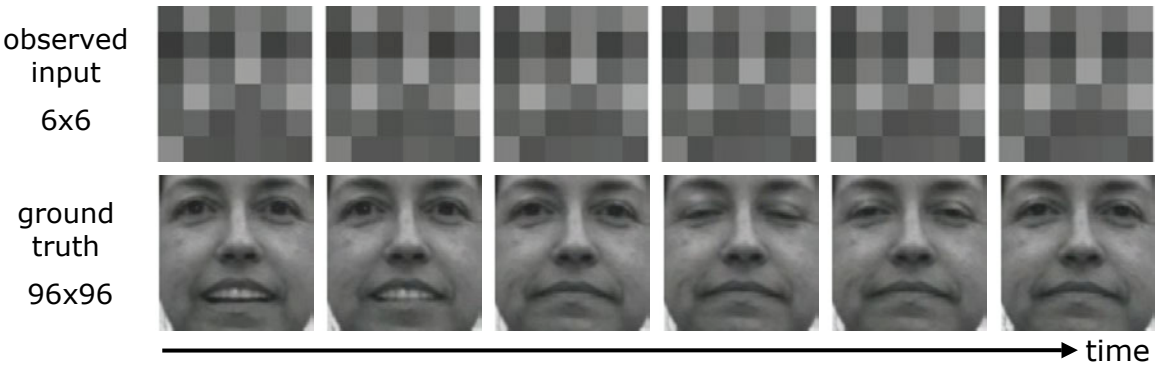

Figure 1.3: Imagine we are given an extremely low resolution video (top). *Assuming* that there is a human face in these images, can we guess the missing details, and estimate (or "hallucinate") a highly zoomed video that resembles the original (bottom)?

"supplemental information hypothesis" [O'Toole et al., 2002] suggests that humans learn motions of familiar faces as dynamic signatures and exploit them for recognition [Roark et al., 2006]. A number of automated facial analysis systems also model emotions and expressions as temporal patterns [Pantic and Rothkrantz, 2000; Tian et al., 2005]. However, their performance drops considerably under low-resolution conditions [Tian, 2004].

1.1.4 A Challenge for Computer Vision

As humans, we are competent in perceiving and interpreting faces in low-resolution images. Could computers do the same, maybe even better? Could we ask them to draw a picture of "what they saw in low-resolution" for us?

1.2 The Face Hallucination Approach

"Face Hallucination" [Baker and Kanade, 2002] aims to recover high-quality, high-resolution images of human faces from low-resolution, blurred and degraded images or video. Fig. 1.3 illustrates the problem: we are given a low-resolution image sequence (top), where an entire face occupies a 6-by-6 pixel patch only. Can we estimate (or "hallucinate") a zoomed, high-resolution video that resembles the original (bottom)?

Image resolution enhancement is an inverse problem [Bertero and Boccacci, 1998]. Under blur and downsampling operations, distinct high-resolution images (Fig. 1.4, middle)

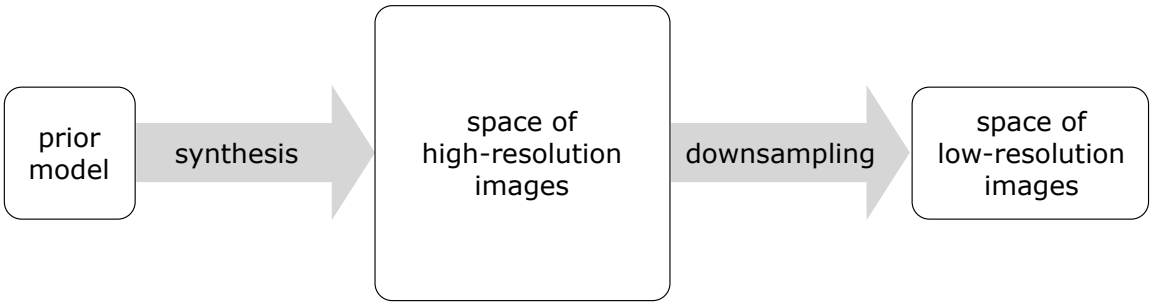

Figure 1.4: Downsampling is a many-to-one mapping from the space of high-resolution images (middle) to that of low-resolution ones (right). This operation cannot be inverted without additional information. Face Hallucination addresses this issue with face priors and models (left) that constrain the set of possible high-resolution images.

produce the same low-resolution image (Fig. 1.4, right). This is a many-to-one mapping that cannot be inverted without additional constraints. Mathematically speaking, the problem is *ill-posed* and does not have a unique solution.

Face Hallucination turns the resolution enhancement problem into a well-posed one by limiting the space of allowed high-resolution images: it *assumes* that a human face is being observed and imposes a face prior (Fig. 1.4, left). Informally speaking, it first estimates the *face content* of the low-resolution observations and then uses this information to reconstruct the high-resolution *image content*. The estimation procedure "skims" low-resolution data for the last bits of facial information that it contains.

1.3 Thesis Statement

This thesis argues for a careful exploitation of space (image) and space-time (video) models when using the Face Hallucination approach to the resolution enhancement problem. In particular, it demonstrates that

1. **Face Hallucination requires carefully crafted metrics and algorithms.** Both the formulation and the numerical optimization of the fitting metric are susceptible to a bias that to date has been ignored.

2. **Face Hallucination can exploit facial dynamics.** Temporal representation and reasoning about facial expressions improve the robustness of recovered video details.

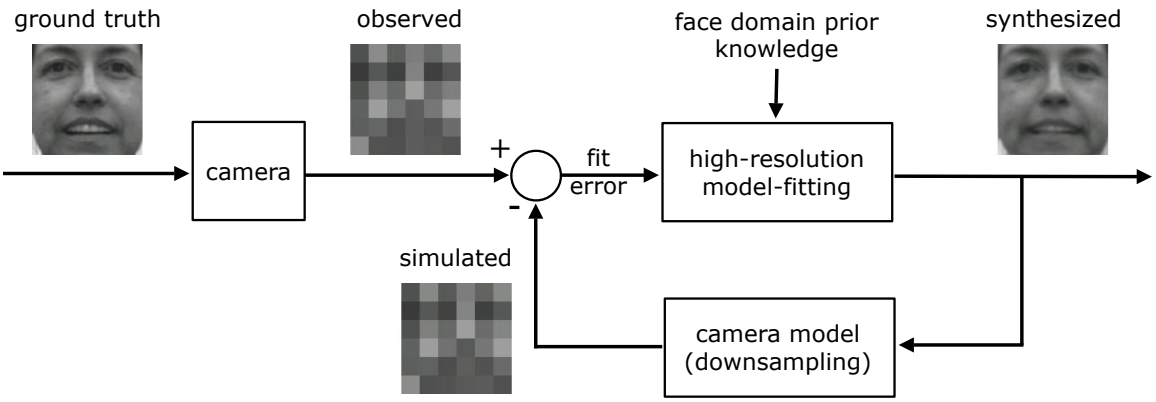

Figure 1.5: Face Hallucination follows the analysis-by-synthesis paradigm: it fits a high-resolution generative face model to low-resolution data. Model-fitting typically involves a search (or optimization) in the state space of the model. Once the best model parameter (or configuration) is found, the model synthesizes high-frequency image details.

1.4 Thesis Overview

While Face Hallucination is a model-based inference problem, it does not prescribe any particular face model or inference algorithm. The representation of the face can be holistic or parts-based, parametric or data-driven, geometric or image-based. If the face is moving, the model can be augmented with geometric transformation parameters to track *and* to hallucinate at the same time. Depending on the application, generic or subject-specific face priors might be more suitable.

The arguments of this thesis span two face models: 1) an Active Appearance Model [Cootes et al., 1998, 2001] that generates face *images* of varying shape and appearance, and 2) a data-driven graphical model [Dedeoğlu et al., 2004] that generates face *videos* of various expressions. Although these models are quite different, Face Hallucination exploits them in the same fashion: following the analysis-by-synthesis paradigm, it fits a (high-resolution) model onto (low-resolution) data. As illustrated in Fig. 1.5, this involves searching for the model parameter (or configuration) that best explains the observations. Once the best model setting is found, the model synthesizes high-frequency image details.

1.4.1 Exploiting an Image Model

Chapter 3 approaches the Face Hallucination problem with an *image* model, and demonstrates that hallucination critically depends on accurate estimates of the underlying facial features. The computational workhorse is the Active Appearance Model [Cootes et al., 1998, 2001], a compact representation of the shape and appearance of objects. This model has been most popular in face modeling, since it can generate faces with varying expression and pose by composing an appearance image and deforming it into different shapes.

The key contribution of Chapter 3 is the discovery of a resolution-induced bias that plagues most model-to-image (or image-to-image) fitting algorithms. This bias affects Active Appearance Models as well. Section 3.2 shows that models and observations should be treated *asymmetrically*, both to formulate an unbiased objective function and to derive an accurate optimization algorithm. Upon this observation, Section 3.3 derives a novel "resolution-aware fitting" algorithm that respects the asymmetry.

The proposed model-fitting (and hallucination) method is experimentally compared against a state-of-the-art fitting algorithm across a variety of resolution and model complexity levels. The results in Section 3.4 show significant improvements in the estimation accuracy of both shape and appearance parameters, yielding more accurate hallucinations. As shown in Fig. 1.6, the novel fitting algorithm is significantly more accurate in estimating and reconstructing faces. This demonstrates the importance of carefully crafted metrics and optimization algorithms in meeting the accuracy challenges in low-resolution.

1.4.2 Exploiting a Video Model

Chapter 4 approaches the Face Hallucination problem with an *video* model, and demonstrates that hallucination can benefit from facial dynamics. To investigate the role of time, a statistical generative model of face videos is proposed. This model treats videos as compositions of space-time patches and encodes visual phenomena in an example-based fashion. The patch-based representation is also used to define a prior in space and time.

To quantify the effect of spatial and temporal models on the hallucination performance, extensive experiments are performed. As illustrated in Fig. 1.7, the proposed algorithm produces high-resolution expressions (bottom) that closely resemble the ground truth (middle). The quantitative results of Section 4.4 demonstrate that temporal representation and reasoning about facial expressions improves robustness by regularizing the Face Hallucination problem.

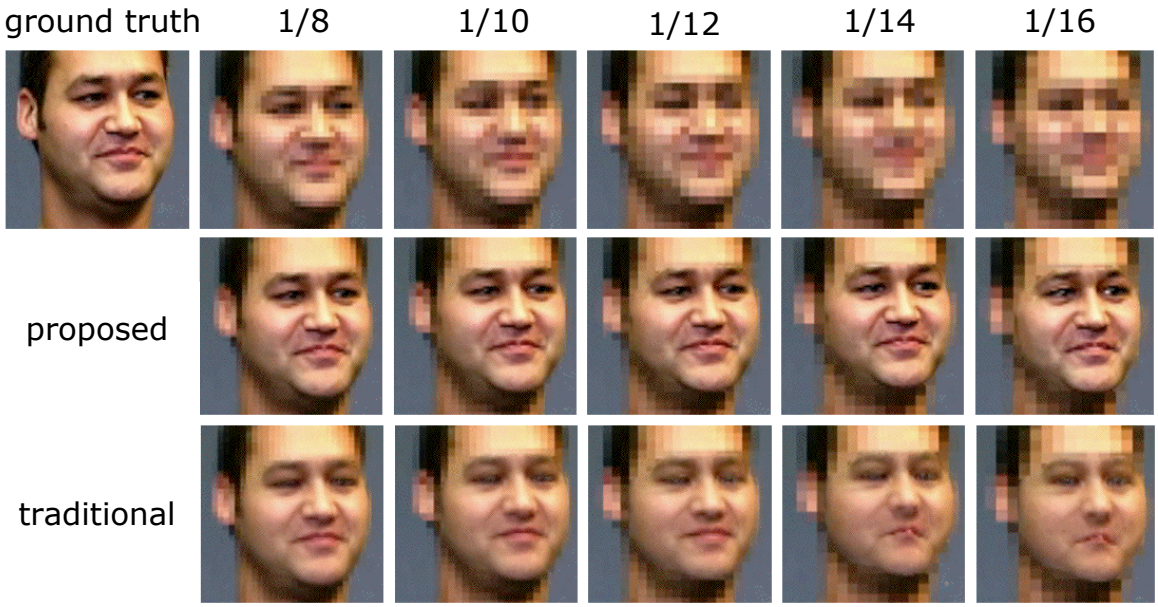

Figure 1.6: The proposed AAM fitting formulation yields significantly more accurate reconstructions of facial details (middle). State-of-the-art algorithms (bottom) that rely on the traditional formulation are shown to exhibit a systematic bias.

Figure 1.7: Exploiting the spatio-temporal dynamics of faces, the video hallucination algorithm (bottom) reconstructs facial expressions that closely resemble the ground truth (middle). Quantitative experiments reveal the smoothing role of temporal dynamics in overcoming 16-fold resolution degradations.

Appearance-based models can be brittle against varying illumination. In Chapter 5, this problem is explicitly addressed by augmenting the video model of Chapter 4 with a low-dimensional illumination subspace. Illumination is treated as a nuisance parameter: its effects are first estimated and then removed from observed videos. This permits face hallucinations beyond the lighting conditions of the training videos.

1.5 Contributions

The contributions of this thesis can be summarized as follows:

- It demonstrates that Face Hallucination critically depends on model-fitting metrics: a resolution-induced bias is shown to affect most model-to-image and image-to-image fitting algorithms operating on low-resolution images. The analysis reveals that models and observations should be treated *asymmetrically* both to formulate an unbiased objective function and to derive an accurate optimization algorithm. The asymmetry leads to a trade-off between computational efficiency and estimation accuracy in low-resolution regimes.

- It proposes a model-fitting algorithm that respects the above-mentioned asymmetry: it adopts the popular Active Appearance Model and derives a novel Face Hallucination and tracking algorithm that proves significantly more accurate than state-of-the-art methods in low-resolution.

- It demonstrates how Face Hallucination can benefit from facial dynamics: a statistical generative model of face videos is proposed to represent and reason about facial expressions. This model treats videos as compositions of space-time patches, efficiently capturing complex visual phenomena such as eye-blinks and the occlusion or appearance of teeth.

- It exploits the space-time representation to define a data-driven face prior on a 3-dimensional Markov Random Field. It poses Face Hallucination as a probabilistic inference problem and demonstrates the crucial role of a video's temporal dimension in hallucinating the correct facial behaviors.

- It proposes an approximate compensation scheme against illumination variation. It augments the generative video model with a low-dimensional illumination subspace, whose parameters are estimated jointly with high-resolution face details. This makes Face Hallucinations beyond the lighting conditions of the training videos possible.
- It achieves person-specific resolution enhancements up to a scaling factor of 16.

Chapter 2

Prior Work on Image Resolution Enhancement

This chapter overviews prior approaches to the general problem of image resolution enhancement. The working principles and assumptions of existing algorithms reveal why enhancing face images can present both a challenge and an opportunity.

2.1 Resolution Enhancement: An Inverse Problem

Cameras cannot capture infinitely detailed images. The optical blur by their lenses and the finite density of their sensing elements limit the resolution of the images they capture. Image quality is further affected by sensor noise and quantization artifacts. In some cases the problem can be alleviated by placing the camera closer to the object of interest or by using a telezoom lens. However, there are a number of scenarios where these approaches would be impractical or prohibitively expensive. These limitations have motivated signal processing and pattern recognition approaches to enhance image resolution, yielding a number of algorithms over the past two decades [Borman and Stevenson, 1998; Chaudhuri, 2001; Park et al., 2003].

In the resolution enhancement or super-resolution (SR) problem, we are given one or more low-resolution images and are asked to produce a higher-resolution one. Let us denote the high-resolution image by H ($M^2N^2 \times 1$ vector), and the low-resolution image by L ($N^2 \times 1$ vector), corresponding to a downscaling factor of M per dimension, where $M > 1$.

From the point of view of image formation, the relationship between H and L is relatively well understood [Andrews and Hunt, 1977]. A local linear averaging and downsampling operator, denoted by A ($N^2 \times M^2N^2$ matrix), maps high-resolution images onto the space of low-resolution ones. A pixel-wise independent Gaussian noise may be added to account for the imaging sensor noise:

$$L = AH + \eta_L. \tag{2.1}$$

Observe that the problem of recovering the high-resolution image H amounts to inverting the operator A. However, because A is a many-to-one mapping, its inversion is mathematically ill-posed [Vogel, 2002], necessitating some form of prior knowledge about the images to constrain the solution for H [Bertero and Boccacci, 1998; Chalmond, 2003]. For example, the smoothness assumption would penalize strong edges, effectively constraining the solution space to that of smooth images (Fig. 1.1, right).

The observation model in (2.1) can generalize to video sequences by simply redefining the variables H and L to be stacked vectors of video frames and by turning A into a block diagonal matrix (ignoring motion-induced blur). As the ill-posed nature of the problem persists, a prior for high-resolution videos would still be needed: a commonly used prior assumption for videos is that of spatio-temporal smoothness [Kokaram, 1998].

A review of the existing literature on this problem identifies two approaches. First, *reconstruction-based* methods aim to increase the effective sampling density. This requires a set of aliased, low-resolution samples of the underlying scene to be fused into a coherent high-resolution estimate. The estimate, in turn, should be able to account for all undersampled low-resolution observations by simulating the image degradation process. The second and more recent approach is *learning-based*, where low-resolution observations are used to predict lost high-resolution details using a training set.

Fig. 2.1 compares the SR methods from an inference point of view. Observe that both attempt to infer a point in the space of high-resolution images (middle) that *explains* the observed low-resolution image (right) through downsampling. In addition, they may both impose generic image priors such as local smoothness, edge preservation, and non-negativity. Learning-based approaches bring in additional, oftentimes domain-specific priors (left) to further constrain the space of possible solutions.

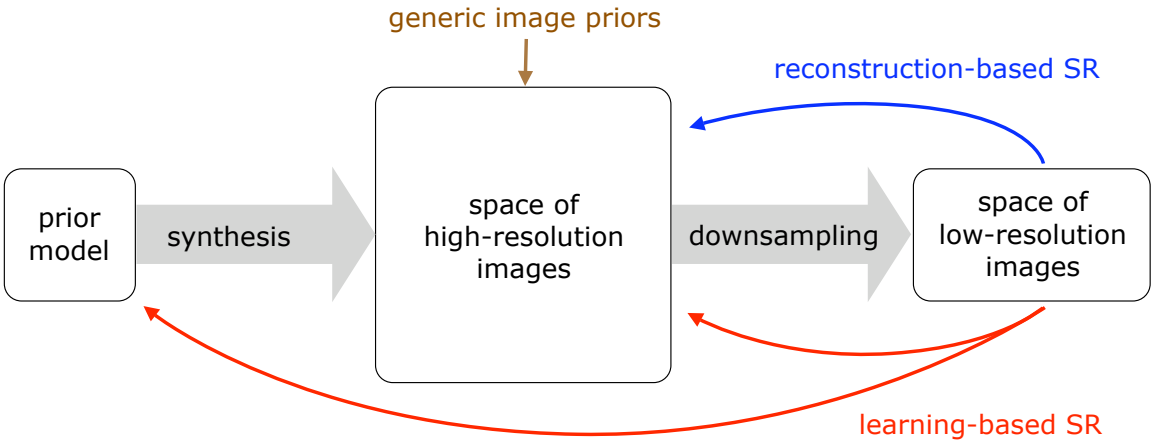

Figure 2.1: Both reconstruction- and learning-based approaches try to infer a high-resolution image that *explains* the low-resolution observation when downsampled. Learning-based methods bring in additional, oftentimes domain-specific priors to further constrain the space of possible solutions.

2.2 Reconstruction-based Approaches

The fundamental problem with low-resolution images and videos is that they are severely undersampled, *i.e.*, aliased. Since the early 80s, much research has been dedicated into alleviating this aspect by means of increasing the effective *sampling density*. The starting point is a set of sub-pixel shifted low-resolution observations of a scene (Fig. 2.2, left). Provided that we can accurately bring these images into a common coordinate frame, it becomes possible to estimate a *super-resolved* (SR) image, defined over a sampling grid finer than in any of the original observations (Fig. 2.2, right). The theoretical foundation of these approaches can be found in the Generalized Sampling Theory [Papoulis, 1977].

The intuition above leads to the *reconstruction constraint* that the high-resolution estimate would need to satisfy: after being warped back to the coordinate frame of each observation, followed by blurring, downsampling and decimation, the estimate should be able to regenerate all low-resolution images up to the noise level. In the case of a video sequence, temporally adjacent frames are candidates for providing largely overlapping observations. Provided that the video is smooth enough to reliably and accurately recover the relative motion between frames, the reconstruction constraint can be imposed.

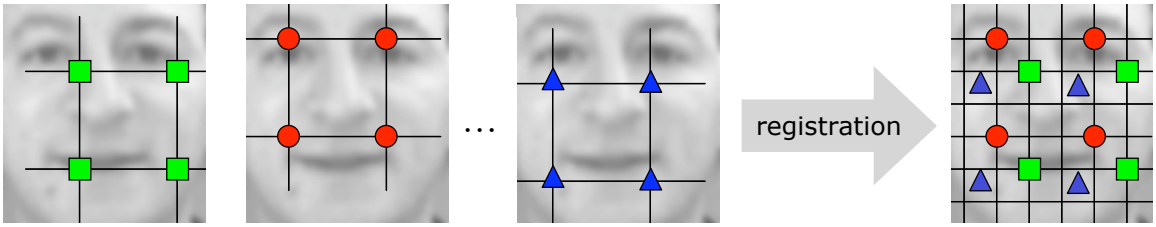

Figure 2.2: Reconstruction-based approaches aim to increase the spatial density of samples. Provided that multiple low-resolution images (left, 2×2 pixels) can be aligned correctly into a common coordinate frame, interpolation over a finer sampling grid can yield a "super-resolved" image (right, 4×4 pixels).

2.2.1 Computational Tools

The reconstruction-based line of research dates back to the frequency-domain technique of [Tsai and Huang, 1984]. Besides restricting the relative motion between low-resolution images to global translations, their work did not model optical blur or observation noise. Since then, a variety of models and computational tools have been developed, gradually expanding allowed motion types and incorporating more realistic observation models. For a comprehensive review, see [Borman and Stevenson, 1998; Park et al., 2003]. The following is a summary of the computational tools used in the reconstruction-based paradigm.

Least-squares (LS) methods have been used in both frequency [Kim et al., 1990] and spatial domains [Elad and Feuer, 1997, 1999]. A recent LS formulation simultaneously enhanced both spatial and temporal resolution of videos via spatio-temporal regularization [Shechtman et al., 2002]. The Iterated Back Projection (IBP) algorithm [Irani and Peleg, 1991] was inspired by Computed Tomography techniques. A similar algorithm was based on non-uniform interpolation and deblurring [Ur and Gross, 1992]. Using Projection-Onto-Convex-Sets (POCS) methods, various image priors could be imposed efficiently [Stark and Oskoui, 1989]. Subsequent work took into account optical blur, nonzero aperture time and sampling on arbitrary lattices [Patti et al., 1997].

Probabilistic inference tools have also been applied to SR problems: spatial-domain Bayesian formulations offered a rigorous estimation framework with regard to spatial priors [Schultz and Stevenson, 1996; Cheeseman et al., 1996; Bascle et al., 1996]. Using probabilistic graphical models such as Markov Random Fields (MRF) [Li, 2001], desired local image properties such as smoothness and edge-preservation could be conveniently expressed. An extension of MRFs into the temporal domain naturally captured spatio-

temporal video priors [Borman and Stevenson, 1999]. Other variations on the Bayesian approach include Gaussian process priors [Tipping and Bishop, 2003] and dynamic tree inference algorithms [Storkey, 2003]. Bayesian inference leads to well-defined Maximum A Posteriori estimates of high-resolution images. In contrast, POCS and IBP methods may not have unique solutions.

2.2.2 The Need for Accurate Motion Estimation

Accurate image alignment plays a critical role in reconstruction-based methods. Prior work explicitly addressed issues related to motion ambiguities and partial occlusions [Irani and Peleg, 1993]. In a similar vein, "validity and segmentation maps" were constructed to eliminate inaccurate motion estimates and to enable object-based tracking [Eren et al., 1997]. Observing the importance and difficulty of estimating the motion accurately, registration and SR problems were cast jointly and solved iteratively [Hardie et al., 1997]. To avoid early commitment to potentially erroneous initial motion estimates, a set of feasible motions were maintained in [Shah and Zakhor, 1999]. Recent work has analyzed the influence of image alignment and warping errors on the quality of super-resolved images [Lin and Shum, 2001, 2004; Zhao and Sawhney, 2002].

Dense optical flow [Horn and Schunck, 1981] has also been used in SR [Baker and Kanade, 1999a; Jiang et al., 2003]. Enforcing the consistency of recovered flow fields has been shown to improve the robustness of SR results [Zhao and Sawhney, 2002]. The challenge in estimating such high-dimensional motion models is that each flow vector has to be estimated from a small number of pixels. This is in contrast to more global, parametric motion models [Bergen et al., 1992] that are reliable and accurate as long as the underlying scene motion is approximated well. For instance, the feature-point approach of [Capel, 2001; Capel and Zisserman, 2003] modeled inter-frame motions as homographies and estimated their parameters using sampling procedures that were robust against outliers.

Reconstruction-based SR requires the registration of independently acquired images into a common coordinate frame. When this crucial step is not well-defined or is simply too difficult due to blur and noise, SR cannot be performed.

2.2.3 What's So Hard About Faces?

The high-resolution video frames in Fig. 1.3 demonstrate how changing facial expressions modify appearance: a surprised look causes the eyebrows to rise and slightly bend. Cheeks, lips and the lower contour of the chin move and deform as one speaks. Dimples and wrinkles appear and disappear as a result of facial muscle activity. Particularly dramatic are eyeblinks that within a fraction of a second occlude and then reveal the eyes. Similarly, parts of a speaker's teeth and tongue become intermittently visible during speech.

Unfortunately, from an image registration point of view, the rich visual phenomena of faces represent the very cases for which the motion estimation problem is not well-defined: the pixel intensities over consecutive frames are not displaced versions of each other. An inspection of the low-resolution version of the same video reveals that the downsampling process has largely destroyed the complex visual phenomena described above.

One may wonder whether nonparametric motion models such as optical flow would be able to recover smooth deformations in certain areas of the face, and thus make SR feasible, at least locally. Unfortunately, since the effects of occluded or newly appeared visual structures get irreversibly mixed in with their neighboring pixels, a good portion of face pixels is unavoidably contaminated with such unmodeled variations. Recovering the *correct* sub-pixel displacements on such a small scale does not seem practical.

The correctness argument about registering video frames may seem too conservative. After all, there is a large body of literature and working systems that estimate complex motion fields for classification and detection purposes [Shah and Jain, 1997; Cédras and Shah, 1995; Efros et al., 2003]. The crucial factor that distinguishes the reconstruction-based SR problem is the *accuracy* requirement on the motion estimation, whereas in other domains, *repeatability* takes priority. Unless the low-resolution images are registered accurately, SR estimates will simply be incorrect. In a motion-based classification task, as long as the same estimation error is consistently reproduced, the performance of classifiers or detectors would suffer minimally.

2.3 Learning-based Approaches

Learning-based approaches rely on the premise that a low-resolution observation contains enough information to make reasonable predictions about its high-resolution counterpart or

features thereof (such as edges). The essence of these techniques is to use a training set of high-resolution images and their low-resolution versions to learn a joint occurrence model. This model can take a variety of forms: a set of learnt interpolation kernels, a look-up table of low-high resolution image patches, or their coefficients in alternative representations.

Learning-based algorithms for SR are relatively recent and have mostly been restricted to static images. It was hypothesized in [Candocia, 1998; Candocia and Principe, 1999] that similar image neighborhoods remained similar across scales, and a local learning from training samples was proposed. A set of interpolation kernels was extracted in an unsupervised fashion, and resolution enhancements by a factor of 2 per dimension were reported. Similar results were achieved through Tree-Based Resolution Synthesis [Atkins, 1998; Atkins et al., 1999], which learned various interpolation filters from training data and applied them selectively upon local classifications. The wavelet-domain formulation of [Daniell and Matic, 1999] exploited SR for compression: their multiresolution image coder used a neural network to predict upper frequency band coefficients from low frequency information. The work by [Freeman and Pasztor, 1999; Freeman et al., 2000, 2002] represented the scene as a Markov network of image patches and used sample-based probabilistic inference algorithms, yielding an enhancement factor of 4. Primal sketches [Marr, 1982] were used for both recognition and enhancement purposes in [Sun et al., 2003], yielding a 3-fold increase in resolution.

When images are limited to a particular domain, learning-based approaches can be very powerful: the seminal work on Face Hallucination [Baker and Kanade, 1999b, 2002] considered super-resolving human faces only, and furthermore, employed inhomogeneous (*i.e.*, location-specific) priors. Their recognition algorithm referred to a database of registered face images and selected those training patches that best matched a given input, producing convincing results with zoom factors of up to 8. The two-step procedure of [Liu et al., 2001, 2007] first estimated a global face via Principal Component Analysis (PCA) and then fit a nonparametric local model. The "eigenface" representation was also exploited in [Capel and Zisserman, 2001; Gunturk et al., 2003; Wang and Tang, 2005]. The work of [Jia and Gong, 2005] extended this approach to the multilinear case and took into account viewpoint and illumination variations. In a similar vein, morphable models [Vetter and Troje, 1997] that encode the shape and texture of faces jointly were also used for SR enhancement [Park and Lee, 2003, 2004].

Learning-based SR has been applied to videos as well: the method of [Freeman et al.,

2000] was applied to image sequences, but severe video artifacts were observed [Bishop et al., 2003]. To achieve more coherent videos, the heuristic of *re-using* the high-resolution solutions of preceding frames was proposed. The temporal redundancy of videos was also targeted in [Jia and Gong, 2006], who fused multiple, partially occluded face observations in a video sequence.

Frame-to-frame smoothness and redundancy is not the only characteristic of videos. For instance, the video of a human face that speaks and displays natural expressions over time will contain very specific temporal regularities. Our previous work on Face Hallucination in videos [Dedeoğlu et al., 2004] exploited these *spatio-temporal signatures* for recognition[1]: the complex occlusion and re-appearance events that haunt alignment algorithms were used as rich temporal signatures that distinguished facial expressions from each other.

In principle, the performance of learning-based techniques is limited by the amount of discriminative information that "sneaks" from high-resolution training samples to their low-resolution counterparts during the downsampling process. The challenge for these algorithms is to retain as much of this information as possible while generalizing to other samples drawn from the domain of interest.

2.4 Summary

This chapter surveyed existing methods for resolution enhancement and discussed their suitability for human faces. First, it described the working principles of reconstruction-based methods and underlined how critically they depended on accurate registration of low-resolution images: this turned out to be a serious hurdle for low-resolution faces, where the eye and mouth pixels can rarely be explained with motion models. Second, it presented the philosophy of learning-based methods, such as Face Hallucination.

[1] This was an early version of the model presented in Chapter 4.

Chapter 3

Face Hallucination with an Image Model

Face Hallucination exploits high-resolution face priors to interpret and to enhance low-resolution images. Its success hinges upon *accurately* estimating facial features, independent of how these may be represented or parametrized in the prior model. When the estimation accuracy drops, hallucinated details become incorrect and unrealistic. Fig. 3.1 illustrates this point on three hallucination examples: when there is a large error, hallucinated faces may not even look human (middle). Moreover, a small advantage in estimation accuracy can yield a drastic improvement in hallucination (right). This chapter demonstrates the importance of carefully crafted metrics and optimization algorithms in meeting the accuracy challenges in low-resolution.

The computational workhorse of this chapter is the Active Appearance Model [Cootes et al., 1998, 2001]. It is a compact representation of the shape and appearance of objects

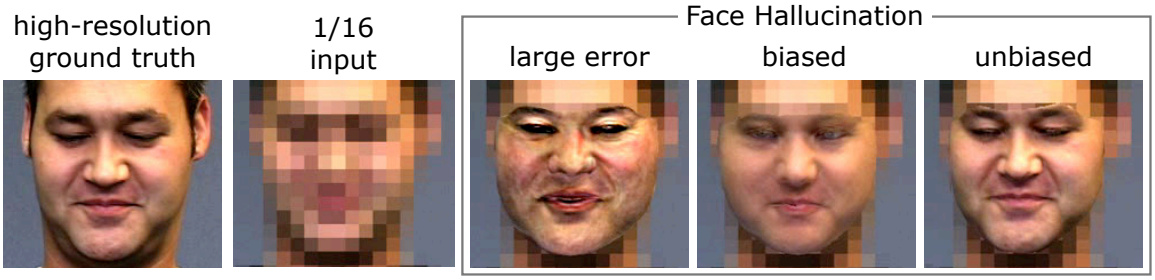

Figure 3.1: Face Hallucination critically depends on accurate estimates of the underlying facial features (left-most). When the estimation accuracy drops, hallucinated details become incorrect and even unrealistic (middle). We show that a small improvement in estimation accuracy can dramatically boost the hallucination quality (right-most).

and has been most popular in face modeling: it can generate faces with varying expression and pose by modifying its appearance image and warping it into different shapes. Section 3.1 defines this model mathematically and describes how it has traditionally been fit to observed data.

Section 3.2 reveals a resolution-induced bias that plagues most model-to-image (or image-to-image) fitting algorithms, including Active Appearance Models. This implies that models and observations should be treated *asymmetrically*, both to formulate an unbiased objective function and to derive an accurate optimization algorithm. Upon this observation, Section 3.3 formulates a novel *resolution-aware fitting* algorithm that respects the asymmetry and incorporates an explicit model of the blur caused by the camera's sensing elements.

Section 3.4 experimentally compares the new fitting and hallucination algorithm against a state-of-the-art algorithm across a variety of resolution and model complexity levels. The results show significant improvements in the estimation accuracy of both shape and appearance parameters, yielding more accurate hallucinations using RAF.

Section 3.5 discusses some practical and algorithmic consequences of the asymmetry principle and identifies directions for future investigation.

3.1 The Active Appearance Model

Active Appearance Models (AAM) are compact, parametric representations of the shape and appearance of objects [Cootes et al., 1998, 2001]. Because they can efficiently encode non-rigid shape deformations and appearance variations, they have been very popular in face tracking applications.

Recall that Face Hallucination exploits a high-resolution face prior model to interpret low-resolution data. The AAM must be trained on face images of high-resolution where landmarks such as eyebrows and lips need to be manually labeled. Once the AAM has been learned, it can be fit to novel (low-resolution) images for interpretation and (high-resolution) synthesis.

An AAM consists of two models, the *shape* and *appearance* of an object. Each of these is a linear Principal Components model learned from training data. The shape of an AAM

is defined by a set of 2D landmark locations

$$\mathbf{s} \;=\; (x_1, y_1, x_2, y_2, \ldots, x_v, y_v)^{\mathrm{T}}. \tag{3.1}$$

The shape model, parametrized with $\mathbf{p} = (p_1, p_2, \ldots, p_n)$, expresses any shape as a linear combination of basis shapes added onto a base shape:

$$\mathbf{s}(\mathbf{p}) \;=\; \mathbf{s}_0 + \sum_{i=1}^{n} p_i \mathbf{s}_i. \tag{3.2}$$

An AAM is defined in the coordinate system of the object being modeled. To generate object instances in arbitrary poses, a global transform is needed. Following [Matthews and Baker, 2004], we define four special shape bases to account for similarity transforms (scale, rotation, and two translations) and compose them with the shape model. We denote the combined geometric deformation by $\mathbf{W}(\mathbf{x}; \mathbf{p})$, where $\mathbf{x}$ is a model point coordinate being mapped onto an image coordinate.

The appearance model consists of the mean and basis images. The basis images are shape-normalized, *i.e.*, they are defined within the base shape $\mathbf{s}_0$. The appearance model is linear, parametrized with $\boldsymbol{\lambda} = (\lambda_1, \lambda_2, \ldots, \lambda_m)$ as

$$H(\mathbf{x}; \boldsymbol{\lambda}) \;=\; H_0(\mathbf{x}) + \sum_{i=1}^{m} \lambda_i H_i(\mathbf{x}) \qquad \forall\, \mathbf{x} \in \mathbf{dom}\, \mathbf{s}_0, \tag{3.3}$$

where $\mathbf{x}$ is a pixel coordinate in the domain of $\mathbf{s}_0$. For a face, appearance basis images of `100x100` pixels are deemed to carry sufficient facial detail.

Given the parameters $\mathbf{p}$ and $\boldsymbol{\lambda}$, an object instance is generated in two steps: First, the image $H(\mathbf{x}; \boldsymbol{\lambda})$ is computed according to appearance coefficients. Next, this image is geometrically warped using the shape parameters $\mathbf{p}$. We use the notation $H\big(\mathbf{W}(\mathbf{x}; \mathbf{p}); \boldsymbol{\lambda}\big)$ to express the resulting warped image.

In this thesis, we consider the simpler case of *independent* AAMs [Matthews and Baker, 2004], where the statistical dependence between the shape and appearance parameters is ignored. While such dependencies have been exploited in prior work [Cootes et al., 1998, 2001], their advantages remain largely separate of the current discussion.

3.1.1 Traditional Fitting Formulation

Given a set of AAM parameters, the linear generative equations (3.2) and (3.3) can uniquely synthesize an object instance [Matthews and Baker, 2004]. Model-based image analysis deals with the inverse of this process: it aims to recover those AAM parameters which *best* explain a given image. For this end, one needs to define a similarity metric to quantify what constitutes a "good" match, and a *fitting* algorithm to compute the parameter values which optimize the similarity metric.

The original AAM work [Edwards et al., 1998; Cootes et al., 1998, 2001] as well as its computationally efficient reformulation [Matthews and Baker, 2004] define the fitting criterion as the sum of squared intensity differences between the synthesized model template and the *warped input image* L:

$$\sum_{\mathbf{x}\in\mathbf{dom}\,\mathbf{s}_0} \Big[L\big(\mathbf{W}(\mathbf{x};\mathbf{p})\big) - H(\mathbf{x};\boldsymbol{\lambda}) \Big]^2. \tag{3.4}$$

Since this objective function is highly nonlinear in its parameters, iterative gradient-descent methods are typically used: in each iteration, updates $\Delta\mathbf{p}$ and $\Delta\boldsymbol{\lambda}$ are computed and added to (or composed with) current estimates of $\mathbf{p}$ and $\boldsymbol{\lambda}$, respectively. Early work in AAMs [Cootes et al., 1998; Edwards et al., 1998; Cootes et al., 2001] assumed a constant relationship between the error image and the additive updates: this mapping was learned through regression on perturbation-based training data. Later, the work by [Matthews and Baker, 2004] showed that in general there is no constant linear relationship between the error image and the update in the additive case, but that there is in the (inverse) compositional case. Based on this insight, along with the independence of the shape and appearance models in an independent AAM, efficient AAM fitting algorithms were developed, running at over 200 frames per second (standard PC, circa 2003) on typically sized AAM.

3.1.2 The Unsuspected Culprit in Low-Resolution Problems

Any search method for optimizing the criterion (3.4) would suffer from a large number of local minima. In some cases, the solution might even be ambiguous. To make matters worse, these difficulties are only exacerbated when the available data is noisy and low-resolution, as in Face Hallucination.

As illustrated in Fig. 3.2, we have to deal with two image coordinates. The first one,

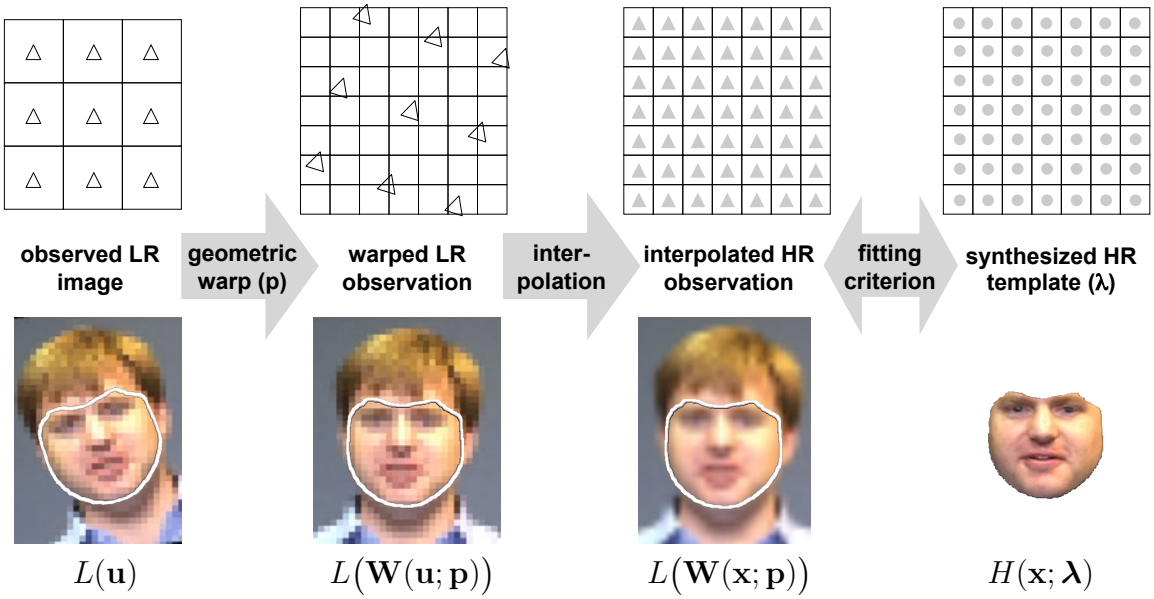

Figure 3.2: Graphical representation of the traditional fitting criterion of (3.4). From left to right, observed images are warped, interpolated, and finally compared against the synthesized model instance. When the input image is low in resolution, significant interpolation is needed to warp it onto the model coordinate frame.

denoted by $\mathbf{u}$, is the pixel coordinate of the low-resolution image L. The second one, denoted by $\mathbf{x}$, is the pixel coordinate in the shape-normalized template image $H(\boldsymbol{\lambda})$. Note that the summation in (3.4) is defined over $\mathbf{x}$, *i.e.*, the fitting criterion prescribes *first warping and interpolating* the image L, and *then* comparing it against the synthesized template. The latter is normalized to shape $\mathbf{s}_0$ at the AAM's native resolution, and remains fixed in size. Consequently, when objects appear small in comparison to the AAM, they need to be enlarged through interpolation.

The traditional formulation's reliance on interpolation turns out to be its *Achilles' heel* in low-resolution regimes. The next section will show that the fitting *criterion* itself becomes increasingly suboptimal (in accuracy) with higher scaling factors. This is an artifact of formulation and represents a serious deficiency for Face Hallucination: if the face content of the low-resolution images is not recovered accurately, reconstructed image details will be erroneous.

3.2 The Asymmetry of Model and Image Fitting Problems

This section reveals a fundamental flaw in traditional AAM fitting; the criterion (3.4) yields a biased estimator when the observations are lower in resolution than the model. Without loss of generality, we analyze this bias in the framework of image-to-image registration. The images can be real (captured by a camera) or synthesized by a generative model such as AAM.

The image registration problem underlies many computer vision applications, such as motion estimation, tracking, model-based recognition and change detection [Brown, 1992; Maintz and Viergever, 1998; Zitova and Flusser, 2003]. It is usually tackled by first defining a geometric deformation scheme and then warping one image onto another such that they become as *similar* as possible according to some criterion.

We address the following questions: When registering two images, can we treat them equally and interchangeably? What are the conditions under which a symmetric treatment is possible? Do these conditions impose any restrictions upon the applicable algorithms? Such questions are relevant to both the *formulation* step and the *numerical optimization* step of the registration task.

The Problem Formulation Step

Consider for example the popular "sum of normed differences" objective function [Irani and Anandan, 2000; Modersitzki, 2004]

$$\sum_{\mathbf{y}\in\mathbf{dom}I_1}\Big[I_1(\mathbf{y})-I_2\big(\mathbf{W_{12}}(\mathbf{y})\big)\Big]^p, \tag{3.5}$$

where I_1 and I_2 are images, $\mathbf{y}$ is a pixel coordinate in the domain of I_1, and $\mathbf{W_{12}}$ is the geometric mapping from the coordinate frame of I_1 to that of I_2. For $p=2$, this amounts to modeling the pixel intensities of I_1 as i.i.d. Gaussian noise added versions of those of the warped I_2. Therefore, the warp that minimizes (3.5) is the Maximum-Likelihood (ML) estimate, known to be asymptotically unbiased [Casella and Berger, 1990].

The formulation above is asymmetric: I_2 is regarded as *template* and is warped onto I_1. Indeed, a survey of existing methods reveals that most image registration problems are formulated in this way, and that there is rarely any discussion as to which image ought to be the template.

The asymmetry of (3.5) has been addressed in prior work [Christensen, 1999; Cachier and Rey, 2000; Rogelj and Kovacic, 2003; Skrinjar and Tagare, 2004], where, in an attempt to remove it, the objective functions were symmetrized [1], yielding

$$\sum_{\mathbf{y}\in\mathbf{dom}I_1} \underbrace{\Big[I_1(\mathbf{y}) - I_2\big(\mathbf{W_{12}}(\mathbf{y})\big)\Big]^p}_{\text{from } I_2 \text{ onto } I_1} + \sum_{\mathbf{z}\in\mathbf{dom}I_2} \underbrace{\Big[I_2(\mathbf{z}) - I_1\big(\mathbf{W_{21}}(\mathbf{z})\big)\Big]^p}_{\text{from } I_1 \text{ onto } I_2}. \tag{3.6}$$

A symmetric form for the geometric warp priors have also been proposed [Ashburner et al., 1999]. In some cases, to further impose symmetry, an additional *consistency* term on $\mathbf{W_{12}}$ and $\mathbf{W_{21}}$ has been used, such as

$$\sum_{\mathbf{y}\in\mathbf{dom}I_1} \Big[\mathbf{y} - \mathbf{W_{21}}(\mathbf{W_{12}}(\mathbf{y}))\Big]^p + \sum_{\mathbf{z}\in\mathbf{dom}I_2} \Big[\mathbf{z} - \mathbf{W_{12}}(\mathbf{W_{21}}(\mathbf{z}))\Big]^p.$$

These past approaches have essentially regarded the asymmetry as an opportunity to incorporate more data and regularization priors into the registration problem at hand.

The Numerical Optimization Step

Independent of the *definition* of an objective function, its *numerical optimization* (*i.e.*, the fitting algorithm) has also been treating the two images in an asymmetric fashion. For example, the original Lucas-Kanade algorithm [Lucas and Kanade, 1981] used a Taylor expansion of the warp around its current estimate, yielding

$$\sum_{\mathbf{y}\in\mathbf{dom}I_1} \Big[I_1(\mathbf{y}) - I_2\Big(\big(\mathbf{W_{12}} + \Delta\mathbf{W_{12}}\big)(\mathbf{y})\Big)\Big]^p,$$

and iteratively solved for the warp updates $\Delta\mathbf{W_{12}}$. Observe that only image I_2 is warped in this scheme. In contrast, the "inverse compositional" algorithm [Baker and Matthews, 2001, 2004] performed the expansion on I_1 and minimized

$$\sum_{\mathbf{y}\in\mathbf{dom}I_1} \Big[I_1\big(\Delta\mathbf{W_{21}}(\mathbf{y})\big) - I_2\big(\mathbf{W_{12}}(\mathbf{y})\big)\Big]^p$$

[1] Re-expressing (3.5) in the domain of I_2 would introduce the Jacobian $|J(\mathbf{W_{12}})|$ as a weighting term. However, the symmetrized form is not necessarily limited to the original noise model. It may instead combine two noise models.

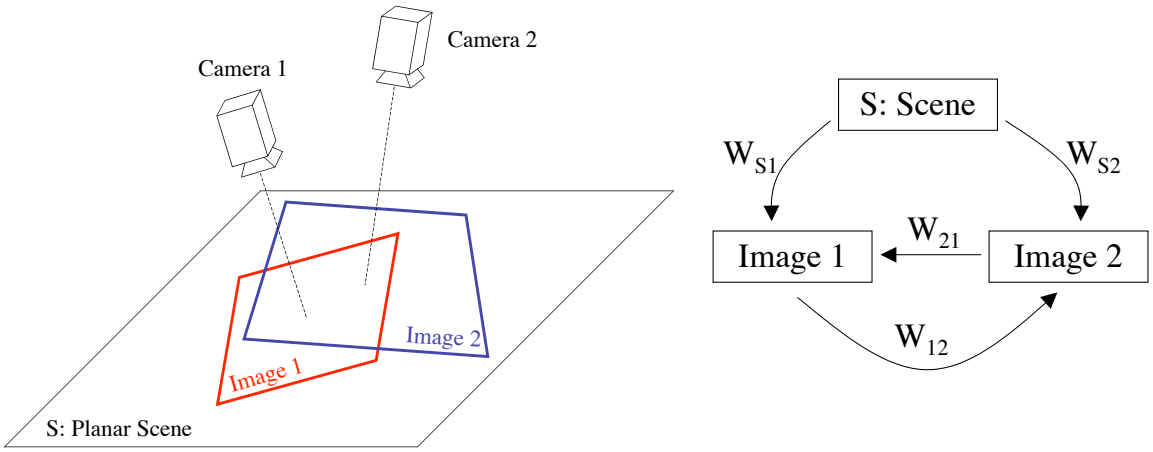

Figure 3.3: A planar scene is observed by pinhole cameras (left). Under central projection, scene-to-image and image-to-image transformations (right) are homographies.

with respect to $\Delta\mathbf{W_{21}}$, resulting in higher efficiency. Note how the inverse compositional formulation warps both images simultaneously, albeit to different degrees.

3.2.1 Analysis in a Simplified Scenario

Consider the simplified scenario shown in Fig. 3.3 (left), in which a planar scene S is observed by two pinhole cameras which capture continuous images. Under the central projection model, scene-to-image and image-to-image coordinate transformations will be homographies [Hartley and Zisserman, 2000]. Note that this class of geometric transformation will account for observed images *exactly*. In order to avoid complications arising from non-corresponding image points, we assume that both images have infinite extent and are free of occlusion.

As shown in Fig. 3.3 (right), the domains of the scene radiance S, image I_1 and image I_2 are related by homographies. $\mathbf{W_{S1}}$ and $\mathbf{W_{S2}}$ denote transformations which take scene coordinates, and compute their corresponding image point locations in I_1 and I_2, respectively. $\mathbf{W_{12}}$ denotes the transformation from I_1 to I_2, and $\mathbf{W_{21}}$ from I_2 to I_1. To render the problem well-posed, all transforms are assumed to be invertible, *i.e.*, $\mathbf{W_{12}} = \mathbf{W_{21}}^{-1}$. Thus, the image registration task is to estimate the homography $\mathbf{W_{12}}$ (or $\mathbf{W_{21}}$) between I_1's and I_2's coordinate frames based on image intensity measurements.

We will use two equivalent notations to express the fact that one image is a geometri-

cally transformed version of another. The first one is $I_1(\mathbf{y}) = I_2\big(\mathbf{W_{12}}(\mathbf{y})\big)$. Using point coordinates, this notation indicates where a particular image point maps onto the other image, and states how those image intensities relate to each other. Alternatively, we will use $I_1 = warp(I_2; \mathbf{W_{21}})$. This notation refers to an entire domain's transformation. It states that I_1 is the image obtained by transforming every point in the domain of I_2 by $\mathbf{W_{21}}$; note the use of $\mathbf{W_{21}}$ here instead of $\mathbf{W_{12}}$, since the transformed points are in I_2.

3.2.2 Theoretical Case: Ideal Camera and Known Scene

We start our discussion with an idealized case. Suppose that we have full knowledge of the underlying scene radiance function S, and both cameras are ideal; their lenses precisely focus incoming light rays parallel to the optical axis onto the camera's image plane, and their photo-receptive fields are continuous (*i.e.*, they have infinite resolution). We model the intensity at an image point as a noisy (independent and identically-distributed, additive Gaussian) observation of the corresponding scene point's radiance,

$$I_1(\mathbf{y}) = S\big(\mathbf{W_{1S}}(\mathbf{y})\big) + \epsilon(\mathbf{y}) \qquad \forall \mathbf{y} \in \mathbf{dom} I_1, \tag{3.7}$$

$$I_2(\mathbf{z}) = S\big(\mathbf{W_{2S}}(\mathbf{z})\big) + \epsilon(\mathbf{z}) \qquad \forall \mathbf{z} \in \mathbf{dom} I_2, \tag{3.8}$$

where $\mathbf{y}$ and $\mathbf{z}$ are points in the domains of I_1 and I_2, respectively. We denote image-to-scene warps (homography) by $\mathbf{W_{1S}}$ and $\mathbf{W_{2S}}$ (Fig. 3.3, right). Using the alternative notation, (3.7) and (3.8) can be also expressed as

$$I_1(\mathbf{y}) = warp(S; \mathbf{W_{S1}})(\mathbf{y}) + \epsilon(\mathbf{y}) \qquad \forall \mathbf{y} \in \mathbf{dom} I_1, \tag{3.9}$$

$$I_2(\mathbf{z}) = warp(S; \mathbf{W_{S2}})(\mathbf{z}) + \epsilon(\mathbf{z}) \qquad \forall \mathbf{z} \in \mathbf{dom} I_2. \tag{3.10}$$

In the following, we present three different methods to estimate $\mathbf{W_{12}}$. Given our assumptions at this moment, these algorithms are rather trivial. Nevertheless, they will be minimally affected while the assumptions are relaxed in Section 3.2.3, allowing us to highlight their applicability to different situations.

A1. Generative Algorithm

Step 1: Find the ML parameters for scene-to-image warps $\mathbf{W_{S1}}$ and $\mathbf{W_{S2}}$:

$$\hat{\mathbf{W}}_{\mathbf{S1}} = \arg\min_{\mathbf{W_{S1}}} \int_{\mathbf{y}\in\mathbf{dom}I_1} \Big[I_1(\mathbf{y}) - warp(S;\mathbf{W_{S1}})(\mathbf{y})\Big]^2 \mathbf{dy}. \tag{3.11}$$

$$\hat{\mathbf{W}}_{\mathbf{S2}} = \arg\min_{\mathbf{W_{S2}}} \int_{\mathbf{z}\in\mathbf{dom}I_2} \Big[I_2(\mathbf{z}) - warp(S;\mathbf{W_{S2}})(\mathbf{z})\Big]^2 \mathbf{dz}. \tag{3.12}$$

Step 2: By the *invariance property* of the ML estimates [Casella and Berger, 1990], the relative warp computed by composition is the ML estimate for $\mathbf{W_{12}}$:

$$\hat{\mathbf{W}}_{\mathbf{12}} = \hat{\mathbf{W}}_{\mathbf{1S}} \circ \hat{\mathbf{W}}_{\mathbf{S2}} = (\hat{\mathbf{W}}_{\mathbf{S1}})^{-1} \circ \hat{\mathbf{W}}_{\mathbf{S2}}.$$

B1. Forward Algorithm

Step 1: Find $\hat{\mathbf{W}}_{\mathbf{S1}}$ and $\hat{\mathbf{W}}_{\mathbf{S2}}$ by (3.11) and (3.12).

Step 2: Based on the scene function S and ML estimates $\hat{\mathbf{W}}_{\mathbf{S1}}$ and $\hat{\mathbf{W}}_{\mathbf{S2}}$, set up a new Least-Squares estimation problem for the relative warp $\mathbf{W_{12}}$:

$$\hat{\mathbf{W}}_{\mathbf{12}} = \arg\min_{\mathbf{W_{12}}} \int_{\mathbf{z}\in\mathbf{dom}I_2} \Big[\underbrace{warp(S;\hat{\mathbf{W}}_{\mathbf{S2}})}_{\hat{I}_2}(\mathbf{z}) - warp\big(\underbrace{warp(S;\hat{\mathbf{W}}_{\mathbf{S1}})}_{\hat{I}_1};\mathbf{W_{12}}\big)(\mathbf{z})\Big]^2 \mathbf{dz}. \tag{3.13}$$

By computing $\hat{I}_1 = warp(S;\hat{\mathbf{W}}_{\mathbf{S1}})$ and $\hat{I}_2 = warp(S;\hat{\mathbf{W}}_{\mathbf{S2}})$, this method essentially *simulates* the formation of ML images of I_1 and I_2. In other words, the registration problem is posed in terms of ML images:

$$\hat{\mathbf{W}}_{\mathbf{12}} = \arg\min_{\mathbf{W_{12}}} \int_{\mathbf{z}\in\mathbf{dom}I_2} \Big[\hat{I}_2(\mathbf{z}) - warp(\hat{I}_1;\mathbf{W_{12}})(\mathbf{z})\Big]^2 \mathbf{dz}. \tag{3.14}$$

Note the similarity between (3.14) and (3.5): they are both asymmetric and warp only one of the images. Indeed, one can use images I_1 and I_2 as plug-in estimates of $\hat{I}_1$ and $\hat{I}_2$, and directly estimate $\mathbf{W_{12}}$. This seems to be exactly the idea behind commonly used objective functions such as (3.5).

C1. Backward Algorithm

Step 1: Find $\hat{\mathbf{W}}_{\mathbf{S1}}$ and $\hat{\mathbf{W}}_{\mathbf{S2}}$ by (3.11) and (3.12).

Step 2: Just as in the *forward* algorithm B1, set up a Least-Squares warp estimation problem. This time, however, solve for the warp in the opposite direction, by warping the other ML image:

$$\hat{\mathbf{W}}_{\mathbf{21}} = \arg\min_{\mathbf{W_{21}}} \int_{\mathbf{y}\in\mathbf{dom}I_1} \Big[\underbrace{warp(S;\hat{\mathbf{W}}_{\mathbf{S1}})}_{\hat{I}_1}(\mathbf{y}) - warp\big(\underbrace{warp(S;\hat{\mathbf{W}}_{\mathbf{S2}})}_{\hat{I}_2};\mathbf{W_{21}}\big)(\mathbf{y})\Big]^2 \mathbf{dy}. \tag{3.15}$$

We have intentionally defined both algorithms to be asymmetric: the *forward* algorithm B1 warps $\hat{I}_1$ onto $\hat{I}_2$, and the *backward* algorithm C1 does the opposite. Using this setup, we can investigate whether there is a fundamental difference between the two. In Appendix A, we show that their optimization criteria differ from each other due to a spatially-varying weighting term. This discrepancy stems from the choice of the noise model in formulating the Least-Squares problem in Step 2.

Choosing an Algorithm

We proposed three algorithms to estimate the geometric warp between two images. The *generative* one requires the knowledge of the scene, but the asymmetric *forward* and *backward* methods do not, because their Step 1 can be skipped and I_1 and I_2 used as proxies for $\hat{I}_1$ and $\hat{I}_2$ instead.

In the next section, we weaken our assumptions and revisit the algorithms above. We show that while the generative formulation can still give us a ML estimate, the asymmetric algorithms need to modify how they use the observed images.

3.2.3 Practical Case: Real Camera and Unknown Scene

A real camera has blur effects. The response of a camera to an ideal point light source is characterized by its Point Spread Function (PSF). This means that the warped scene irradiance will be subject to a convolution with the PSF. For convenience, we still assume the images to be continuous. Instead of (3.9) and (3.10), we have

$$I_1(\mathbf{y}) = B\big(warp(S;\mathbf{W_{S1}})\big)(\mathbf{y}) + \epsilon_1(\mathbf{y}) \qquad \forall \mathbf{y} \in \mathbf{dom} I_1, \tag{3.16}$$

$$I_2(\mathbf{z}) = B\big(warp(S;\mathbf{W_{S2}})\big)(\mathbf{z}) + \epsilon_2(\mathbf{z}) \qquad \forall \mathbf{z} \in \mathbf{dom} I_2, \tag{3.17}$$

where the blur operator $B(\cdot)$ indicates a convolution with the PSF:

$$B\big(S\big)(\mathbf{x}) = \int\limits_{\mathbf{w}\in\mathbf{dom}S} S(\mathbf{w}) PSF(\mathbf{w}-\mathbf{x}) \mathbf{dw}.$$

Due to imperfect lenses and density constraints on photo-receptive sensing elements, the PSF of a real camera is not a delta function [Barbe, 1980]. In fact, the PSF is closely related to measurement noise characteristics. In order to operate at prescribed frame rates and signal-to-noise ratio levels, CCD cameras accumulate photon counts over a finite spatial extent, a procedure called *binning*. The blur model must not only account for realistic lens optics, but also capture the binning operations which take place at the sensing elements.

Also in real situations, we do not know the scene radiance S. Assuming a blurry camera and unknown scene, let us discuss the three algorithms corresponding to those considered in Section 3.2.2 for the ideal case.

A2. Generative Algorithm

Step 1:
$$\hat{\mathbf{W}}_{\mathbf{S1}} = \arg\min_{\mathbf{W_{1S}}} \int\limits_{\mathbf{y}\in\mathbf{dom}I_1} \Big[I_1(\mathbf{y}) - B\big(warp(S;\mathbf{W_{S1}})\big)(\mathbf{y})\Big]^2 \mathbf{dy}, \tag{3.18}$$

$$\hat{\mathbf{W}}_{\mathbf{S2}} = \arg\min_{\mathbf{W_{2S}}} \int\limits_{\mathbf{z}\in\mathbf{dom}I_2} \Big[I_2(\mathbf{z}) - B\big(warp(S;\mathbf{W_{S2}})\big)(\mathbf{z})\Big]^2 \mathbf{dz}. \tag{3.19}$$

Step 2:
$$\hat{\mathbf{W}}_{\mathbf{12}} = \hat{\mathbf{W}}_{\mathbf{1S}} \circ \hat{\mathbf{W}}_{\mathbf{S2}} = (\hat{\mathbf{W}}_{\mathbf{S1}})^{-1} \circ \hat{\mathbf{W}}_{\mathbf{S2}}$$

Since we do not know the scene radiance S, we need to estimate it jointly with the warps. This approach was proposed in the past as part of a super-resolution problem in [Hardie et al., 1997]. Although theoretically sound and elegant, the *generative* algorithm is rarely used in registering images. Instead, *forward* or *backward* algorithms that perform image-to-image comparisons as in (3.5) are used, presuming their equivalence. In the presence of camera blur, however, this turns out to be incorrect.

B2. Forward Algorithm

The corresponding algorithm to that in B1 is

Step 1: Find $\hat{\mathbf{W}}_{\mathbf{S1}}$ and $\hat{\mathbf{W}}_{\mathbf{S2}}$ using (3.18) and (3.19).

Step 2:

$$\hat{\mathbf{W}}_{\mathbf{12}} = \arg\min_{\mathbf{W_{12}}} \int_{\mathbf{z}\in\mathbf{dom}I_2} \Big[\underbrace{B\big(warp(S;\hat{\mathbf{W}}_{\mathbf{S2}})\big)}_{\hat{I}_2}(\mathbf{z}) - \underbrace{B\Big(warp\big(warp(S;\hat{\mathbf{W}}_{\mathbf{S1}});\mathbf{W_{12}}\big)\Big)}_{T}(\mathbf{z}) \Big]^2 \mathbf{dz}. \tag{3.20}$$

This algorithm could estimate the warp only if the scene S was known. The immediate question is whether we can follow the same steps as before, and use the images I_1 and I_2 in place of the warped scene. In the presence of blur, this turns out to be not always possible.

Note that the observed image I_2 is the ML estimate for

$$\hat{I}_2 = B\big(warp(S;\hat{\mathbf{W}}_{\mathbf{S2}})\big).$$

Suppose we denote by T the following "imaging" function

$$T = B\Big(warp\big(warp(S;\hat{\mathbf{W}}_{\mathbf{S1}});\mathbf{W_{12}}\big)\Big).$$

Then the image registration problem of (3.20) becomes

$$\hat{\mathbf{W}}_{\mathbf{12}} = \arg\min_{\mathbf{W_{12}}} \int_{\mathbf{z}\in\mathbf{dom}I_2} \Big[I_2(\mathbf{z}) - T(\mathbf{z}) \Big]^2 \mathbf{dz}. \tag{3.21}$$

Since T is still a function of the unknown S, it cannot be readily computed. For the sake of argument, consider changing the order of warp and blur operators in T, and define a new imaging function

$$T' = warp\Big(\underbrace{B\big(warp(S;\hat{\mathbf{W}}_{\mathbf{S1}})\big)}_{\hat{I}_1}; \mathbf{W_{12}} \Big).$$

The observed image I_1 is the ML estimate for $\hat{I}_1 = B\big(warp(S;\hat{\mathbf{W}}_{\mathbf{S1}})\big)$, and therefore,

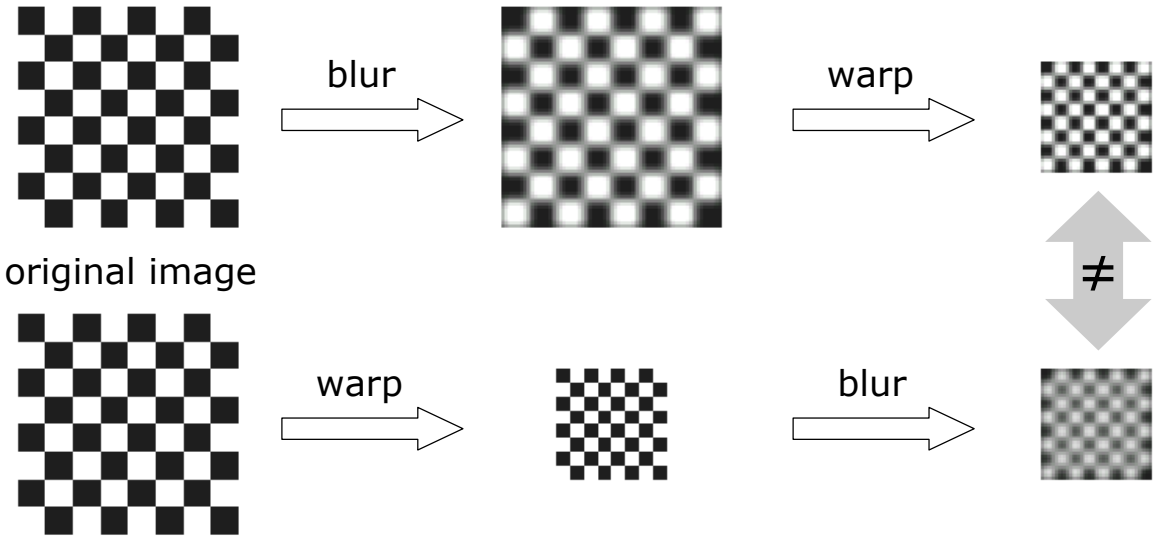

Figure 3.4: The ordering of blur and geometric warp operations is important: In this example, we used the same Gaussian blur kernel ($\sigma = 2$ pixels) before (top row) or after (bottom row) geometric scaling by a factor of $1/2$. Resulting images, shown on the right, differ from each other.

$T' = warp(I_1; \mathbf{W_{12}})$. That is, if we replace T by T' in (3.21), we would arrive at the commonly used form (3.5) of objective function in image registration (for $p = 2$):

$$\hat{\mathbf{W}}'_{\mathbf{12}} = \arg\min_{\mathbf{W_{12}}} \int_{\mathbf{z} \in \mathbf{dom} I_2} \Big[I_2(\mathbf{z}) - warp(I_1; \mathbf{W_{12}})(\mathbf{z}) \Big]^2 \mathbf{dz}. \tag{3.22}$$

However, the warp and blur operations do not commute in general. Fig. 3.4 illustrates this fact with a simple example. We therefore have $T \neq T'$, resulting in $\hat{\mathbf{W}}'_{\mathbf{12}} \neq \hat{\mathbf{W}}_{\mathbf{12}}$. Since $\hat{\mathbf{W}}'_{\mathbf{12}}$ does not coincide with the ML solution $\hat{\mathbf{W}}_{\mathbf{12}}$, it will be a *biased* estimator.

Compensating for the Bias

While $\hat{\mathbf{W}}'_{\mathbf{12}}$ is biased, there exist conditions under which T' can help us compute the unbiased estimate $\hat{\mathbf{W}}_{\mathbf{12}}$. To reveal when this would be possible, we express the blur operators in T and T' explicitly as convolution integrals. For notational conciseness, let us define $S' = warp(S; \hat{\mathbf{W}}_{\mathbf{S1}})$.

$$\begin{aligned} T(\mathbf{x}) &= B\Big(warp\big(warp(S;\hat{\mathbf{W}}_{\mathbf{S1}});\mathbf{W_{12}}\big)\Big)(\mathbf{x}) \\ &= B\big(warp(S';\mathbf{W_{12}})\big)(\mathbf{x}) \\ &= \int\limits_{\mathbf{w}\in\mathbf{dom}\,warp(S';\mathbf{W_{12}})} warp(S';\mathbf{W_{12}})(\mathbf{w})PSF\big(\mathbf{w}-\mathbf{x}\big)\mathbf{dw}. \end{aligned} \tag{3.23}$$

On the other hand,

$$\begin{aligned} T'(\mathbf{x}) &= warp\Big(B\big(warp(S;\hat{\mathbf{W}}_{\mathbf{S1}})\big);\mathbf{W_{12}}\Big)(\mathbf{x}) \\ &= warp\big(B(S');\mathbf{W_{12}}\big)(\mathbf{x}) \\ &= B\big(S'\big)\big(\mathbf{W_{12}}^{-1}(\mathbf{x})\big) \\ &= \int\limits_{\mathbf{v}\in\mathbf{dom}S'} S'(\mathbf{v})PSF\big(\mathbf{v}-\mathbf{W_{12}}^{-1}(\mathbf{x})\big)\mathbf{dv}. \end{aligned}$$

To rewrite the integral above in the domain of $warp(S';\mathbf{W_{12}})$, we define $\mathbf{w}=\mathbf{W_{12}}(\mathbf{v})$. As $\mathbf{dv}=\big|J(\mathbf{W_{12}}^{-1})\big|\mathbf{dw}$, changing the variable of integration of $\mathbf{v}$ to $\mathbf{w}$ will yield

$$T'(\mathbf{x}) = \int\limits_{\mathbf{w}\in\mathbf{dom}\,warp(S';\mathbf{W_{12}})} warp(S';\mathbf{W_{12}})(\mathbf{w})PSF\big(\mathbf{W_{12}}^{-1}(\mathbf{w})-\mathbf{W_{12}}^{-1}(\mathbf{x})\big)\big|J(\mathbf{W_{12}}^{-1})\big|\mathbf{dw}. \tag{3.24}$$

Intuition for the Restricted Case of a Similarity Transform

Observe the following two differences between T in (3.23) and T' in (3.24): first, the argument of the PSF in (3.24) is subject to a transformation. Second, the determinant of the warp's Jacobian appears as a multiplier in (3.24). Before discussing general properties of this difference and providing a concrete method for its elimination, we develop an intuition for a restricted case.

Let us consider $\mathbf{W_{12}}$ to be a similarity transformation, which can be parametrized using scale s, rotation θ, and translation (t_x, t_y) variables. The argument of the PSF in (3.24) is then

$$\begin{aligned}
\mathbf{W_{12}}^{-1}(\mathbf{w}) - \mathbf{W_{12}}^{-1}(\mathbf{x}) &= \left(\begin{bmatrix} s\cos\theta & -s\sin\theta \\ s\sin\theta & s\cos\theta \end{bmatrix}^{-1} \begin{bmatrix} w_x \\ w_y \end{bmatrix} - \begin{bmatrix} t_x \\ t_y \end{bmatrix} \right) - \left(\begin{bmatrix} s\cos\theta & -s\sin\theta \\ s\sin\theta & s\cos\theta \end{bmatrix}^{-1} \begin{bmatrix} x_x \\ x_y \end{bmatrix} - \begin{bmatrix} t_x \\ t_y \end{bmatrix} \right) \\
&= \begin{bmatrix} \frac{\cos\theta}{s} & \frac{\sin\theta}{s} \\ \frac{-\sin\theta}{s} & \frac{\cos\theta}{s} \end{bmatrix} \begin{bmatrix} w_x \\ w_y \end{bmatrix} - \begin{bmatrix} \frac{\cos\theta}{s} & \frac{\sin\theta}{s} \\ \frac{-\sin\theta}{s} & \frac{\cos\theta}{s} \end{bmatrix} \begin{bmatrix} x_x \\ x_y \end{bmatrix} \\
&= \begin{bmatrix} \frac{\cos\theta}{s} & \frac{\sin\theta}{s} \\ \frac{-\sin\theta}{s} & \frac{\cos\theta}{s} \end{bmatrix} \begin{bmatrix} w_x - x_x \\ w_y - x_y \end{bmatrix} \\
&= \mathbf{W}'(\mathbf{w} - \mathbf{x}),
\end{aligned}$$

where $\mathbf{W}'$ is a similarity transform with scale $\frac{1}{s}$, rotation θ, and zero translation. Furthermore, if the camera's PSF is rotation-invariant (*i.e.*, isotropic),

$$PSF\big(\mathbf{W}'(\mathbf{w} - \mathbf{x})\big) = PSF(\frac{\mathbf{w} - \mathbf{x}}{s}).$$

In summary, when $\mathbf{W_{12}}$ is limited to similarity transforms and the PSF is isotropic, (3.24) becomes

$$T'(\mathbf{x}) = \int_{\mathbf{w} \in \mathbf{dom}\, warp(S';\mathbf{W_{12}})} warp(S';\mathbf{W_{12}})(\mathbf{w}) PSF\big(\frac{\mathbf{w}-\mathbf{x}}{s}\big) \big| J(\mathbf{W_{12}}^{-1}) \big| \mathbf{dw}. \tag{3.25}$$

A comparison of (3.25) with (3.23) reveals how the imaging functions T' and T relate to each other. Although they are both obtained by blurring $warp(S';\mathbf{W_{12}})$, the actual blur kernels are different. Imagine that T has the blur kernel $PSF(\cdot)$, shown in the middle of Fig. 3.5. Since the blur kernel of T' is $PSF(\frac{\cdot}{s})$, it will have a dilated or compressed shape, and it will be area-normalized through the multiplicative term $\big|J(\mathbf{W_{12}}^{-1})\big| = \frac{1}{s^2}$. For $0 < s < 1$, the kernel gets compressed (Fig. 3.5, left), resulting in a T' less blurry than T. For $s = 1$, we have equality between T and T'. Finally, for $s > 1$, the effective blur kernel becomes wider (Fig. 3.5, right), causing T' to be even more blurred than T.

The analysis above provides the conditions under which T' can be used in emulating T, and the *forward* algorithm B2 still work even if the scene function S is unknown:

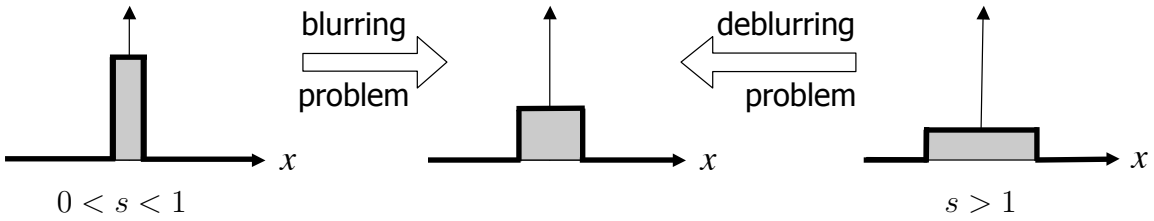

Figure 3.5: Given I_1, we can compute $T' = \mathbf{W_{12}}(I_1)$. However, to find the ML warp estimates, we need to evaluate T, which originally results from a convolution operation with the PSF of the camera (middle). Depending on the value of s, estimating T from T' turns out to be a blurring ($0 < s < 1$) or deblurring ($s > 1$) problem.

- For $s = 1$, T' can readily replace T.

- For $0 < s < 1$, we may blur T' further to make up for the difference in blur kernels $PSF(\cdot)$ and $\frac{1}{s^2}\ PSF(\frac{\cdot}{s})$. Only after this blur compensation would the minimizer of (3.22) correspond to *forward* algorithm's warp estimate.

- For $s > 1$, the wider blurring kernel produces an overblurred T', and emulating T is therefore a deblurring problem: this is in general an ill-posed inverse problem, and difficult to solve [Banham and Katsaggelos, 1997].

Note that the quantities T and T' in (3.23) and (3.25) were derived for the *forward* algorithm. By definition, when the *forward* algorithm scales up (*i.e.*, $s > 1$), the *backward* algorithm scales down ($0 < \frac{1}{s} < 1$). Therefore, in situations where $s > 1$, the deblurring problem can be avoided by simply switching to the algorithm which solves for the warp in the opposite direction. Hence, for obtaining an unbiased estimate of the warp between two images, there is a *natural choice* between the *forward* and *backward* algorithms: One should pick the direction of warp such that, after necessary blurring, it scales one image down onto the other, *i.e.*, the higher-resolution image should be warped onto the lower[2]. Since the two images are not interchangeable, their registration is an *asymmetric* problem.

[2]Could one still blur the high-resolution image, but warp the low-resolution image onto the higher-resolution one instead? It is hard to tell which optimization criterion this approach would be minimizing, and whether its solution would correspond to a Least-Squares estimate of the warp.

More General Cases

For more complex warps than similarity, the blur varies spatially and the analysis above does not apply. The inequality between T and T' is still due to the difference between the effective blur kernels in (3.23) and (3.24), but the analysis becomes harder. Probably the general solution is to use the *generative* algorithm and explicitly recover S.

In specific cases, however, it may be possible to derive an algorithm, but this has to be done on a case-by-case basis. In section 3.3, we do this for the case of piecewise affine warps used in AAMs.

C2. Backward Algorithm

The above analysis also applies to the *backward* algorithm: if it happens to be scaling down the higher-resolution image onto the lower-resolution one, incorporating a blurring step will ensure that the warp estimate will be unbiased. However, if it happens to be scaling the lower-resolution image onto the higher, it will remain biased unless we deblur the low-resolution image.

3.2.4 The Asymmetry Principle

We argued that the problem of image registration is inherently asymmetric, and that ignoring this fact leads to biased estimates. We used a simple yet illuminating scenario starting with an idealized setting wherein the underlying scene radiance field was known. We presented three algorithms (*generative*, *forward* and *backward*) to estimate the geometric warp which maps between the image coordinate frames. We then investigated how these algorithms could be used in the absence of scene information. Our analysis exposed the conditions under which *forward* and *backward* algorithms could estimate the warp based on the images only, and prescribed a specific blurring step in the presence of relative scaling between images. Such cases turned out to impose a particular warp direction for ensuring unbiased estimates.

Our asymmetry claim is based on the scaling-induced extra blurring that neither a *forward* nor a *backward* algorithm can overcome. It depends upon whichever happens to be warping the lower-resolution image onto the higher-resolution one. We have shown that an inability to deblur, which is an ill-posed inverse problem, results in a bias that is inde-

pendent of the assumed observation noise model. The analysis remains applicable to other cases where blur-related discrepancies result not necessarily from camera poses and zoom levels, but from imaging modalities or instrument characteristics.

Quantification of the scaling-induced bias is not trivial because blur effects are ultimately related to image content: while blurring (*i.e.*, low-pass filtering) visually rich and detailed images would produce a significant effect, it would barely alter already smooth images. In Appendix B, we quantify the magnitude of the bias in translation-only registration problems between face images of different resolutions.

What does the asymmetry imply for Face Hallucination with AAMs? Note that the traditional fitting criterion (3.4) completely overlooks the bias problem: it warps and interpolates low-resolution observations to compare them with the model template. This is an artifact of formulation and it will cause the fitting accuracy to drop with higher scaling factors. As remedy, the next section formulates a novel algorithm that respects the asymmetry and incorporates an explicit model of the blur into the fitting formulation.

3.3 Resolution-Aware Fitting

3.3.1 Formulation

We propose an alternative to the fitting criterion (3.4). Recognizing the asymmetry of the problem, we not only change the warp direction, but also introduce a model of the blur/image formation process. From a generative point of view, this simulates the pixel-wise image formation process in a CCD camera [Barbe, 1980]. We feed the AAM and its current parameters into a camera model, and compare the outcome against the observed low-resolution image. This process is illustrated in Fig. 3.6.

Mathematically, the proposed fitting criterion is

$$\sum_{\mathbf{u}\in\mathbf{dom}L} \big[L(\mathbf{u}) - B\big(\mathbf{u}; H(\mathbf{W}(\mathbf{p});\boldsymbol{\lambda})\big)\big]^2 , \tag{3.26}$$

where the summation is now over pixel coordinates $\mathbf{u}$ of the observed image L. That is, if (3.4) was the *forward* algorithm of Section 3.2, (3.26) is the *backward* algorithm with an additional blur operator B. This blur simulates a low-resolution image of the object, believed to be what the camera would have captured under current AAM parameters[3].

[3] If the camera is expected to have aliasing, our method should simulate that as well.

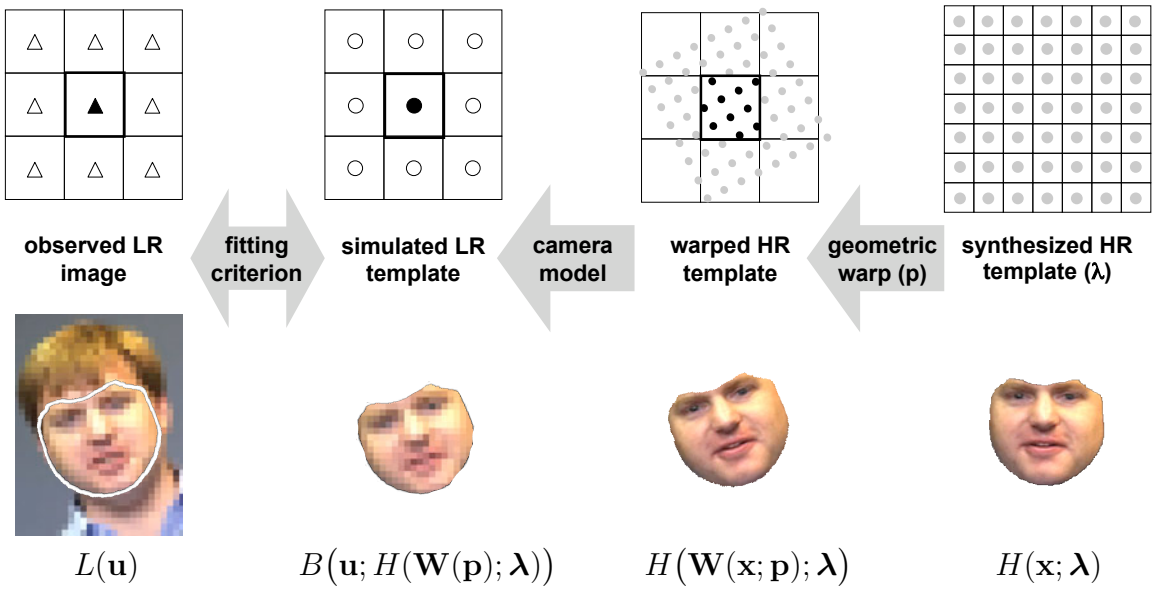

Figure 3.6: The Resolution-Aware Fitting algorithm simulates the formation of low-resolution images in a digital camera. In contrast to the traditional formulation shown in Fig. 3.2, the fitting criterion is defined between observed and simulated image pixels.

Although this formulation can accommodate arbitrary camera models and point spread functions, in this thesis, we use the rectangular PSF

$$B\big(\mathbf{u}; H(\mathbf{W}(\mathbf{p}); \boldsymbol{\lambda})\big) = \frac{1}{area(bin(\mathbf{u}))} \int_{\mathbf{u}' \in bin(\mathbf{u})} H\big(\mathbf{W}^{-1}(\mathbf{u}'; \mathbf{p}); \boldsymbol{\lambda}\big) d\mathbf{u}',$$

where the continuous integral is defined over $bin(\mathbf{u})$, the sensing area of the discrete pixel $\mathbf{u}$. As illustrated in Fig. 3.6, the blur operator itself is independent of AAM parameters. It simply averages out those template pixel intensities which map into a low-resolution pixel's sensing area under the current warp $\mathbf{p}$. To express the integral above in the shape-normalized coordinate frame $\mathbf{s}_0$, we observe that $\mathbf{u}' = \mathbf{W}(\mathbf{x}; \mathbf{p})$, and consequently, $d\mathbf{u}' = \big|J\big(\mathbf{W}(\mathbf{p})\big)\big| d\mathbf{x}$,

$$B\big(\mathbf{u}; H(\mathbf{W}(\mathbf{p}); \boldsymbol{\lambda})\big) = \frac{1}{area(bin(\mathbf{u}))} \int\limits_{\substack{\mathbf{x} \in \mathbf{dom}\, s_0 \text{ s.t.} \\ \mathbf{W}(\mathbf{x}; \mathbf{p}) \in bin(\mathbf{u})}} H(\mathbf{x}; \boldsymbol{\lambda}) \big| J\big(\mathbf{W}(\mathbf{p})\big) \big| d\mathbf{x}.$$

We implement this integration as a discrete, Jacobian-weighted sum over template pixels,

$$B\big(\mathbf{u}; H(\mathbf{W}(\mathbf{p}); \boldsymbol{\lambda})\big) = \frac{1}{area(bin(\mathbf{u}))} \sum_{\substack{\mathbf{x} \in \mathbf{dom}\, s_0 \text{ s.t.} \\ \mathbf{u} - \left[\begin{smallmatrix} .5 \\ .5 \end{smallmatrix}\right] < \mathbf{W}(\mathbf{x}; \mathbf{p}) < \mathbf{u} + \left[\begin{smallmatrix} .5 \\ .5 \end{smallmatrix}\right]}} H(\mathbf{x}; \boldsymbol{\lambda}) \big| J\big(\mathbf{W}(\mathbf{p})\big) \big|. \tag{3.27}$$

Observe that our formulation avoids interpolating low-resolution data, and models the object appearance, geometric deformation, and image formation processes simultaneously.

3.3.2 Algorithm Derivation

We now present a Gauss-Newton gradient-descent algorithm for the minimization of the fitting criterion (3.26) with respect to $\mathbf{p}$ and $\boldsymbol{\lambda}$. This algorithm gives up the computational efficiency of [Matthews and Baker, 2004] in exchange for a more accurate/unbiased estimate of the parameters. Until convergence, updates $\Delta\mathbf{p}$ and $\Delta\boldsymbol{\lambda}$ are iteratively computed and added to the current estimates. The derivation below closely follows that of the *simultaneous* algorithm in [Baker et al., 2003]. Expressing $A(\boldsymbol{\lambda})$ as a sum of the mean and linearly weighted basis images, the fitting criterion is

$$\sum_{\mathbf{u} \in \mathbf{dom} L} \left[L(\mathbf{u}) - B\Big(\mathbf{u}; H_0\big(\mathbf{W}(\mathbf{p})\big) + \sum_{i=1}^{m} \lambda_i H_i\big(\mathbf{W}(\mathbf{p})\big)\Big) \right]^2.$$

Consider the Taylor expansion

$$\sum_{\mathbf{u} \in \mathbf{dom} L} \left[L(\mathbf{u}) - B\Big(\mathbf{u}; H_0\big(\mathbf{W}(\mathbf{p}+\Delta\mathbf{p})\big) + \sum_{i=1}^{m} (\lambda_i + \Delta\lambda_i) H_i\big(\mathbf{W}(\mathbf{p}+\Delta\mathbf{p})\big)\Big) \right]^2.$$

Ignoring its second-order terms, the fitting criterion is approximately

$$\sum_{\mathbf{u} \in \mathbf{dom} L} \left[L(\mathbf{u}) - B\Big(\mathbf{u}; H_0\big(\mathbf{W}(\mathbf{p})\big) + \nabla H_0 \frac{\partial \mathbf{W}}{\partial \mathbf{p}} \Delta\mathbf{p} + \sum_{i=1}^{m} (\lambda_i + \Delta\lambda_i) \Big(H_i\big(\mathbf{W}(\mathbf{p})\big) + \nabla H_i \frac{\partial \mathbf{W}}{\partial \mathbf{p}} \Delta\mathbf{p}\Big)\Big) \right]^2.$$

For notational conciseness, denote $n+m$ steepest-descent images as

$$\mathbf{SD}_{sim}=\left[\left(\nabla H_0+\sum_{i=1}^{m}\lambda_i\nabla H_i\right)\frac{\partial\mathbf{W}}{\partial p_1},\ldots,\left(\nabla H_0+\sum_{i=1}^{m}\lambda_i\nabla H_i\right)\frac{\partial\mathbf{W}}{\partial p_n},H_1\big(\mathbf{W}(\mathbf{p})\big),\ldots,H_m\big(\mathbf{W}(\mathbf{p})\big)\right].$$

We can now compactly rewrite the fitting criterion as

$$\sum_{\mathbf{u}\in\mathbf{dom}L}\left[L(\mathbf{u})-B\left(\mathbf{u};H_0(\mathbf{W}(\mathbf{p}))+\sum_{i=1}^{m}\lambda_iH_i(\mathbf{W}(\mathbf{p}))-\mathbf{SD}_{sim}\begin{pmatrix}\Delta\mathbf{p}\\ \Delta\lambda\end{pmatrix}\right)\right]^2.$$

Observing that B is a linear operator, the objective function to be minimized is

$$\sum_{\mathbf{u}\in\mathbf{dom}L}\left[L(\mathbf{u})-B\Big(\mathbf{u};H_0\big(\mathbf{W}(\mathbf{p})\big)\Big)+\sum_{i=1}^{m}\lambda_iB\Big(\mathbf{u};H_i\big(\mathbf{W}(\mathbf{p})\big)\Big)-B\big(\mathbf{u};\mathbf{SD}_{sim}\big)\begin{pmatrix}\Delta\mathbf{p}\\ \Delta\lambda\end{pmatrix}\right]^2,$$

whose minimum is given by

$$\begin{pmatrix}\Delta\mathbf{p}\\ \Delta\lambda\end{pmatrix}=-H_{sim}^{-1}\sum_{\mathbf{u}\in\mathbf{dom}L}B\big(\mathbf{u};\mathbf{SD}_{sim}^{\mathrm{T}}\big)\left[L(\mathbf{u})-B\Big(\mathbf{u};H_0\big(\mathbf{W}(\mathbf{p})\big)\Big)+\sum_{i=1}^{m}\lambda_iB\Big(\mathbf{u};H_i\big(\mathbf{W}(\mathbf{p})\big)\Big)\right],$$

where H_{sim} is the Hessian with appearance variation:

$$H_{sim}=\sum_{\mathbf{u}\in\mathbf{dom}L}B\big(\mathbf{u};\mathbf{SD}_{sim}^{\mathrm{T}}\big)B\big(\mathbf{u};\mathbf{SD}_{sim}\big).$$

3.4 Experiments

We will compare the RAF formulation (3.26) to the traditional formulation in (3.4). In particular, we will compare the algorithm detailed in Section 3.3.2 (referred to as RAF) with the simultaneous algorithm of [Gross et al., 2005] (referred to as AAMR-SIM), which optimizes (3.4). Earlier work [Baker et al., 2003; Gross et al., 2005] empirically showed that simultaneously solving for the shape and appearance parameters performs better than projecting out the appearance, although at a greater computational cost. Comparing with the simultaneous algorithm [Gross et al., 2005] is therefore a fairer comparison than with the project-out algorithm [Matthews and Baker, 2004].

We performed two types of experiments: synthetic and real. First, we synthetically downsampled images, and compared our fitting (and hallucination) results against high-resolution "ground truth" fits. We generated a variety of input test sequences by a range

of scaling factors, and measured each algorithm's accuracy at lower input resolutions. These will be presented as quantitative results. Second, we ran our algorithms on real low-resolution images to capture their performance under real noise processes. We present these results qualitatively.

Independently of the resolution of a given test sequence, we initialized all algorithms with fitting results at the highest resolution. This allowed us to discard initialization quality as a confounding factor when comparing performances across resolution levels. While manual initialization is reasonable at higher resolutions, it becomes increasingly sub-optimal in lower resolutions, jeopardizing the fairness of comparisons across scales. Once in tracking mode, the fitting of each frame was initialized with the parameters of the preceding frame.

3.4.1 Metrics of Fit and Hallucination Accuracy

The most appropriate metric of fit quality depends on applications. Face Hallucination aims to synthesize detailed face images and their expressions correctly; therefore accurate estimates are required for both shape and appearance parameters. Other domains may be less demanding: in object tracking, only the global pose (*i.e.*, the similarity transform parameters) may be of interest. In lip-reading, non-rigid deformations of a speaker's lips (encoded by a facial AAM's shape coefficients) may carry all the necessary information.

We defined two metrics, illustrated in Fig. 3.7, to summarize the fitting accuracy of the RAF and AAMR-SIM algorithms. The *tracking error* is the average of positional error of landmarks (such as the corner of nostrils): this error is a combined effect of both similarity transform (scale, rotation, and translation) and non-rigid deformation parameters, as encoded by the estimate $\hat{\mathbf{p}}$. The *reconstruction error*, on the other hand, is computed by comparing the synthesized (hallucinated) model instance, parametrized by $\hat{\boldsymbol{\lambda}}$, against the ground truth image in terms of RMS error of intensities. In addition, we report estimation errors for the coefficients of the top four principal shape and appearance modes.

For all test sequences included in this chapter, only the landmark coordinates were available as hand-labeled, high-resolution ground truth data. To infer the ground truth values for the similarity, non-rigid shape and appearance variables, we ran the AAMR-SIM tracker at the original resolution of the videos, and verified its convergence (each landmark's tracking error smaller than 1 high-resolution pixel). The resulting parameters were then regarded as "ground truth" in subsequent low-resolution tests.

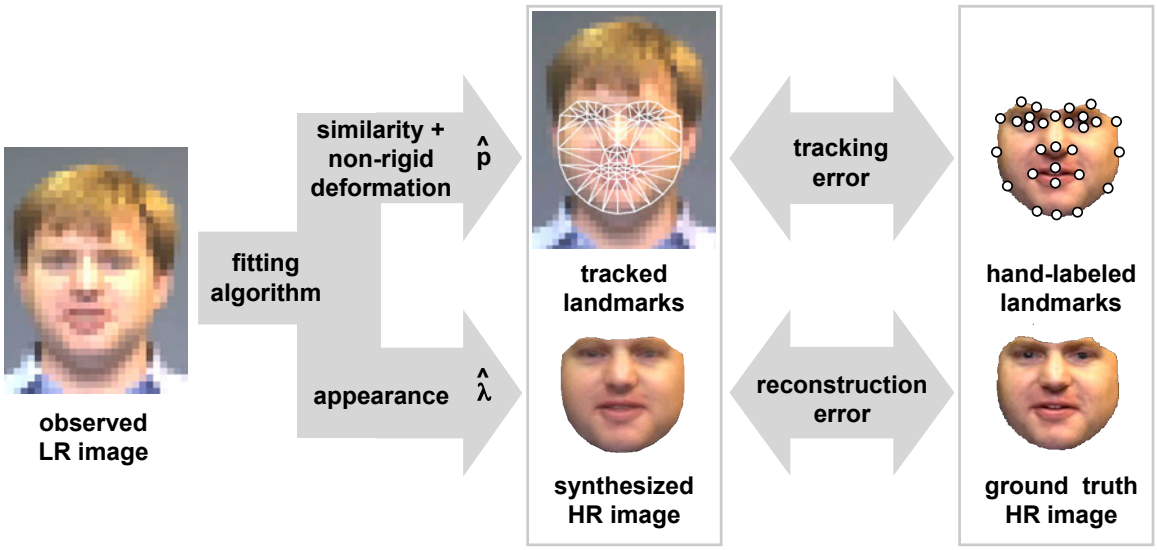

Figure 3.7: We define two metrics to compare the fitting accuracy of algorithms. The average landmark tracking error combines the estimation accuracy of the similarity and non-rigid shape parameters. The reconstruction (*i.e.*, hallucination) error quantifies how well the underlying high-resolution face was inferred from the low-resolution data.

3.4.2 Examples

Before presenting extensive quantitative results, we begin with some examples of our error metrics and their temporal behavior. In reporting Euclidean distance metrics (as in translation parameters or landmark tracking error), we scale-normalize the estimates so that their numerical values are in high-resolution pixel units. Similarly, we normalize each shape and appearance coefficient according to its mode's variance and report them in units of their standard deviation.

Fig. 3.8 plots error trajectories of a low-resolution tracking experiment, where the subject's speaking and eye blinking were the major sources of motion. The input sequence was 10 times lower in resolution than the AAM. The error metrics indicate that RAF tracked the face consistently better than AAMR-SIM. To provide further evidence, Fig. 3.9 shows temporal trajectories of selected variables. Those estimated by AAMR-SIM do not follow the ground truth values, and remain mostly constant. In contrast, RAF can recover the non-rigid deformations and appearance changes, amounting to a more accurate recovery of the facial expressions[4].

[4]Demonstration videos are available at http://www.cs.cmu.edu/~dedeoglu/thesis

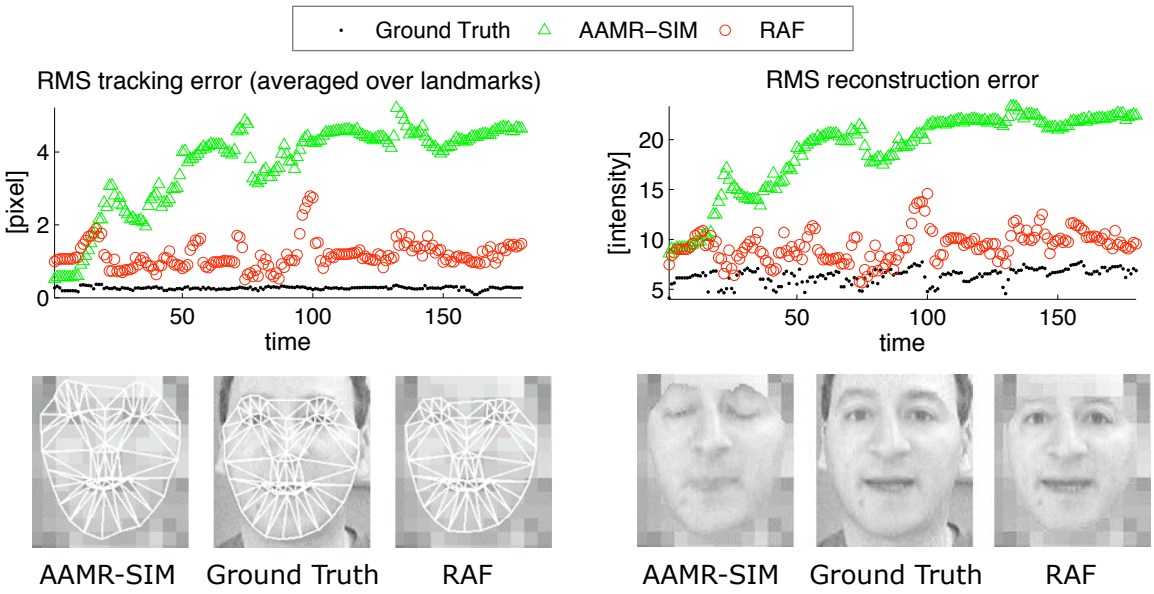

Figure 3.8: The landmark tracking (upper left) and reconstruction/hallucination (upper right) error metrics are plotted as a function of time for a 10-fold resolution degraded tracking experiment. Included images (bottom, captured at frame no. 102) display the mesh fits (lower left) as well as synthesized (hallucinated) model images (lower right). We overlay the latter onto pixel-replicated low-resolution inputs to demonstrate how well the underlying high-resolution image could be inferred.

3.4.3 Quantitative Evaluation

It would be impractical to report time trajectories for all our experiments. In the following, we report the temporal mean and standard deviation of the Root Mean Squared (RMS) errors of selected variables. Since lost trackers can easily corrupt these statistics with outliers, we required both trackers to produce valid results (*i.e.*, not have lost track of the face) for a fitting instance to be included in these statistics. This was achieved by manually inspecting all experiments and verifying that faces were tracked reasonably well.

Recall that each tracking experiment was initialized with the highest-resolution fitting results. At lower input resolutions, such an optimistic initialization would cause the fitting performance to be overestimated at the beginning. To minimize this effect, we discarded the fitting results of the first 20 frames of each sequence.

Fig. 3.10 compares the AAMR-SIM and RAF algorithms for fitting a single-person AAM. The list at upper-left corner provides a brief summary of experimental conditions:

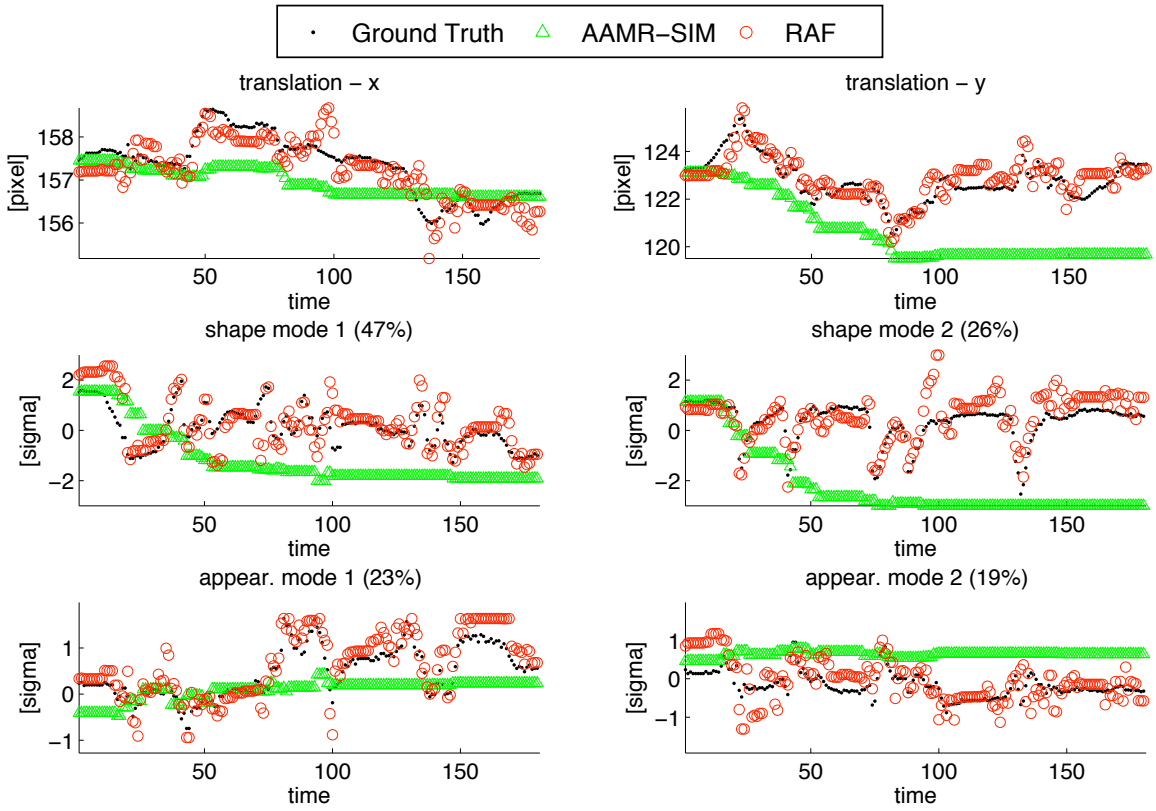

Figure 3.9: Selected temporal trajectories are shown for a 10-fold resolution degraded face tracking experiment. As the supplemental video material shows, the main source of motion were the subject's speaking and eye blinking. See Fig. 3.8 for one example frame of this sequence. The estimates of AAMR-SIM do not follow the ground truth, and remain mostly constant. In contrast, RAF remains close to ground truth in all trajectories, indicating that it is able to extract the underlying facial expressions correctly.

this AAM was built using 31 training images and was tested on a set of 180. These were 8-bit grayscale images and the AAM's native resolution was `100x104` pixel. We retained 95% of the total variation, yielding 11 shape and 23 appearance principal components.

The plots in Fig. 3.10 present extensive quantitative comparisons between the fitting algorithms. They are organized to show RMS error metrics as a function of downscaling factor. Observe how AAMR-SIM and RAF perform equally well at downsampling factor 2. This case corresponds to a minor degradation in resolution, and the fact that both algorithms perform similarly confirms the correctness of our derivations as well as implementations. Starting from downsampling factor 4, RAF brings substantial accuracy improvements over

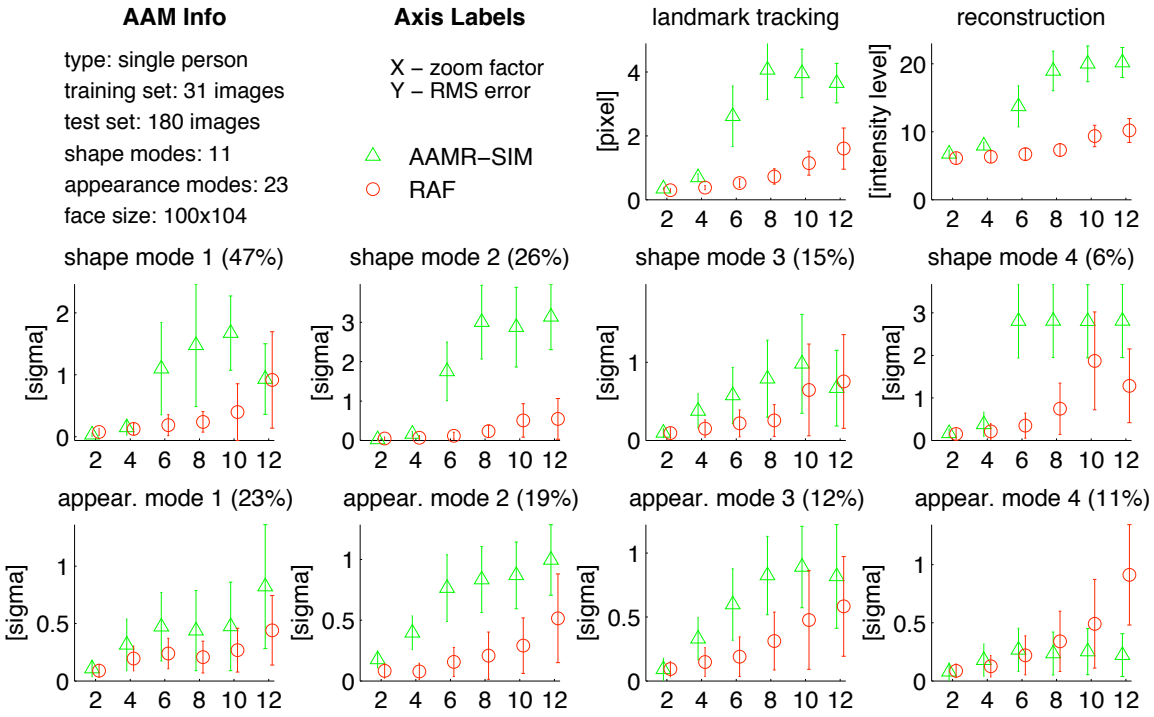

Figure 3.10: Quantitative comparison between the AAMR-SIM and RAF algorithms for fitting the single-person AAM to a 180 frame-long sequence. The horizontal axis is the downscaling factor of the input data. Both algorithms perform well at half-resolution, validating the derivation and implementation of RAF. The latter brings substantial improvements across all metrics for downscaling factors 4 and higher. The principal modes are displayed in order of % energy (*i.e.*, variation) they capture.

AAMR-SIM across all metrics and variables.

The performance of a model-based method ultimately depends on the quality of the available model. In order to investigate how the AAM fitting accuracy varies with model complexity, we also ran our experiments on a multiperson AAM, which we built using data from 5 subjects. Details of this AAM are provided in the upper-left corner of Fig. 3.11, organized in the same fashion as Fig. 3.10. The multiperson appearance model has almost twice the number of appearance modes compared to the single-person case, indicating a richer subspace being modeled. Again, RAF is consistently superior to AAMR-SIM in accuracy with regard to both tracking and reconstruction.

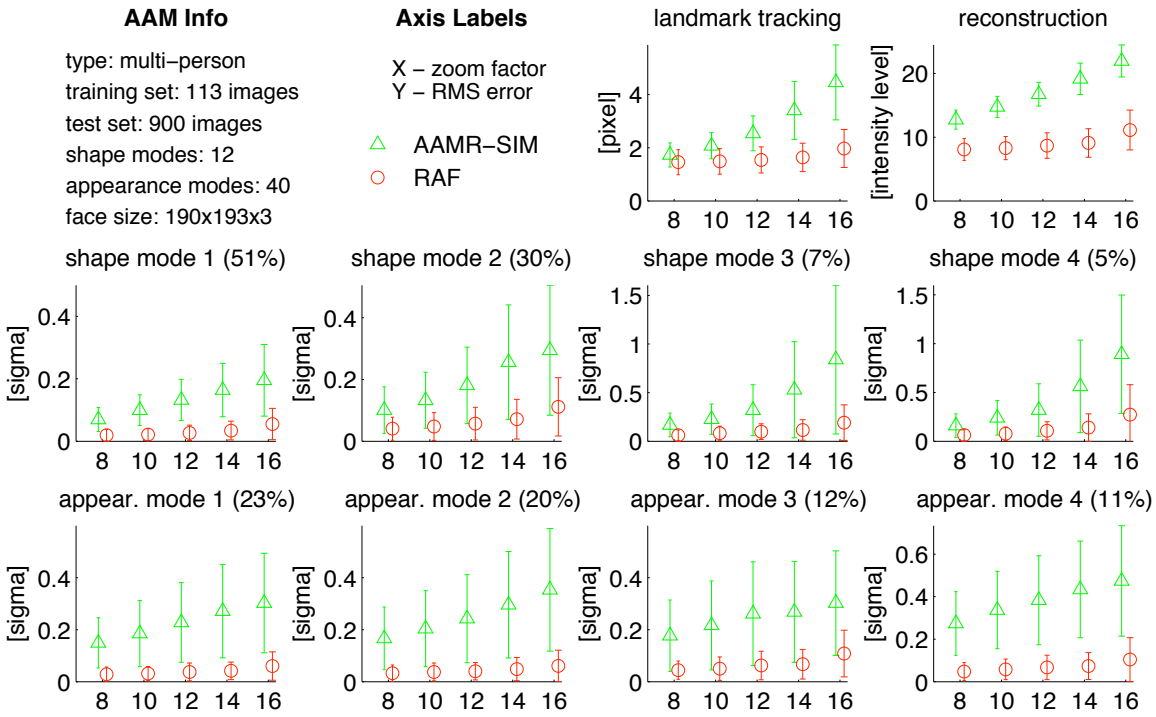

Figure 3.11: Quantitative comparison between the AAMR-SIM and RAF algorithms for fitting the multiperson (5 subjects) AAM. The horizontal axis is the downscaling factor of the input data. Each reported mean and standard deviation is calculated over 900 frames, comprising 180 frames for each of 5 subjects. RAF improves the tracking, reconstruction, non-rigid shape, and appearance estimates considerably.

3.4.4 Qualitative Results

Simulated Low-Resolution Data

We can qualitatively evaluate a fitting result by inspecting its reconstructed/hallucinated face image. In the following, we overlay such reconstructions on pixel-replicated original low-resolution inputs at where the trackers thought the faces were.

Fig. 3.12 shows every second frame in a sequence of the single-person AAM tracking experiment. Observe that RAF correctly extracts the eye blink and mouth opening, whereas AAMR-SIM does not. Fig. 3.13 offers a visual alternative for assessing how the trackers degrade with increased downscaling: it displays the single-person AAM results for frame no. 102 across various scales. While RAF can consistently recover the open eyes and mouth, AAMR-SIM's estimates degrade quickly: starting from downsampling factor 6,

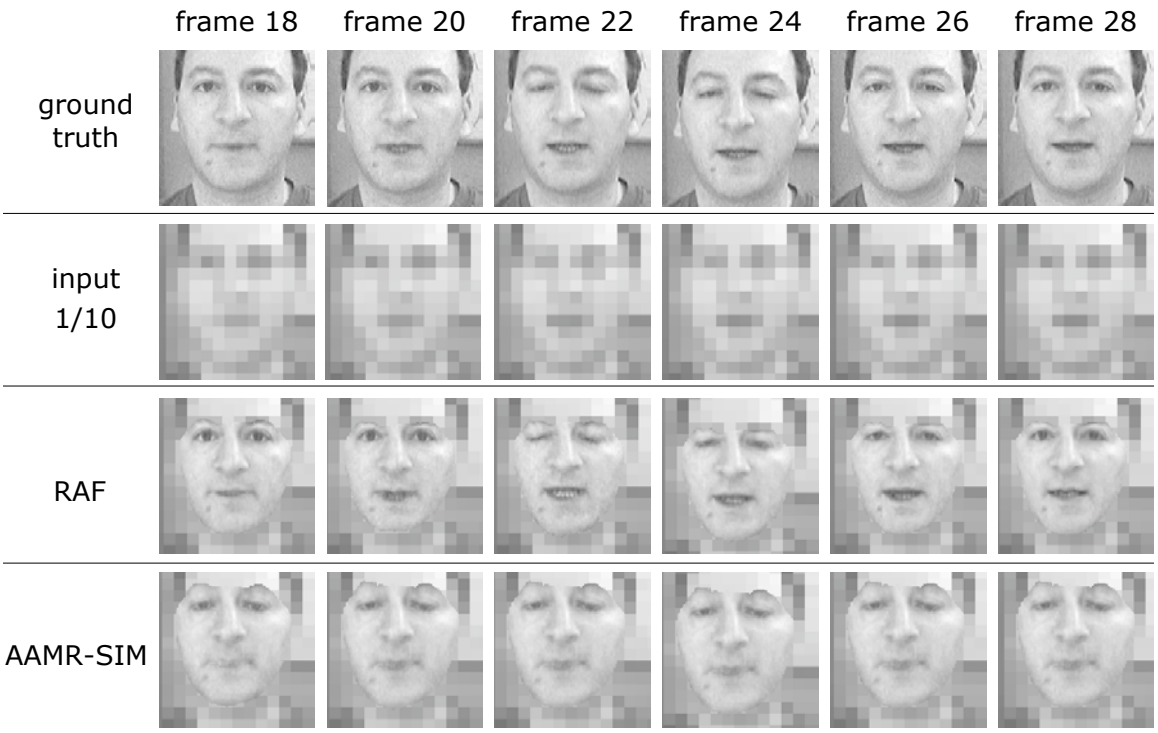

Figure 3.12: Exemplar subsequence of high-resolution reconstructions (*i.e.*, hallucinations), obtained by fitting the single-person AAM. Observe how RAF correctly extracts the eye blink and mouth opening, whereas AAMR-SIM does not.

the eyes and mouth are first estimated to be half-open, and then totally closed. Similarly, Fig. 3.14 displays snapshots of different test subjects, all tracked using the multiperson AAM of Fig. 3.11. In both single- and multiperson AAMs, we find the visual reconstruction quality of RAF to be consistently superior to that of AAMR-SIM.

Real Low-Resolution Data

We also compared the two AAM fitting algorithms on real low-resolution videos. Using a Sony DCR-VX2000 camera (15 fps in progressive mode and DV-format compression), we video-taped a particular subject's face at various distances, yielding face heights between 20 and 120 pixels in images. At each camera distance, the subject uttered the sequence "left-right-up-down-smile" and moved her face accordingly. We built a face AAM using 43 of high-resolution frames, and verified its tracking and reconstruction performance in that resolution. The AAM was `110x114x3` pixels, with 14 non-rigid shape and 27 ap-

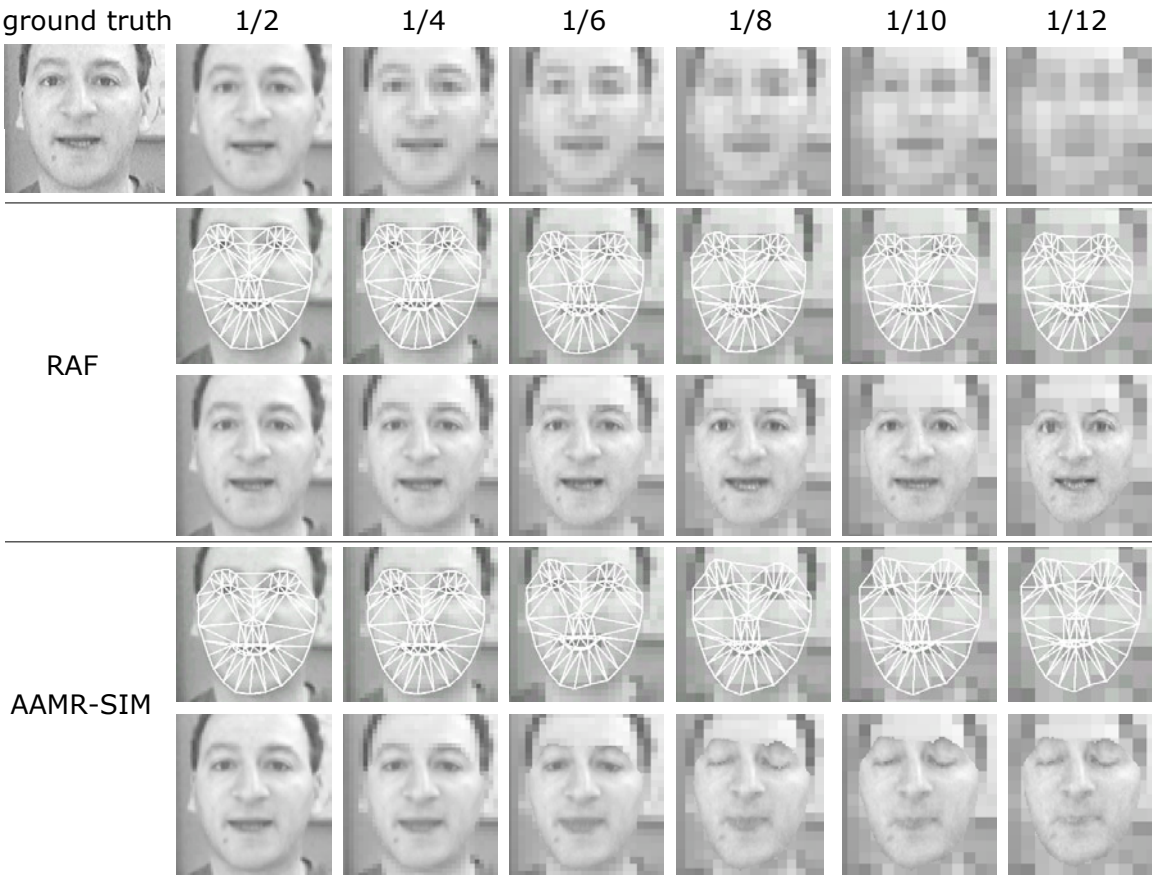

Figure 3.13: We compared the AAMR-SIM and RAF algorithms over a range of scales. Lower and lower resolution versions of input frame no. 102 are shown in the top row. While AAMR-SIM degrades quickly, RAF maintains a reasonable estimate of the face.

pearance modes. We fit this AAM to videos using the AAMR-SIM and RAF algorithms. Initialization was done manually, by scaling and positioning the AAM.

Fig. 3.15 compares face reconstructions for an eye-blink subsequence. The observed face is 33 pixels high, corresponding to a downscaling factor of about 3. Note the sharpness of RAF hallucinations. In contrast, AAMR-SIM misses the eye-blink, and hallucinates blurrier faces.

On all video sequences with downscaling factors 3.5 and higher (where the face height ranged from 30 to 20 pixels), AAMR-SIM consistently lost track of the face. In contrast, RAF kept tracking and reconstructing the face reasonably well. In Fig. 3.16, we include selected frames of RAF hallucinations. The faces are approximately 22 pixels high

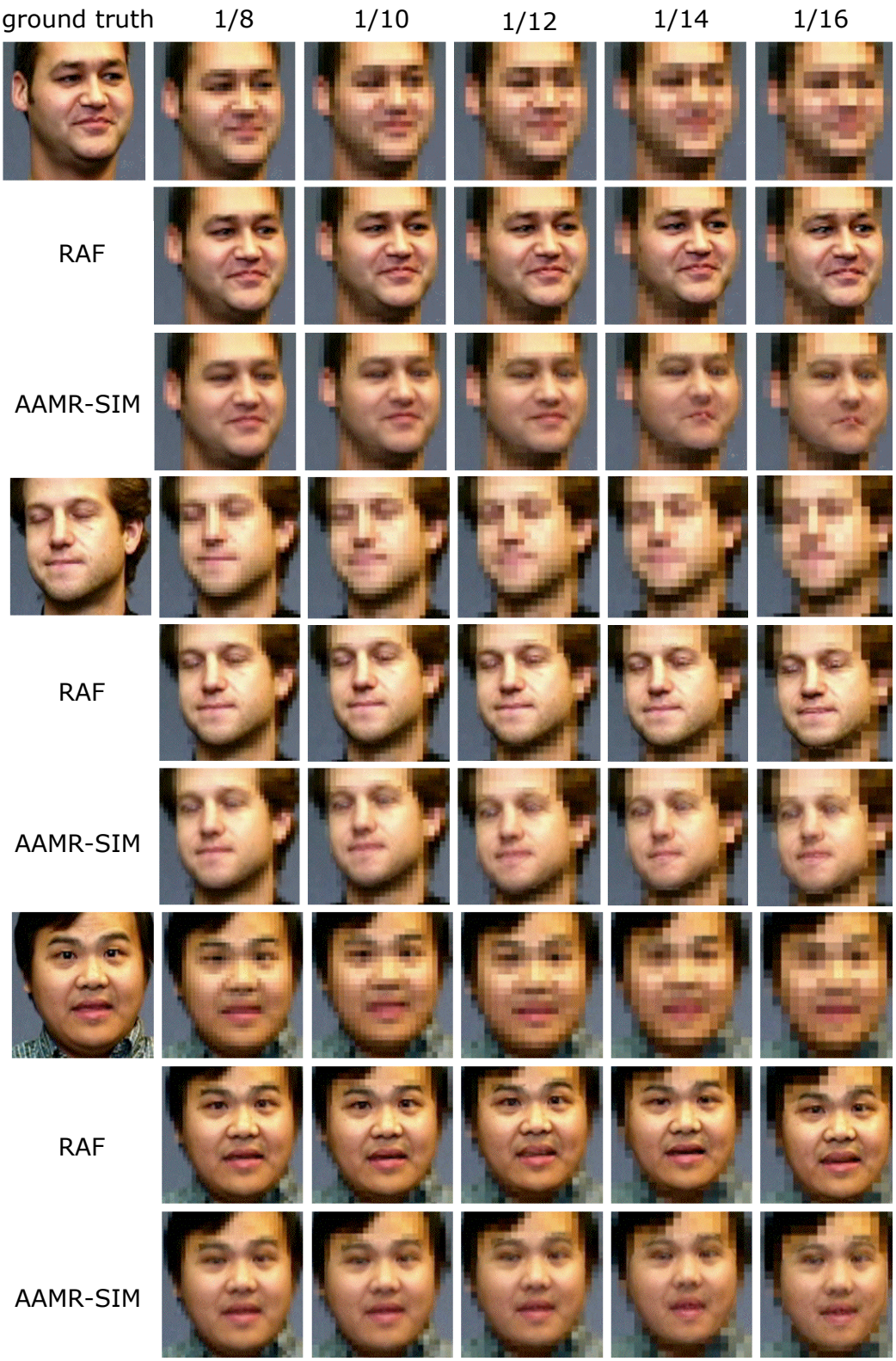

Figure 3.14: Selected frames are shown to visually compare hallucinated faces by fitting a multiperson AAM. The quantitative improvement in appearance estimates (Fig. 3.11) has visible effects. Mesh displays are omitted due to a lack of significant difference.

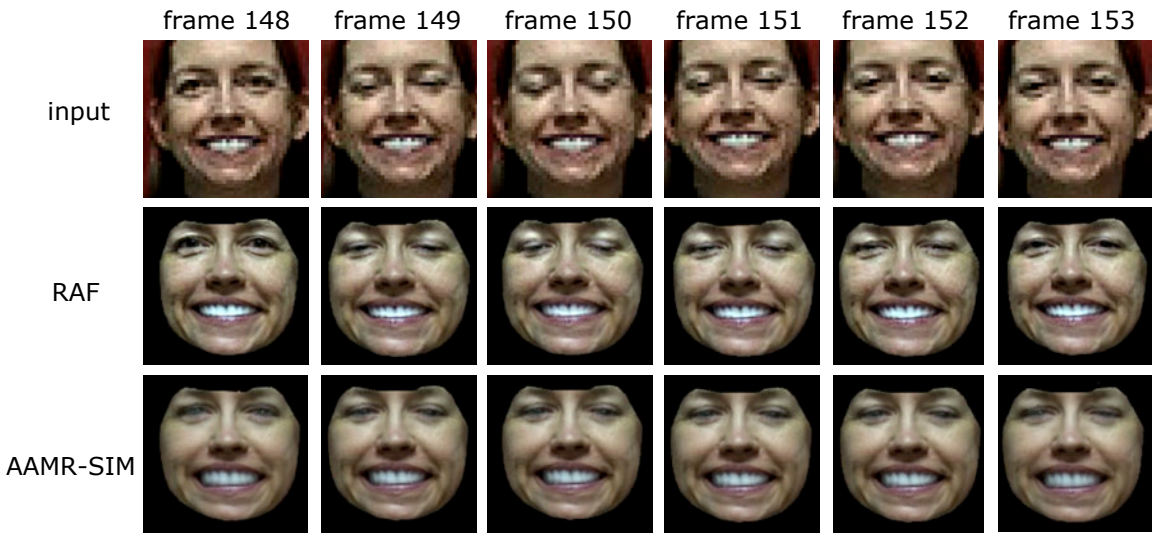

Figure 3.15: The top row shows DV-compressed video frames. The face is about 33 pixels high (downscaling factor $\sim$3). AAMR-SIM (bottom row) misses the eye-blink, and reconstructs overly smooth faces. Indeed, AAMR-SIM fails to track faces any smaller than this size. In contrast, RAF (middle row) infers and reconstructs the underlying facial expression with crisp details.

(downscaling factor $\sim$5). At this resolution, RAF can still recover the underlying facial expression, but the reconstructions start growing unstable.

3.5 Discussion

3.5.1 Performance Metrics

When facial features are not estimated correctly, Face Hallucination produces high-resolution, yet incorrect and unrealistic face images. Since hallucination fidelity to the underlying face was our priority, we exclusively focused on *accuracy* measures in comparing AAM fitting algorithms. However, in other applications (*e.g.*, non-rigid registration of medical images), criteria such as the repeatability, robustness and efficiency may be as important.

In extremely low resolutions, we found the AAMR-SIM algorithm to be more robust than RAF. Given the smoothing effect of (bilinear) interpolation, this does not seem surprising. While RAF struggles among the many parameter settings which yield almost the same low-resolution images, AAMR-SIM commits to an interpolated high-resolution observation and pursues the fit. In the future, we plan to study RAF's robustness more closely.

Figure 3.16: Selected reconstructions on DV-compressed video. Although the face is only 22 pixels high (downscaling factor $\sim$5), RAF can still track the face, recover its expressions, and reconstruct it reasonably well. The temporal jitter and instability observed at this resolution can be seen in videos available at http://www.cs.cmu.edu/~dedeoglu/thesis.

3.5.2 Computational Implications of the Asymmetry Principle

Constraints on the direction of the geometric warp have important consequences. Recall how the traditional AAM fitting criterion (3.4) had conveniently defined the summation over the model template pixels. Since the latter do not change as a function of the input, computational savings become possible. For instance, the inverse-compositional tracker of [Matthews and Baker, 2004] considers the Taylor expansion for the warp parameters over the model (*i.e.*, AAM appearance basis) and pre-computes all associated Jacobians and Hessians. Unfortunately, the RAF formulation of Section 3.3 does not benefit from such pre-computations. Implemented in MATLAB, the average fitting time (in tracking mode) of AAMR-SIM is 3 seconds, whereas RAF takes about 10 times longer. One area for future work is to investigate how such savings may be possible.

Unless deblurring becomes possible, either the *forward* or the *backward* algorithm of Section 3.2 will remain biased under relative scaling. By their design, symmetric objective functions of the form (3.6) can never get rid of this bias; it will always appear in one of the two terms. Nevertheless, if there is not a large scaling, the symmetric formulation is still practical.

3.5.3 The Bootstrapping Problem

It has been previously argued that "without any other prior knowledge, the registration problem is symmetric" [Cachier and Rey, 2000]. We claim that the blurry nature of real images breaks this symmetry as soon as there is relative scaling. Yet, we don't know a priori whether to expect any relative scaling between two images, and if so, which of them ought to be downscaled. This uncertainty raises the question, which warp direction (*i.e.*, *forward* or *backward* algorithm of Section 3.2) should be employed initially to guess the scaling. The empirical evidence we gathered in the face domain suggests that the bias induced by using the non-optimal warp algorithm is not big enough to instigate a wrong decision about the direction of the scaling. In other words, we expect both algorithms to be acceptably correct in hinting at the relative scaling, and based on this initial result we could commit to the *correct* warp direction and obtain unbiased estimates.

3.5.4 Heuristics and Regularization

One way of dealing with low-resolution data is to construct a scale-space pyramid of AAMs and to model multiple resolutions at once [Liu et al., 2006]. Due to blur, higher-level (*i.e.*, lower-resolution) AAMs would have more compact appearance models and would therefore be easier to fit. However, their reconstructions would also be blurry and therefore not well-suited for hallucination. In our work we only fit high-resolution AAMs, independently of how much lower in resolution the observations were. This allowed us to hallucinate faces in high-resolution.

We have exclusively dealt with the *formulation* of the image registration problem. The two AAM fitting algorithms compared in Section 3.4 use exactly the same Gauss-Newton minimization method and parameters such as step size, number of iterations, etc. As such, our discussion remains orthogonal to practical search heuristics such as multiresolution, hierarchical and progressive methods [Anandan, 1989; Bergen et al., 1992]. We can still exploit the advantages of these: for instance, a pyramid-style fitting algorithm would increase the robustness of RAF, complementing its accuracy at the bottom level.

Finally, recall that both AAMR-SIM and RAF are Gauss-Newton gradient-descent methods that iteratively update the entire set of AAM parameters. The single and multi-person AAMs have 34 and 52 parameters respectively, all being estimated simultaneously. Beyond a certain downscaling factor, there are so few face pixels that singularities arise

while inverting the Hessians. This practically limited our scaling range to factors of 12 and 16 in the above cases. Parameter scheduling and regularization techniques would help, but such issues are beyond the scope of this thesis.

3.5.5 Related Issues in Computer Vision

A relevant discussion on symmetric vs. asymmetric formulations can be found in [Horn, 1987], where the goal was to estimate the similarity transformation between two sets of points with known correspondence. It was observed that the sensitivity of the scale estimate depended on the direction of the transformation. Consequently, an asymmetric objective function was recommended for those cases where one set of coordinates might be known with much greater precision than the other.

Earlier work on image matching with point features [Hansen and Morse, 1999; Dufournaud et al., 2000] observed that interest points were not invariant to scale. As a remedy, these points were computed for a variety of scale (*i.e.*, blur) levels, which parallels the extra blurring advocated in this work.

Previously, systematic biases in optical flow methods were shown to stem from errors in image gradient estimation [Kearney et al., 1987; Brandt, 1994; Nagel and Haag, 1998; Fermueller et al., 2001; Bride and Meer, 2001] or from the data-dependence of the noise process [Kanatani, 1996; Bride and Meer, 2001]. In contrast, we explored a potential bias arising from the resolution-limited nature of real images.

From a practical point of view, we would expect to have difficulties if the two cameras were defocused by different degrees: Since our blur compensation step estimates the amount of necessary blurring from the relative scaling factor, it would not be able to account for the total blur accurately. We plan to explore this phenomenon in future work.

3.5.6 Imposing Priors onto AAM Parameters

Recall that both the shape and appearance components of an AAM are characterized through Principle Components Analysis, which assumes that the data follows a Gaussian distribution. According to this model, a multivariate Gaussian, centered at the mean of the training data, would be the natural prior for AAM parameters. This idea was explored by [Cootes and Taylor, 2001], who reported mixed results: while the shape prior reduced the positioning error of the landmarks, it also deteriorated the appearance estimates.

A weaker variant of the Gaussian prior clamps the AAM parameters such that they remain within a prescribed hyper-cube or -ellipsoid [Stegmann, 2004, p. 145]. This imposes a uniform prior within the allowed parameter hyper-volume, and completely bans regions outside of it. We used this hard constraint on all fitting algorithms reported in this chapter, and found that it improved fitting robustness.

As part of this thesis research, we also characterized and imposed more sophisticated priors on AAM parameters, but our attempts did not lead to significant improvements. For instance, we investigated the use of temporal smoothness priors. In a separate approach, we approximately specified which nonlinear regions of the parameter space produced human-like faces, and banned those regions that did not. Overall, we found that our priors did not render the fitting problem more robust or accurate. In our experience, the data term (and its local minima) dominates the fitting process in single-hypothesis trackers, and the addition of the prior does not help regularize the search. When applied to AAMs, multi-hypothesis trackers such as [Isard and Blake, 1998] might be able to benefit from these priors.

3.6 Conclusion

This chapter demonstrated the importance of carefully crafted metrics and algorithms in meeting the challenges of resolution degradation in Face Hallucination.

The key observation is a resolution-induced asymmetry in model-to-image or image-to-image registration problems: under relative scaling, one must start with the higher-resolution image (or model) and warp it onto the lower-resolution one while incorporating a blur-formation process in the fitting criterion. If the scaling-induced bias is ignored, or the lower-resolution image is warped (and interpolated) onto the higher-resolution one, one should expect the warp estimated to be biased.

The asymmetry has tangible consequences. We demonstrated and quantified its detrimental effects in Face Hallucination using AAMs. We showed how the traditional fitting formulation overlooks the asymmetry issue, causing the fitting accuracy to degrade quickly when the observed faces are smaller than their model. We then formulated a novel algorithm that respected the asymmetry, and incorporated an explicit model of the blur into the fitting formulation. We compared this algorithm against a state-of-the-art method across a variety of resolutions and AAM complexity levels. Experimental results showed significant accuracy improvements in both shape and appearance estimates when fitting to low-resolution data. This resulted in much more detailed and accurate Face Hallucinations.

Chapter 4

Face Hallucination with a Video Model

In a video sequence, the appearance of a face does not change randomly. On the contrary, the head motion, facial expressions and speech produce a rich set of temporal dynamics. These can be as smooth as raising eyebrows, or as abrupt as eye blinks. This chapter develops a Face Hallucination framework that exploits the spatio-temporal dynamics of faces. It demonstrates that a *video-based* approach to hallucination performs better than an *image-based* one that operates frame-by-frame.

Section 4.1 underlines the shortcomings of the parametric *image*-based approach of Chapter 3 and motivates a non-parametric *video* representation. Section 4.2 introduces the main computational tool of this chapter: a statistical generative model of face videos. By treating a video as a composition of space-time patches, this model can efficiently represent and reason about the complex visual phenomena to be hallucinated. The patch-based representation is further exploited to define a data-driven prior on a 3-dimensional Markov Random Field in space and time.

Section 4.3 poses Face Hallucination as a probabilistic inference problem over the proposed graphical model and presents an algorithm for solving it. The experimental results of Section 4.4 demonstrate that temporal dynamics regularize the hallucination problem. Finally, Section 4.5 discusses the limitations and potential extensions of the approach.

4.1 Motivation

As in Chapter 3, we follow the *analysis-by-synthesis* paradigm in solving the hallucination problem. First, we briefly revisit the *image* model of Chapter 3 and unveil the key aspects of our *video* model.

4.1.1 Parametric vs. Data-Driven Models

Chapter 3 treated Face Hallucination as a parametric model-fitting problem . The approach involved recovering the (face image) model parameters that best explained the observed low-resolution data, and then using these parameters to synthesize a model instance, *i.e.*, a high-resolution face image. This turned a difficult high-resolution image estimation problem (with 100x100=10000 unknown pixel intensities) into a more manageable parameter estimation one (with only 50 shape/appearance coefficients).

The number of face pixels drops with image size. Consequently, model-fitting approaches run out of data beyond a certain resolution. In a parametric optimization process, the problem exhibits itself as a numerical one. For instance, the Hessian of the Gauss-Newton search algorithm of Section 3.3 becomes singular and cannot be inverted. To proceed, one has to introduce regularization and randomization strategies, but these are far more complex than an analytical gradient-descent.

The difficulties above raises the question as to whether parameter estimation has to be an integral part of hallucination. Would it be possible to infer high-resolution intensities *directly* from their low-resolution counterparts, without having to recover intermediate parameters? This chapter provides a positive answer. The proposed approach is data-driven, *i.e.*, it uses a training database of low- and high-resolution image pairs. The low-to-high inference problem is then tackled via look-ups among available examples. Although computationally more expensive, this approach does not suffer from the singularity issues of parameter estimation.

4.1.2 Global vs. Local Representation

In many problems, parametric models built from a small training set can generalize and successfully predict novel data. For instance, given a set of shape and appearance parameters, the AAMs of Chapter 3 can synthesize high-resolution face images for a continuous

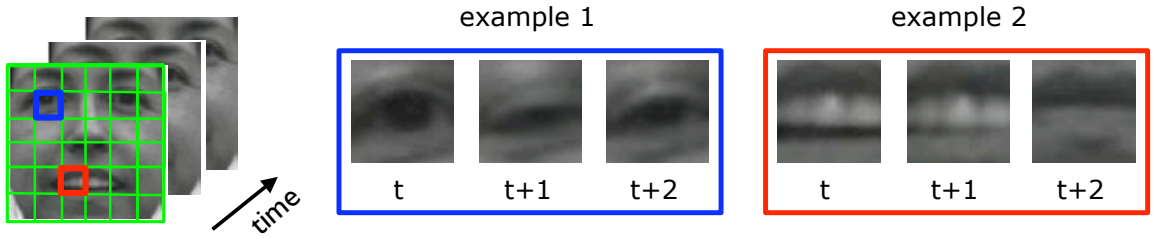

Figure 4.1: We decompose face videos into space-time patches and thereby reduce the dimensionality of the data to be modeled. The two examples shown here capture the visual phenomena of the eye-blink and the occlusion of the teeth. These patches are 16×16 pixel wide and have a temporal support of 3 frames.

range of expression and pose variation. Building a non-parametric, example-based model to capture the same variation is a daunting task: it requires collecting image samples for all possible facial expressions under each and every pose. If the space is not sampled densely, a data-driven model would overfit to the training data and not generalize [Wasserman, 2006].

Our strategy to avoid the overfitting problem is to decompose face images into a collection of patches and thereby reduce their dimensionality. Unfortunately, this destroys the global structure of faces and necessitates a co-occurrence model among the patches, which is a challenge by itself. As a trade-off between treating face patches all independently and building a full statistical co-occurrence model among them, we model a more limited interaction and consider only local couplings between neighboring patches.

4.1.3 Image vs. Video

The image-based approach of Chapter 3 ignored the temporal aspect of videos and hallucinated faces in a frame-wise fashion. Faces do not change randomly. On the contrary, the motion of the head, speech and facial expressions produce a rich set of temporal dynamics. The main goal of this chapter is to exploit these dynamics in the hallucination process. For this end, we extend the patch-based image representation into the temporal domain and compose videos out of space-time patches. As illustrated in Fig. 4.1, these patches capture the complex visual phenomena directly while still benefiting from locality (*i.e.*, reduced dimensionality) in both space and time.

The video model we propose for Face Hallucination is inspired by the key aspects of the following earlier work. First, by using a spatially varying prior as in [Baker and

Kanade, 2002], the computational requirements are kept relatively low. Furthermore, the spatial couplings in [Freeman et al., 2000] are extended to capture both spatial *and* temporal consistencies in the hallucinated videos. In contrast to [Bishop et al., 2003], we do not resort to re-seeding our high-resolution hypothesis space with earlier solutions, but instead model and deal with *temporal visual phenomena* directly.

4.2 A Graphical Model for Face Videos

We aim to integrate our domain knowledge about the videos of human faces with the physical principles of image formation. In the following, we first introduce our graphical model for the formation of low-resolution observations. For clarity, Section 4.2.1 first describes this generative model for the static image case, and Section 4.2.2 extends the latter to the temporal dimension.

4.2.1 Generative Image Model

A graphical model is a concise tool for expressing causal and statistical dependence relationships between random variables of interest. Specifically, two nodes which are not connected by a link are independent when conditioned upon their neighbors. Such conditional independencies will play a crucial role in Section 4.3, where we will articulate our probabilistic inference algorithm.

We use the graphical model illustrated in Fig. 4.2 to integrate our domain knowledge about the images (or videos) of human faces with the physical principles of image formation. Our model for low-resolution observations comprises two steps. Starting from the bottom, it prescribes to:

1. Generate a high-resolution template T to be imposed as prior on the hallucination H.

2. Blur and downsample the hallucination H to simulate the formation of the low-resolution image L.

The starting point is a high-resolution template image T, generated following a prior model about possible images in the domain. Building a generative statistical model of T that can account for all possible face images (and videos) represents a formidable challenge.

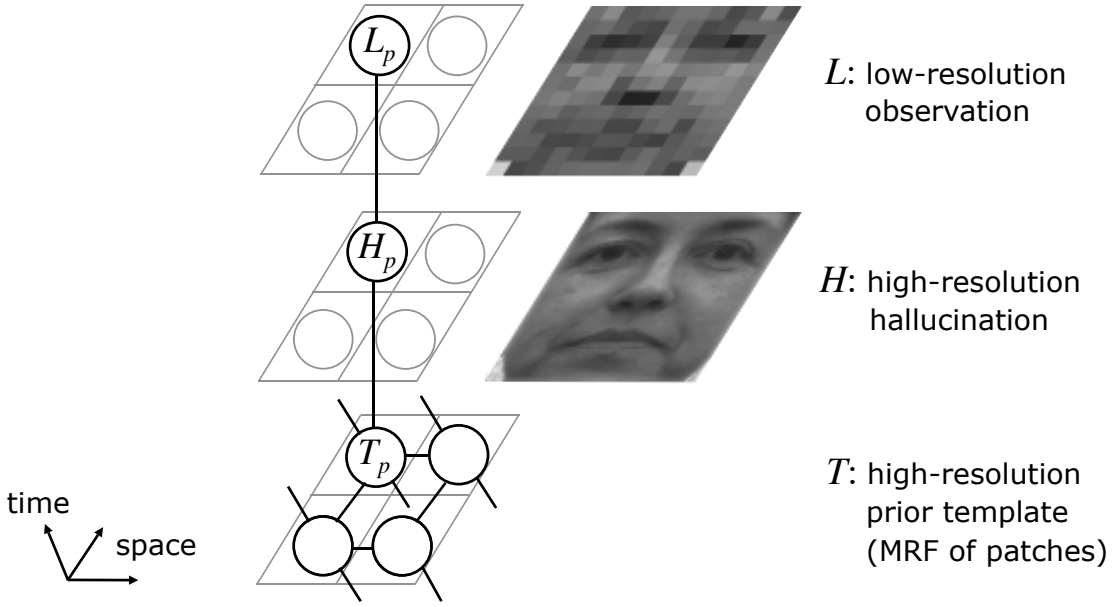

Figure 4.2: Starting from a low-resolution face L, Face Hallucination estimates a high-resolution face H. In modeling this problem, we integrate our domain knowledge about the images (or videos) of human faces with the physical principles of image formation. The nodes in this graphical model correspond to space-time patches. The prior template T composes facial expressions from a large database of training examples.

In order to circumvent this modeling problem, we take a non-parametric approach and draw samples from a large database of examples. Since capturing all possible variations of facial expressions and features requires a very large number of examples to be stored, one can adopt local models, defined over image patches, and treat them independently, as in [Baker and Kanade, 2002]. However, such a choice fails to capture those events which span multiple patches. As a computational trade-off between treating these patches all independently and building a full statistical co-occurrence model, we impose compatibility constraints only between neighboring patches (Fig. 4.3, left). In particular, we use a Markov Random Field (MRF) to model spatial interactions, allowing us to compose face template images without noticeable artifacts.

In our model, the template image T acts as a strong prior for the hallucinated image H. To simulate the formation of the low-resolution observation L, we model the blur and downsampling operations by a linear, local-averaging operator [Andrews and Hunt, 1977].

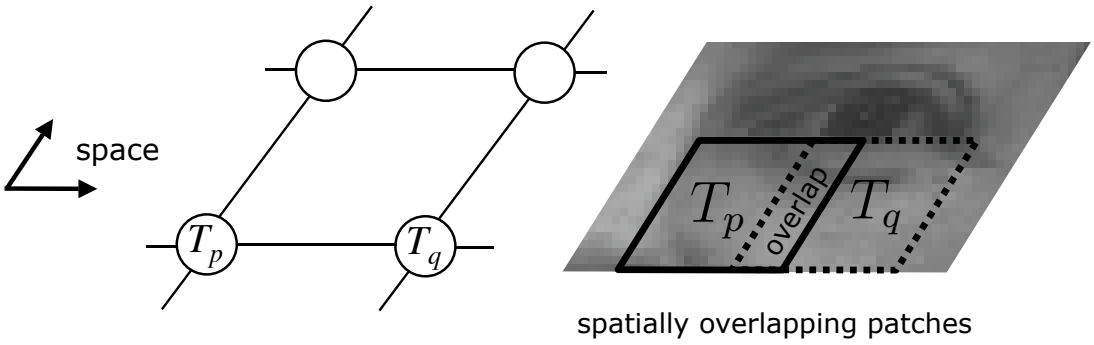

Figure 4.3: In a Markov Random Field, global properties are obtained through local interactions between neighboring sites. In this work, we encourage those template configurations where neighboring patches contain similar pixel intensities in their overlapping area. This results in face template images without noticeable patch artifacts.

4.2.2 Exploiting Time

Just as neighboring pixels in natural images tend to be highly correlated, so too are consecutive frames in video sequences. In this chapter, we exploit these temporal dependencies to further constrain the space of high-resolution solutions. As illustrated in Fig. 4.4 (a), we model couplings between consecutive frames by extending the MRF framework into the time dimension. This results in a three-dimensional network of *video patches*, defined as data structures spanning multiple consecutive frames. For instance, as shown in Fig. 4.4 (b), we can choose a temporal support of 3 frames for the nodes in T and make consecutive nodes overlap by one frame. This is equivalent to stating that the underlying video sequence is first-order Markov in time.

Our scheme gives the temporal dimension an unconventional role compared to earlier approaches to super-resolution: in the reconstruction-based resolution enhancement literature, the relative motion between frames is estimated, then eliminated through warping or optical flow. Reconstruction-based approaches are essentially two dimensional, treating time as a nuisance parameter to be compensated for.

The very small size of inputs (6×6 pixels) considered in this chapter would make the recovery of facial motions (*e.g.*, the opening and closing of the eyelids and mouth, and the appearance of pupils and teeth) particularly difficult. Avoiding this motion estimation problem, we take advantage of the richer local signature that the combination of space *and* time provides. Our model captures and reasons about occlusions, appearance of new

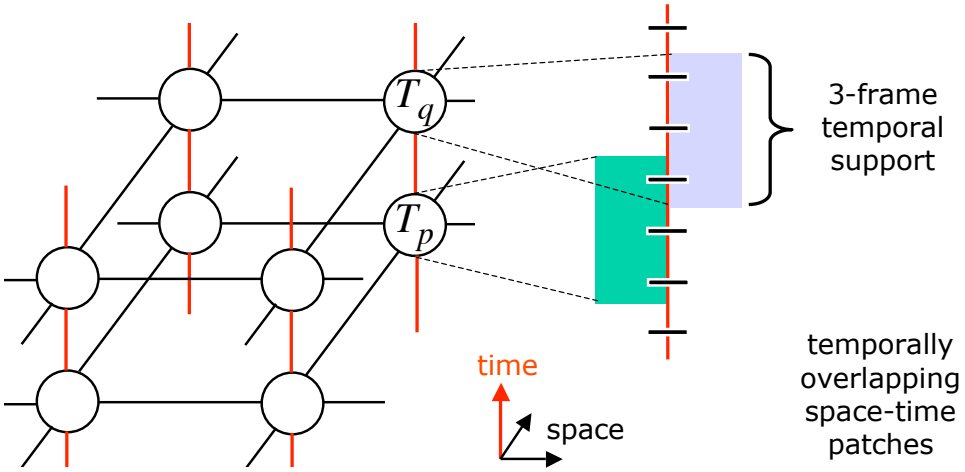

Figure 4.4: (a) We model couplings between consecutive frames by extending the MRF framework into the time dimension. (b) The space-time patches that constitute the nodes of the MRF span multiple consecutive frames. To evaluate compatibilities between neighboring nodes, we compute pixel differences over whole patches of overlapping video frames.

structures, and non-diffeomorphic deformations naturally, in terms of interacting chunks of high-resolution template videos. As such, we exploit the complex visual phenomena as *spatio-temporal signatures*.

Recently, psychological studies have shown that motion helps human face perception under non-optimal (*e.g.*, viewing blurred or negative images) conditions [Bruce and Valentine, 1988; Pike et al., 1997; Lander et al., 2001]. The "supplemental information hypothesis" [O'Toole et al., 2002] suggests that humans learn motions of familiar faces as dynamic signatures and exploit them for recognition [Roark et al., 2006]. As the experimental results will attest, our video hallucination framework mimics this capability.

4.3 Hallucination as a Probabilistic Inference Problem

Now we will formulate the problem of Face Hallucination in videos. Using our graphical model, we pose the problem as one of finding the Maximum A Posteriori (MAP) high-resolution video H_{MAP} given the low-resolution video L:

$$H_{MAP} \triangleq \arg\max_{H} P(H \mid L).$$

To express the MAP estimate in terms of known quantities, we first marginalize over the unknown template video T:

$$P(H \mid L) = \sum_T P(H, T \mid L).$$

The discrete summation above is due to the sample-based representation for T. By applying the chain rule[1], the posterior can be expressed as

$$\sum_T P(H \mid T, L)\ P(T \mid L). \tag{4.1}$$

At this point, we would like to tease out a premise that underlies the entire enterprise of super-resolution. The very assumption that we can succeed at the task of super-resolution (*i.e.*, estimate H uniquely and to arbitrary resolution) implies that the underlying distribution $P(T \mid L)$ is peaked around the true high-resolution solution. As an approximation, we assume that this posterior is a delta-function at the true configuration, which we estimate using the input.

Unique Template Assumption

Assume that the posterior $P(T \mid L)$ is highly concentrated around some $T^* = T^*(L)$. In other words,

$$P(T \mid L) \approx \delta(T - T^*). \tag{4.2}$$

Deferring the computation of T^* until section 4.3.1, we substitute (4.2) into (4.1) so that $P(H \mid L)$ is approximately

$$P(H \mid T^*, L).$$

Using Bayes rule, the posterior can be written as

$$\frac{P(L \mid T^*, H)\ P(H \mid T^*)}{P(L \mid T^*)}.$$

[1]The chain rule asserts that $P(X, Y) = P(X \mid Y)P(Y)$.

The graphical model of Fig. 4.2 entails the conditional independence[2] $P(L \mid H, T) = P(L \mid H)$. Capturing the denominator by a constant C, we rewrite the posterior as

$$C\ P(L \mid H)\ P(H \mid T^*). \tag{4.3}$$

Taking the logarithm of (4.3), H_{MAP} approximately maximizes

$$\log P(L \mid H) + \log P(H \mid T^*). \tag{4.4}$$

Note the trade-off in the computation of H_{MAP}. The first term encourages those H that increase the likelihood of the *reconstructed* observation L, while the second term imposes a data-dependent prior T^* on H.

4.3.1 Maximization of the Posterior $P(T \mid L)$

Now we describe our method for computing the peak template T^* in (4.2) by estimating the maximum of $P(T \mid L)$.

Using Bayes rule, we first rewrite this posterior in terms of likelihood and prior terms. Observing that nodes in L are conditionally independent given the high-resolution template T, we obtain a factorized likelihood term

$$P(T \mid L) \propto P(L \mid T)\ P(T) = \prod_{p=1}^{N^2} P(L_p \mid T_p)\ P(T). \tag{4.5}$$

Unfortunately, in the case of extremely blurred images, the likelihood term $P(L_p \mid T_p)$ is too weak; many templates match with a given L_p. One remedy to this problem is based on the observation that there are spatial dependencies in the observed data. Thus, by pooling contextual information about L_p into a local feature vector, one can make the likelihood term more descriptive. The downside of such an extension is that the factorized form of (4.5) will no longer be valid. In section 4.3.3, we will present the details of such a feature vector, and expose our assumptions for achieving a computationally tractable algorithm.

[2]Two nodes which are not connected by a link are independent when conditioned upon their neighbors.

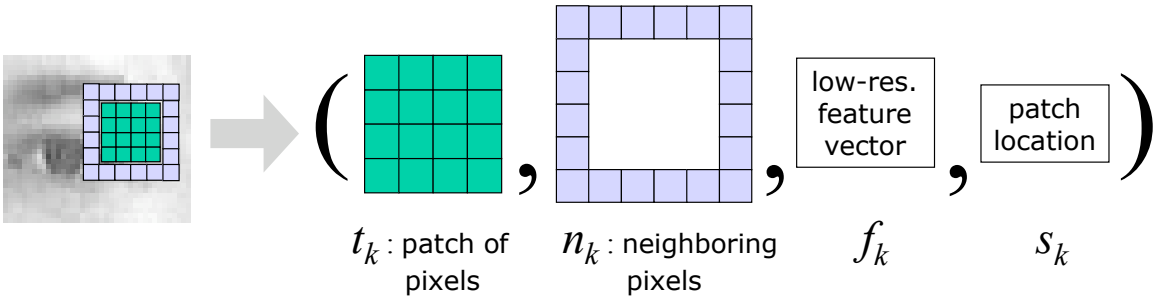

Figure 4.5: Each database entry contains an image patch, the neighboring pixels (for enforcing consistency), a feature vector (for matching to the low-resolution image), and its location (for supporting non-homogeneous spatial statistics). This structure is repeated for all frames within the temporal support of the space-time patch considered.

4.3.2 The Template Prior

We restrict the space of possible T's to a domain-specific collection of example templates. For this end, a database is generated from training data by artificially downsampling high-resolution images and computing their low-resolution feature images. As shown in Fig. 4.5, we store these examples patch-wise, in that each record is a quadruple, (t_k, n_k, f_k, s_k), containing high-resolution template patch pixels t_k, a thin strip n_k of surrounding pixels, the feature vector f_k computed at the corresponding low-resolution pixel, and the location s_k of the template.

The MRF model assigns a probability to each template patch configuration T. According to Hammersley-Clifford theorem, $P(T)$ is a product $\prod_{T_p,T_q} \phi(T_p, T_q)$ of compatibility functions $\phi(T_p, T_q)$ over all pairs of neighboring nodes. We define ϕ using similarity between pixel values in the overlapping areas of example patches. For spatially neighboring patches, it is

$$\phi_s(T_p = t_k, T_q = t_l) \propto exp\Big(- \sum_{u,v \in overlap} \big(t_k(u) - n_l(v)\big)^2 - \sum_{u,v \in overlap} \big(n_k(u) - t_l(v)\big)^2\Big),$$

whereas for temporally neighboring patches it is:

$$\phi_t(T_p = t_k, T_q = t_l) \propto exp\Big(- \sum_{u,v \in overlap} \big(t_k(u) - t_l(v)\big)^2\Big).$$

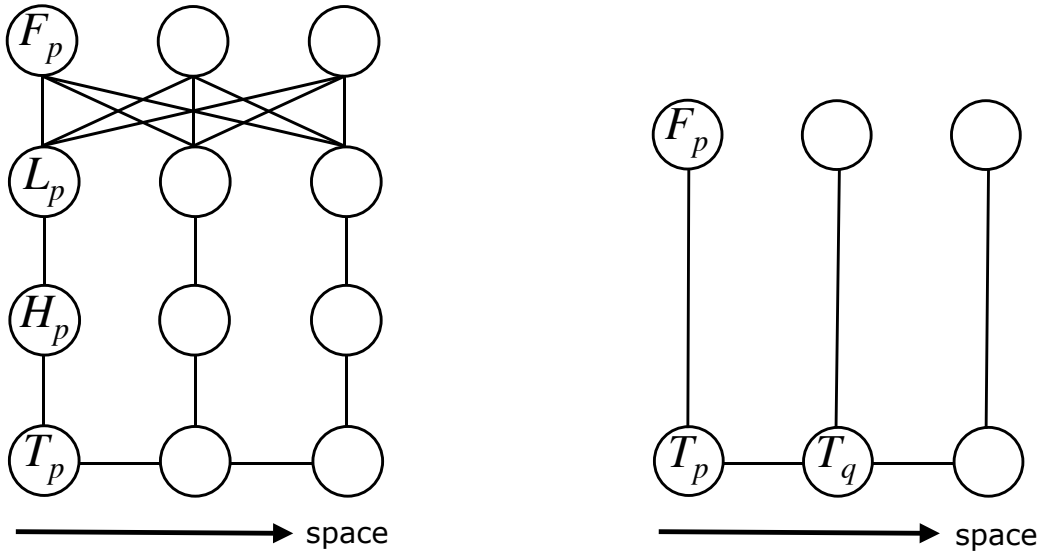

Figure 4.6: Interactions involved in determining the peak template T^*. For illustration purposes, a 1-dimensional version of the model in Fig. 4.2 is shown (left). After applying the factorization assumption, the resulting graph structure (right) becomes tractable enough to apply inference methods such as ICM.

4.3.3 The Feature Vector

To render the likelihood term more descriptive, we use a multiscale feature vector derived from the low resolution observation L. Following [Baker and Kanade, 2002], we adopt the *parent vector* as our feature F_p, which stacks together local intensity, gradient and Laplacian image values at multiple scales [DeBonet and Viola, 1998]. Fig. 4.6 (left) shows a 1-dimensional version of Fig. 4.2, with the feature vector nodes added.

Factorization Assumption

Observe that we have two random fields, F and T that are coupled through the low-resolution image formation model. For computational tractability, we invoke the pseudo-likelihood approximation [Li, 2001] to assume that $P(F \mid T)$ factorizes across feature image pixels:

$$P(F \mid T) \approx \prod_{p=1}^{N^2} P(F_p \mid T). \tag{4.6}$$

Correspondingly, the graphical model of Fig. 4.6 (left) is simplified to Fig. 4.6 (right).

The likelihood $P(F_p = f_p \mid T_p = t_k)$ will be defined using the similarity between the feature vectors $f_p(L)$ and f_k, where k is an index to database entries. For a spatially-varying (*i.e.*, inhomogeneous) prior for T_p, we consider a similarity of the form

$$P(F_p = f_p \mid T_p = t_k) \propto \begin{cases} exp\big(-\|f_p(L) - f_k\|^2\big) & \text{if } s_k = p, \\ 0 & \text{otherwise.} \end{cases}$$

Using the factorized form (4.6), T^* is approximately

$$\arg\max_{T} \prod_{p=1}^{N^2} P(F_p \mid T_p) \prod_{(p,q)} \phi(T_p, T_q). \tag{4.7}$$

Finding the Peak Template T^*

In Alg. 1, we adopt a greedy approach commonly taken in the field of Bayesian image estimation: the Iterated Conditional Modes (ICM) method [Besag, 1986]. This algorithm takes advantage of the Markov structure, and maximizes local conditional probabilities sequentially.

4.3.4 Hallucinating Face Videos

Now we introduce the details of the likelihood models of image formation and observation. We show that, once we have computed T^*, video hallucination (computing H_{MAP}) only requires a quadratic minimization.

Likelihood Models

We model the hallucination H as a noisy version of the template T:

$$H = T + \eta_H.$$

The deviation from the template follows a pixel-wise independent additive Gaussian noise $\eta_H \sim N(0,\ diag(\sigma_H))$:

$$P(H \mid T) = \prod_{h=1}^{M^2N^2} \frac{1}{\sigma_H\sqrt{2\pi}}\, exp\Big(-\frac{\big(T(h) - H(h)\big)^2}{2\ {\sigma_H}^2}\Big).$$

input : observation L
output: peak template T^*

compute the multiscale feature $F = F(L)$

/ initialize T^* with local Maximum Likelihood estimates */*
for *all video patches* p **do**

$$T_p^* \leftarrow \arg\max_{t_k} P(F_p = f_p \mid T_p = t_k)$$

end

/ choose a video patch, and update it using its neighbors */*
repeat

pick a random location p

$$T_p^* \leftarrow \arg\max_{t_k} P(F_p \mid T_p^* = t_k) \prod_{q \in \mathcal{N}(p)} \phi(T_p^* = t_k, T_q^*)$$

until T^* *converges* ;

Algorithm 1: We search for the peak template T^* with ICM.

After the high-resolution hallucination H is blurred and downsampled, additive sensor noise is considered, resulting in our model for the low-resolution observation L:

$$L = AH + \eta_L,$$

where the matrix A is a local averaging operator with N^2 rows and M^2N^2 columns. We assume a pixel-wise independent noise model for L:

$$P(L \mid H) = \prod_{l=1}^{N^2} \frac{1}{\sigma_L \sqrt{2\pi}}\, exp\Big(-\frac{\big(L(l) - (AH)(l)\big)^2}{2\,{\sigma_L}^2}\Big).$$

Computing H_{MAP}

Given the likelihood models above, we can evaluate (4.4): H_{MAP} minimizes

$$\|L - AH\|^2 \; + \; \frac{\sigma_L^2}{\sigma_H^2}\, \|T^* - H\|^2. \tag{4.8}$$

Individual terms above have intuitive interpretations: the first term encourages those high-resolution videos H that can *reconstruct* the observation L. At the same time, the second term states that H cannot be too different from T^*. Consequently, H_{MAP} will be a trade-off between the inferred template T^* and the observation. Finally, we observe that (4.8) is quadratic in the unknown H, and employ a gradient descent scheme for this minimization.

4.4 Experiments

4.4.1 Setup

Training

To build our face template prior, we video-taped a story-telling subject for 10 minutes. The recording took place indoors, under fixed lighting conditions. In a post-processing step, we used a translational tracker to stabilize the face position throughout the video. This yielded a training set with approximately 10,000 high-resolution examples, wherein the face covered a `96x96` pixel area.

Our training database was designed for a 16-fold increase in resolution; it paired individual low-resolution pixels (and their feature vectors) with `16x16` pixel-wide high-resolution template patches. The feature vector was a 15-dimensional measurement (grayscale intensity, horizontal/vertical/temporal derivatives, and the Laplacian, computed over 3 scales) for each time instant within the temporal support of the space-time patch. As depicted in Fig. 4.5, the neighboring pixels came from a 2-pixel wide frame that surrounded each patch.

Testing

As test data, we used approximately 3 second long video sequences of the same subject, under the same illumination condition as the training data. The translational motion of the face was eliminated as above. These high-resolution videos were considered to be the "ground truth" data, and their artificially downsampled low-resolution versions were used for testing. To account for tracking uncertainty in low-resolution, we first added independent and identically-distributed (i.i.d.) translational jitter (zero-mean Gaussian, $\sigma = 1$ high-resolution pixel) to test videos before blurring and downsampling them at a resolution of `6x6` pixels. Finally, we added i.i.d. noise (zero-mean Gaussian, $\sigma = 1$ grayscale level) to low-resolution pixel intensities to simulate sensor characteristics.

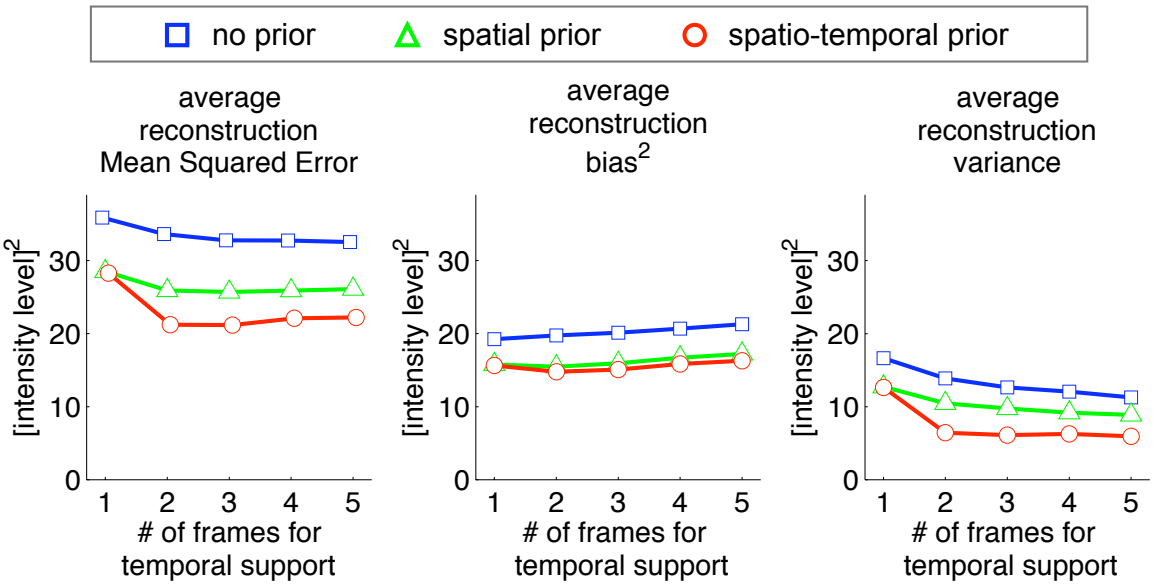

Figure 4.7: We generated and applied independent jitter and additive noise to the same low-resolution input video for 30 experiments, and compared the hallucinations against the ground truth. Enforcing spatio-temporal couplings reduces the MSE (left), primarily by reducing the variance and enhancing the stability of hallucinated videos (right). However, as temporal couplings become stronger, bias magnitudes also increase (middle).

4.4.2 Quantitative Evaluation

To analyze the role of spatial and temporal couplings, we systematically turned the spatial and temporal interactions on and off, and varied the temporal support of space-time patches from one to five frames. At each test configuration we ran 30 hallucination experiments wherein we jittered, blurred and downsampled, and finally added noise to the same 50-frame test sequence.

Signal-Level Reconstruction Error

We first report the statistics of the Mean-Squared-Error (MSE) between the hallucinations and the ground truth. In Fig. 4.7, we plot three curves corresponding to the no-prior, spatial-only prior, and spatio-temporal prior cases as a function of temporal support in space-time patches. In a horizontal arrangement, we decompose the MSE (left) into its squared-bias (middle) and variance (right) components [Casella and Berger, 1990]. For summary, we average these metrics spatially and temporally over the entire video sequence.

Fig. 4.7 shows that imposing priors onto the hallucination problem significantly reduces the MSE. Modeling the spatial couplings between face patches reduces the error, and extending such interactions into the temporal domain pushes the error levels even lower. The improvement can be visually confirmed in sample hallucination videos[3].

The Bias-Variance decomposition of the MSE reveals that temporal models dramatically reduce the variance of hallucinations, but not so much their bias magnitude. However, as the temporal support (hence the dimensionality) of the representation gets larger, the bias slightly increases (Fig. 4.7, middle). Since the size of our training set is fixed, such overfitting effects are to be expected.

Although the MSE is a physically meaningful metric for signal reconstruction, it does not necessarily reflect perceived visual quality by humans [Girod, 1993]. The design of objective image/video quality metrics that mimic the sensitivities of the the Human Visual System (HVS) is an active research area [Winkler, 1999; Pappas and Safranek, 2000]. In the following, we present additional quantitative evaluations based on perceptual metrics.

Structural Similarity Index

Humans are highly adapted to extract the structural information of a scene despite illumination effects. This observation motivated the Structural Similarity (SSIM) Index that aims to primarily measure the structural changes between a reference image and its distorted version [Wang et al., 2004]. The mean SSIM is a scalar that summarizes the effects of local luminance, contrast and correlation distortions between two images, and has been shown to be a competitive predictor of perceived quality by humans.

We computed the mean SSIM[4] quality measure between our hallucinations and the ground truth videos. In Fig. 4.8 (left), we plot three SSIM curves as we did for the MSE values in Fig. 4.7. Note that a larger SSIM score means a higher similarity to the ground truth, hence better hallucination quality. We observe that the relative performance among experimental settings are in agreement between the MSE and SSIM measures: capturing the spatial interdependencies between face patches reduces the error, and extending the model into the temporal domain pushes the error levels even lower.

[3] Videos are available at http://www.cs.cmu.edu/~dedeoglu/thesis

[4] Code from http://www.cns.nyu.edu/~lcv/ssim/ was used with default parameters.

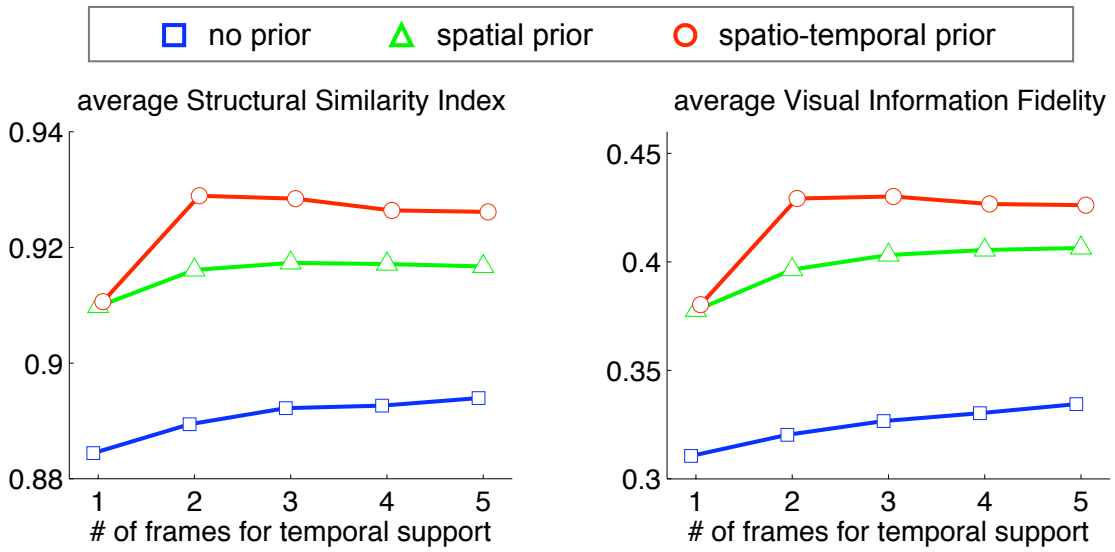

Figure 4.8: In addition to the MSE, we used objective image quality metrics to measure the perceptual similarity between the hallucinated and the ground truth videos. Both the Structural Similarity (left) and the Visual Information Fidelity (right) measures agree with the MSE results of Fig. 4.7: spatial and spatio-temporal priors improve hallucinations.

Visual Information Fidelity

The Visual Information Fidelity (VIF) is an information-theoretic similarity measure that makes use of natural scene statistics [Sheikh and Bovik, 2006]. If the information content of an image is defined as the mutual information between the input and output of the HVS channel, then the VIF measure between two images is their relative image information. In Fig. 4.8 (right), we plot the mean VIF[5] similarity curves between the hallucinated and the ground truth videos. Note the similarity between the SSIM and VIF curves.

Temporal Derivatives

The MSE, SSIM and VIF average over the temporal dimension of videos. In our subjective experience, we found temporal priors to be very effective in reducing disturbing flicker artifacts in hallucinations. To capture this effect, we measured frame-to-frame differences between consecutive time instants (*i.e.*, temporal derivatives) in hallucinated videos, and investigated how well they matched those of the ground truth. Fig. 4.9 shows the temporal

[5]Code from http://live.ece.utexas.edu/research/Quality/VIF.htm was used with default parameters.

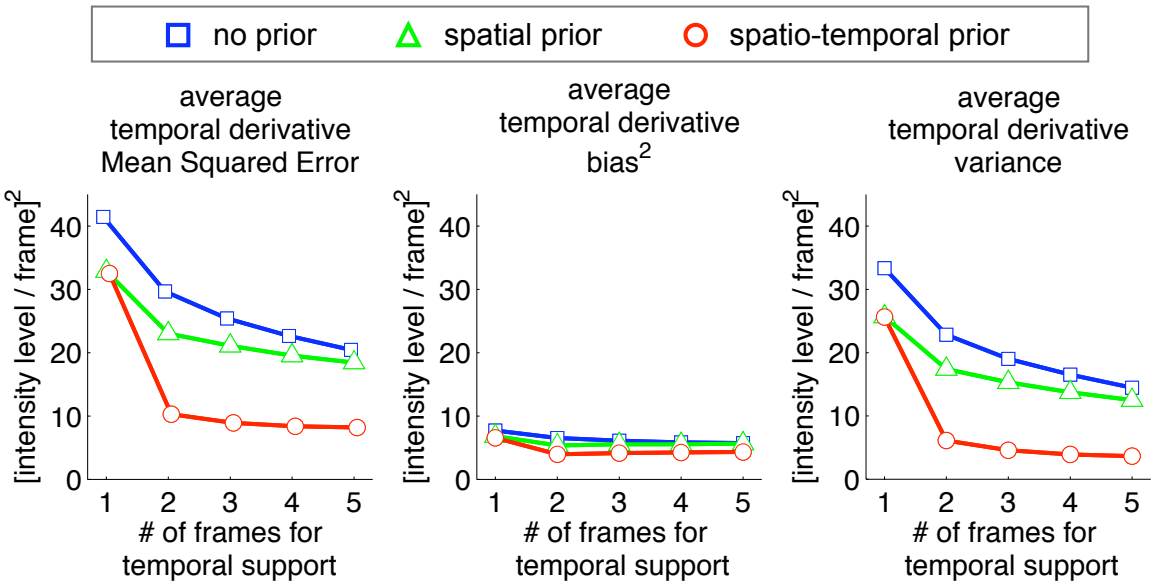

Figure 4.9: To analyze the observed reduction in video flicker artifacts, we compared the temporal derivatives of our hallucinations with those of the ground truth. Temporal couplings dramatically reduce mismatches in temporal derivatives of hallucinations.

derivative MSE curves, organized in the same fashion as the reconstruction MSE curves of Fig. 4.7. In agreement with our perceptual assessment, we observe a reduction in error magnitudes due to spatial couplings, but significantly more so due to temporal ones.

In Fig. 4.10, we plot the temporal derivative MSE as a function of time for the case of 2-frame temporal support. Note that the errors for the no-prior and spatial-only-prior cases are consistently higher compared to the spatio-temporal-prior case. The peaks observed around frames 4 and 13 are due to a smile and blinking eyes, whose exact timing is challenging to replicate in hallucination.

4.4.3 Qualitative Results

In Fig. 4.11, we visually compare the hallucination results for selected test frames. The first column shows the `6x6` pixel input, whereas the last column shows the underlying `96x96` pixel ground truth. The columns in between compare the hallucinations among three different configurations of the model: no prior, spatial-only prior and spatio-temporal prior, all using 2 frames of temporal support.

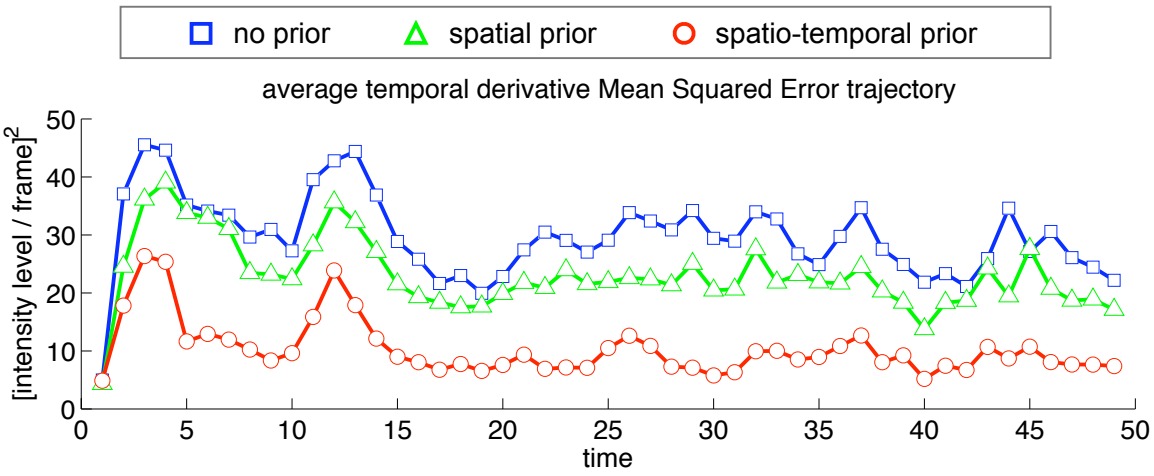

Figure 4.10: Plotting the temporal derivative MSE as a function of time can be revealing: the peaks observed around frames 4 and 13 are due to a smile and blinking eyes, whose exact timing is challenging to replicate in hallucination.

Spatial Interaction

The second column of Fig. 4.11 shows hallucinations with no interaction among the template patches. In other words, each patch in each frame is hallucinated independently using the local Maximum Likelihood template estimate. We observe that the results display many blocking artifacts and extraneous edges.

In the third column of Fig. 4.11, we enforce the interactions in space but ignore those in time. In other words, we estimate each frame independently and hallucinate *frame-wise*. Note that many of the blocking artifacts have disappeared, but unfortunately, hallucinations now contain some incorrect estimates of the underlying facial expressions (*e.g.*, closed vs. open eyelid and mouth).

Spatio-Temporal Interaction

The fourth column of Fig. 4.11 shows hallucinations with full spatio-temporal couplings. Note that facial expressions are recovered more accurately when temporal interactions are allowed (compare the opening of eyelid and mouth with spatial-only hallucinations).

As *static images*, the results in Fig. 4.11 already exhibit considerable improvements due to both spatial-only and spatio-temporal modeling of the problem. We find our results as *video sequences* to be even more compelling: frame-to-frame transitions that are not

directly observable in static images can have perceptually detrimental effects when seen as a time sequence. We observe that such flicker artifacts, amply present in frame-wise hallucinations, vanish by a large extent when temporal couplings are taken into account (*i.e.*, when two or more frames of temporal support are used). These observations confirm that time plays a crucial role as a regularizer in our inference[6].

4.5 Discussion

4.5.1 The Global Tracking Assumption

The AAM-based hallucination approach of Chapter 3 tracked low-resolution faces by estimating a similarity transformation and non-rigid deformation parameters. However, below a certain resolution, parametric tracking suffers from numerical singularities and becomes impractical. For instance, one cannot fit a 50+ parameter AAM to a `6x6=36` pixel image. One option is to solve for a subset of parameters such as translation. While this approach might be able to track the face, hallucination would not be possible unless the full set of face parameters are estimated accurately.

Tracking in low-resolution can be a challenge. Yet, the difficulties of tracking do not have to hinder hallucination as it does for an AAM. In this chapter, our strategy was to decompose the problem into its tracking and resolution enhancement components: we assumed that the low-resolution face was roughly tracked (*i.e.*, "boxed" in the `6x6` input window) and focused on the hallucination problem. We treated the tracking errors within this window as noise, and relied on the smoothing effect of the MRF to overcome its effects.

4.5.2 Estimating the Local Jitter Motion

A promising direction for development is to refine the tracking of low-resolution faces. This could be achieved by augmenting the generative model of Fig. 4.2 with a *jitter motion* variable that would geometrically perturb the (high-resolution) template images before they get blurred and downsampled. As such, the inference algorithm would be jointly solving for three variables: the template, illumination mismatch, and jitter motion. While computationally more expensive, this extension has the potential to reduce the sub-pixel tracking

[6] Videos are available at http://www.cs.cmu.edu/~dedeoglu/thesis

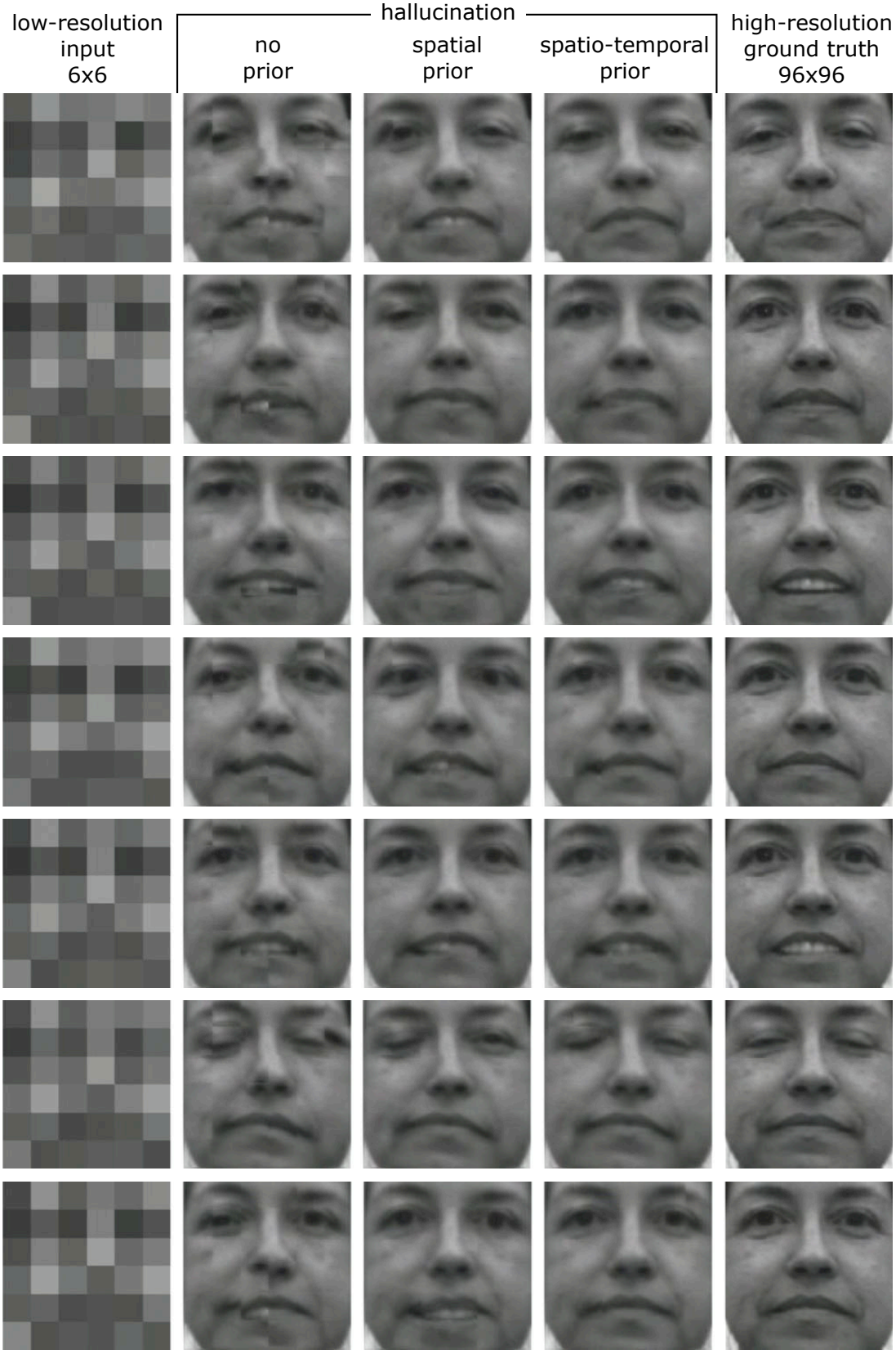

Figure 4.11: Time plays a regularizing role in video hallucination (see Section 4.4.3).

Figure 4.12: For a realistic representation of the unknown background problem, we collected our training and testing videos against a dynamically varying background. We videotaped the subject in a seminar room where a movie was projected behind the subject.

errors that are currently ignored as noise. Eliminating this noise would result in an overall improvement in hallucination.

4.5.3 Background Effects

One of the disadvantages of assuming a fixed-size face window is that one cannot reason precisely about the background: if the `6x6` window is not fully occupied by a face, some edge pixels will contain unmodeled hair, ear and background scene intensities. Indeed, this "background contamination" gets only worse with multiscale features whose span reach even farther away from the face. Note that, if the same background is used for both training and testing, this problem might never be noticed.

In our video model, we did not attempt to reject low-resolution pixels (or their associated features) on the grounds of background effects. Instead, we regarded the variability of the background as a nuisance, and dealt with it as an additional source of noise. To make our training and testing conditions realistic, we collected all our data against a dynamic background: we video-taped the subject in a seminar room where we projected a movie onto a screen behind the subject. Fig. 4.12 shows selected full-frames from our data set.

4.5.4 Exploring the Posterior $P(T \mid L)$

Under our "unique template" assumption, we only sought for the peak of the posterior distribution $P(T \mid L)$. An interesting area for future research would be to explore this posterior through Monte-Carlo Markov Chain sampling techniques. This might reveal, at the price of increased computation, interesting properties of the hallucination problem at hand. For instance, we might find out that the posterior is multimodal, indicating an ambiguity in our inference. In addition, the mean or the mode of the posterior might prove to be more robust solutions than the peak.

4.5.5 Hallucinating Template T^* vs. H_{MAP}

The video hallucination algorithm prescribes first finding the peak template T^* (and the associated illumination mismatch I^*), and then solving the quadratic minimization problem of (5.2) to compute the hallucination H_{MAP}. One might wonder how necessary the second step is: could one directly hallucinate the peak template T^* ?

First, observe that the generative template model may not be general or precise enough to synthesize some of the observed faces. When the template cannot account for the low-resolution observation exactly, H^* settles for a trade-off. As such, H effectively acts as a smoother that masks the imperfections in T.

For comparison, we quantified the reconstruction errors incurred by the peak template T^* and hallucination H_{MAP}. The curves shown in Fig. 4.13 were obtained through the experiments detailed in Section 4.4.2, where both spatial and temporal couplings were used. The MSE measurements (left) indicate a slight advantage (*i.e.*, smaller error) in favor of H_{MAP}. Decomposing the MSE into its squared bias and variance components reveals that H_{MAP} has indeed a larger bias magnitude compared to T^* (middle). In other words, *on average*, hallucinating T^* actually gives a more accurate reconstruction. Nevertheless, the variance of T^* is higher by an even wider margin (right). These results confirm smoothing effect of H_{MAP} over T^*, which is a desirable property.

Finally, recall that the template is a data-driven, appearance-based face model. While it can compose a variety of faces by combining examples it has seen before, it nevertheless cannot generalize to novel images (or videos) below the granularity of its patches. In contrast, the hallucination variable H can benefit from additional priors such as smoothness, gradient distributions, or even simulated lighting effects. H could also accumulate evidence about the underlying scene as done in conventional, reconstruction-based enhancement.

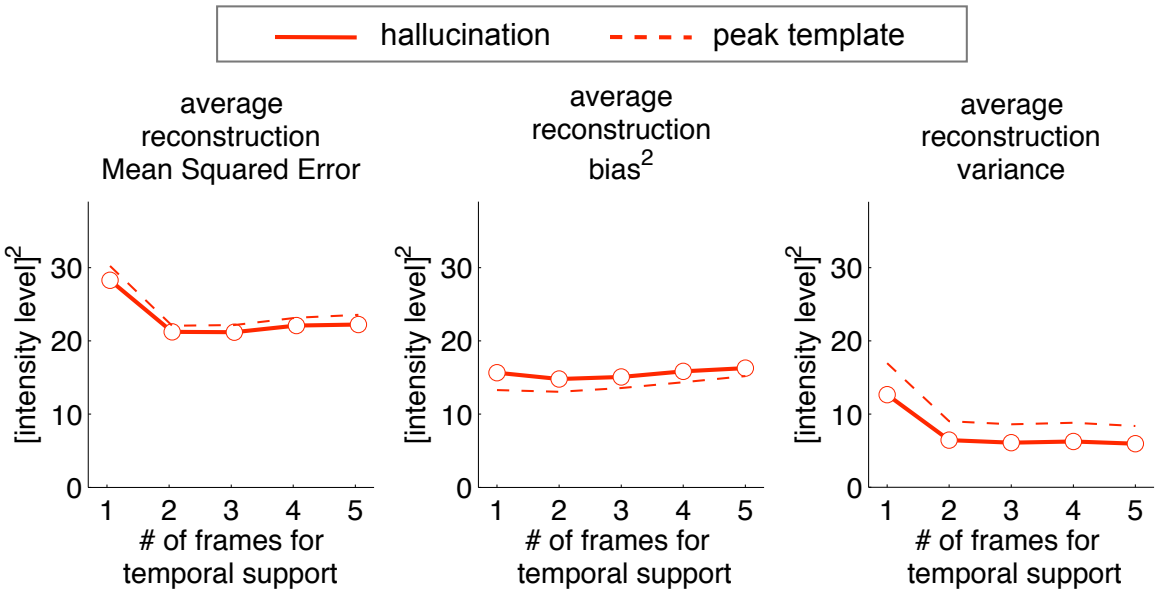

Figure 4.13: We compared the hallucination errors incurred by the peak template T^* and the hallucination H_{MAP}. The latter exhibits slightly lower error levels (left), solely due to its smaller variance (right). These results confirm the smoothing effect of H_{MAP} over T^*.

4.5.6 Sensitivity to Point Spread Function

As an inverse problem, resolution enhancement is highly sensitive to image formation models such as the point spread function (PSF) of the camera [Stark, 1987; Banham and Katsaggelos, 1997]. In Section 4.4, the template training database paired high-resolution space-time patches with their 16-fold resolution degraded versions. In both training and testing, low-resolution data was simulated with a pillbox PSF that was `16x16` pixels wide. In other words, perfect knowledge of the PSF was assumed.

How much would the hallucination performance degrade with an imperfect PSF? To find out, we ran experiments wherein we kept the training PSF intact but modified the PSF of the test data. As in Section 4.4.2, we used a 50-frame test sequence and ran 30 hallucination experiments under translational jitter and additive noise conditions. We hallucinated using the best-known model configuration: space-time patches had a temporal support of 2-frames and the full spatio-temporal interactions were enabled during the inference.

In Fig. 4.14, we plot the average reconstruction MSE as a function of the PSF width, taking integer values from 14 to 18 pixels. As expected, the MSE is lowest when the PSF is characterized accurately as 16 pixels wide. With a PSF mismatch, the reconstruction per-

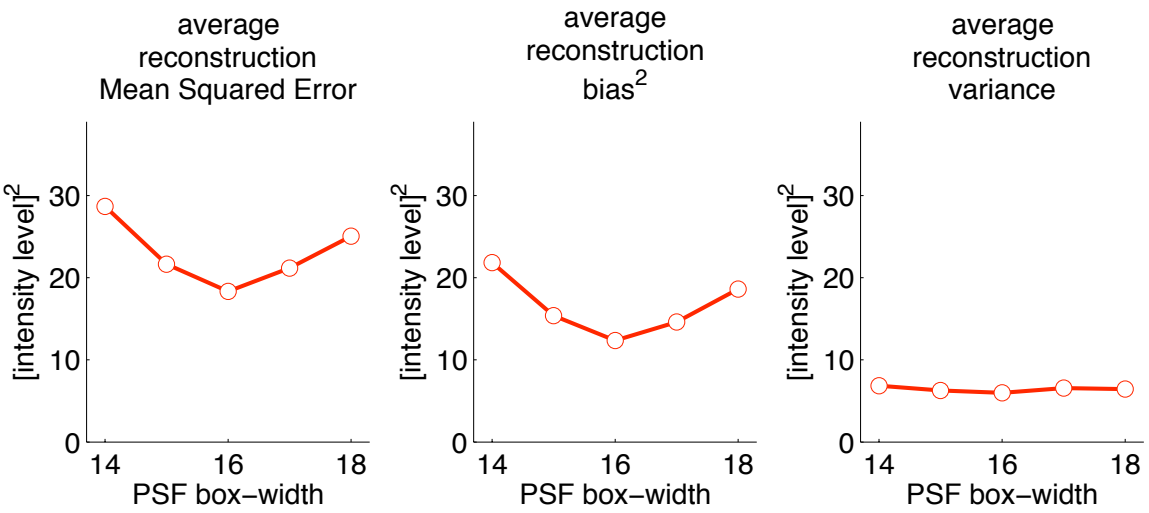

Figure 4.14: To find out how the hallucination performance degrades under an imperfect PSF, we ran experiments wherein we modified the PSF of the test data. We plot the average reconstruction MSE as a function of the PSF-width, for integer values from 14 to 18 pixels. The MSE is lowest when the PSF is characterized accurately as 16 pixels wide. Under a mismatch, the reconstruction performance degrades quickly.

formance degrades quickly: when the test PSF is 14 pixels wide, the relative degradation is comparable to the case where we had ignored the temporal couplings between the patches. The curve of the bias magnitude confirms that the underlying problem is one of inaccurate modeling.

4.5.7 Face-Specific Design

As mentioned in Section 4.1, our graphical model and its specialization to human faces were inspired by the earlier works of [Freeman et al., 2000] and [Baker and Kanade, 2002]. We now revisit some of these design choices and propose alternatives that may further exploit the human face domain.

Graphical Model Topology

Currently, our MRF is defined on a regular 3-dimensional lattice, and clique potentials of order three and higher and assumed to be zero. The connectivity of each node in the model is limited to its 6 neighbors: 4 spatial and 2 temporal. Although theoretically this structure

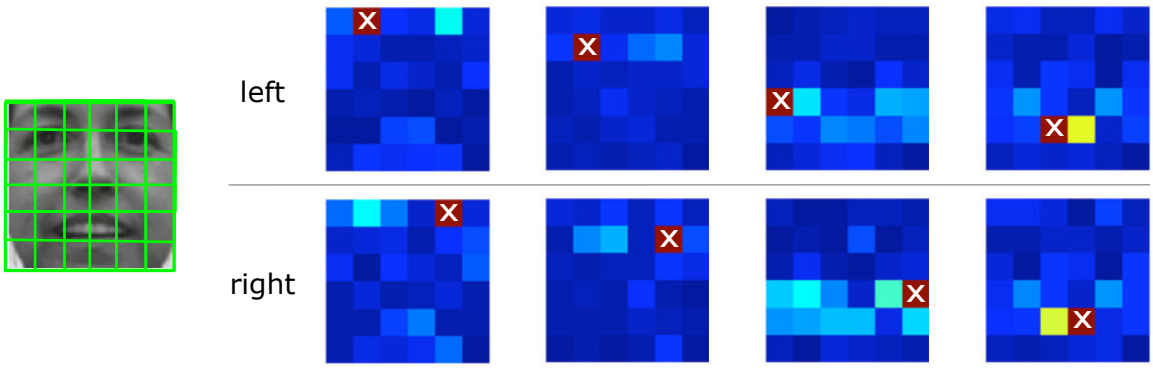

Figure 4.15: We show intensity-mapped estimates of the mutual information between a selected low-resolution pixel (marked with a $\times$) and all others. While there is some degree of mutual dependency between immediately neighboring pixels, much stronger dependencies usually exist between the left and right half pixels of faces.

can already express global statistical properties, one may expect a different connectivity to be better suited to this particular domain. Patches could be defined according to facial features, *e.g.*, as to cover the eyebrows and the mouth completely instead of breaking them into fixed-size patches. Furthermore, the underlying symmetry of faces could be used explicitly, by connecting left-right side pixels and imposing additional interaction constraints.

Feature Selection for Contextual Information

In Section 4.3.3, we adopted the multiscale "parent" feature vector of [DeBonet and Viola, 1998] in an attempt to pool contextual information about a low-resolution pixel. However, the face domain exhibits a lot more structure and regularity than generic textures, for which this feature vector was originally designed. One could expect features derived from the face domain to be more powerful in pooling relevant information. One commonly used measure of the relevancy between variables is mutual information [Cover and Thomas, 1991]. To give an intuition on this, as done by [Schneiderman and Kanade, 2004], we measured and displayed mutual information between low-resolution pixels in Fig. 4.15.

We observe that, while there is usually some degree of mutual dependency between immediately neighboring pixels, much stronger dependencies usually exist between the left and right half pixels of faces. As an extension, one could develop a feature selection framework for designing more effective contextual pooling mechanisms.

4.6 Conclusion

This chapter demonstrated that Face Hallucination can benefit from the spatio-temporal dynamics of faces. To investigate the role of time, we first devised a novel generative model of face videos. This model treated a video as a composition of space-time patches and encoded visual phenomena in a non-parametric, example-based fashion. The patch-based representation was also used to define a prior in both space and time.

We ran extensive hallucination experiments and quantified the effect of spatial and temporal models on the hallucination performance. Our results highlighted the importance of a video's temporal dimension in hallucinating facial expressions correctly.

Chapter 5

Accounting for Changes in Illumination

Images of a face under different lighting conditions exhibit large intensity variations, which pose a serious modeling challenge [Adini et al., 1997]. Since the data-driven template model of Chapter 4 can only compose face videos under the training lighting conditions, it is brittle against illumination variation. Fig. 5.1 illustrates a failure case, where the hallucination does not resemble the ground truth. Observing that it would be impractical to replicate the template database for all possible lighting configurations, we propose a method to reason about the illumination effects explicitly. We develop an approximate compensation scheme against lighting conditions and demonstrate Face Hallucinations beyond the lighting conditions of the training examples.

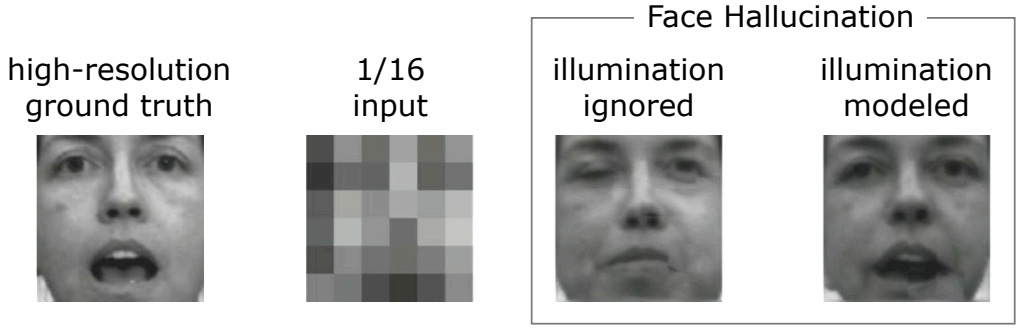

Figure 5.1: Since the data-driven template model of Chapter 4 can only compose face videos under the training lighting conditions, it is brittle against illumination variation. A failure case is shown, where the hallucinated face does not resemble the ground truth. By explicitly modeling the illumination effects and compensating for them, our system will be able to hallucinate more accurately.

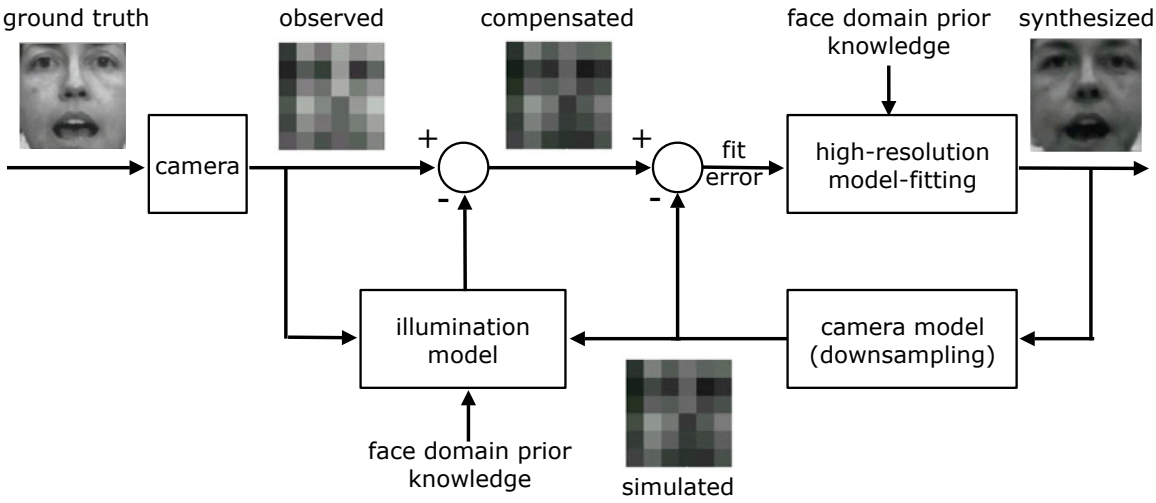

Figure 5.2: We treat illumination as a nuisance parameter and compensate for it. Observed faces are brought to the lighting condition of the face model before the fitting process. This allows the model training to focus on facial expressions rather than illumination artifacts.

We treat illumination as a nuisance parameter. Our strategy is to compute a point estimate of this variable and to eliminate its effect in input videos. Fig. 5.2 illustrates how the observed faces are brought to the (training) lighting condition of the model before the fitting process. Independent of the *observed* illumination, our algorithm hallucinates faces in the *reference* illumination of the training data. This allows our template training to focus on facial expressions rather than lighting artifacts.

5.1 Explaining the Illumination Effects

Assume that T is the perfect template image, *i.e.*, it is the high-resolution version of the observed face L. Thus, any difference between the simulated low-resolution template AT and the observation L must be purely an illumination artifact. This is illustrated in Fig. 5.3, where the "illumination mismatch" term I is a vector that captures the difference between the training and testing lighting conditions. Conversely, subtracting I from L brings the observation into the lighting of the training set, *i.e.*, compensates for illumination.

In practice, one has to estimate both T and I, a non-trivial joint optimization problem. Observe that the mismatch term I has to be constrained; otherwise the observed L could be arbitrarily matched to any template AT. In the following, we propose a constraint for I.

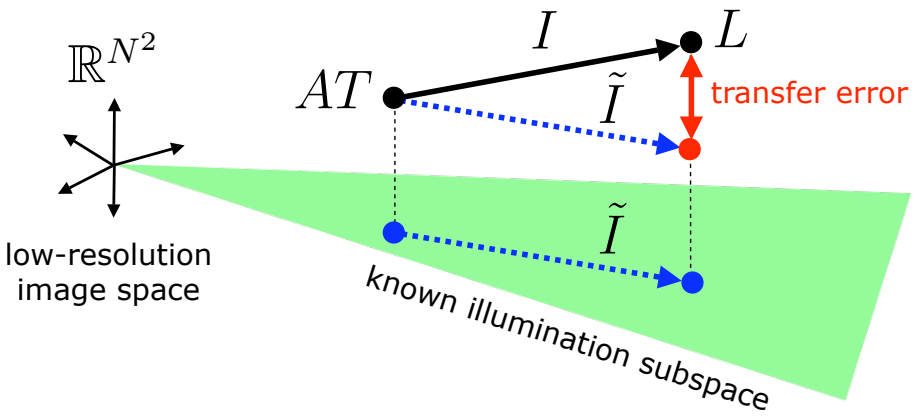

Figure 5.3: We regularize the "illumination mismatch" variable I with a subspace. The constrained estimate, denoted by $\tilde{I}$, is computed between the *projections* of observed (L) and simulated (AT) low-resolution images onto a given illumination subspace.

An Approximate Regularization

Prior work showed that the set of face images obtained by varying the lighting lied in a low-dimensional subspace [Shashua, 1992; Hallinan, 1994; Belhumeur and Kriegman, 1998; Yuille et al., 1999; Basri and Jacobs, 2003]. One can use Singular Value Decomposition to compute a linear (illumination) basis for this subspace from just a few images.

Can we regularize the mismatch term I with an illumination subspace? Note a key restriction here: subspace models only hold for rigid and static objects. This assumption is constantly violated in Face Hallucination due to natural head motion and facial expressions. Unfortunately, learning an illumination subspace for each and every pose/expression configuration would not be practical. Even if one could build such an array of models, indexing the correct subspace would still require perfectly recognizing the facial pose/expression from low-resolution videos, which constitutes the inference sub-problem of Face Hallucination.

Given the difficulty above, we resort to an approximation that ignores the variations in pose and expression. As shown in Fig. 5.3, we choose the constrained $\tilde{I}$ to be the difference vector between the *projections* of AT and L onto a given illumination subspace (*e.g.*, built for a face with a neutral expression). Since this operation transfers the illumination mismatch vector $\tilde{I}$ to a new point in the space of low-resolution images, the approximated compensation incurs a "transfer error" = $I - \tilde{I}$.

How valid is the above approximation? In the following section, we quantify the transfer errors among four exemplar subspaces of a subject, and observe that the relative error magnitudes (with respect to the ground truth I) are acceptably small in low-resolution. The results suggest that, in low-resolution, distinct illumination subspaces of a face —each coming from a different pose and expression— still exhibit similar variation with respect to illumination. The approximate illumination compensation algorithm of Section 5.3 will exploit this structure.

5.2 Quantifying Approximation Errors due to Mismatch of Illumination Subspaces

We proposed a method for compensating low-resolution face images against illumination effects. This involved an approximation when the pose/expression of the observed face did not exactly match that of the assumed illumination subspace. We now present empirical evidence about the magnitude of the approximation errors.

5.2.1 Experimental Setup

For a quantitative analysis, we performed experiments wherein we artificially generated a large number of approximation instances. We applied our approximate method and compared the results against ground truth, where no approximation was necessary. To investigate the role of image resolution, we repeated all procedures for both high-resolution (`96x96` pixels) and low-resolution (`6x6` pixels) faces.

As a representative set of pose/expression mismatches, we used four different facial expressions (neutral, smiling, frowning, and angry) with small pose perturbations. At each of these configurations, we captured a few hundred images under varying illumination. Within each set we randomly selected six images, shown in Fig. 5.4, to characterize the illumination subspace through SVD.

5.2.2 Generating Approximation Instances

Our setup with four illumination subspaces is illustrated in Fig. 5.5. The point clusters overlaid on subspaces represent the face images collected for the corresponding pose and expression. Note that the spread within each cluster is due to lighting variation only.

Figure 5.4: We built four illumination subspaces corresponding to the neutral, smiling, frowning and angry expressions with small pose perturbations. Each row shows the training images used to compute the linear basis for the corresponding subspace.

We define the following "illumination problem": we declare the mean of each cluster to represent the *reference* lighting condition and we use the compensation method to normalize each image. In other words, we try to bring every point to the mean of its cluster. Recall that the compensation method exploits a subspace, assumed to be known *a priori*, in regularizing the illumination mismatch variable I. When the observed pose/expression does not exactly match this model, the method is only approximate and incurs a transfer error.

To get a representative sampling of the transfer error, we apply our algorithm to all available images, using each of the learnt subspaces. This results in a `4x4` matrix of pose/expression (mis)matches. The diagonal of this matrix corresponds to cases where the subspace is exact, *i.e.*, there is no transfer error. Off-diagonal cases are those with a subspace mismatch: for instance, when neutral face image is compensated with the subspace constraint of a smiling face, there will be a non-zero transfer error, *i.e.*, $I \neq \tilde{I}$.

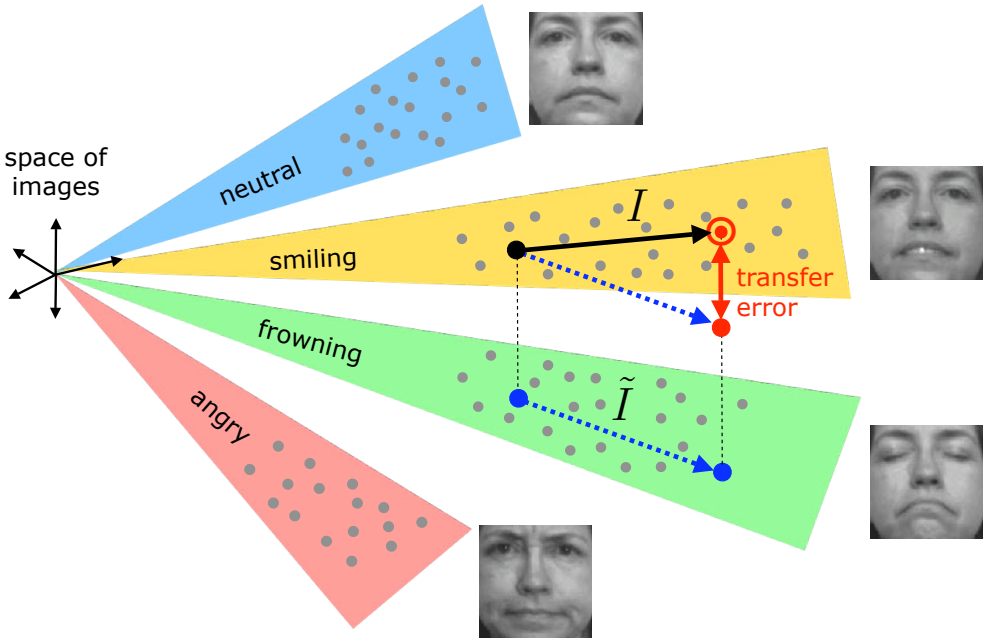

Figure 5.5: Our illumination compensation method incurs a "transfer error" whenever the pose/expression of the observed face does not exactly match that of the assumed illumination subspace. To quantify these errors, we artificially generate problem instances where we apply our method and compare the results against ground truth.

5.2.3 Quantitative Evaluation

Without a reference point, the magnitude of a transfer error is difficult to interpret. A more intuitive measure is the relative magnitude of the transfer error $I - \tilde{I}$ with respect to the ground truth compensation vector I:

$$\% \text{ relative error} = \frac{\|I - \tilde{I}\|^2}{\|I\|^2} \times 100.$$

In Fig. 5.6, we report the mean value of this metric for the off-diagonal entries of the mismatch matrix. We observe that, in low-resolution, the relative error magnitudes remain acceptably small. These results suggest that distinct illumination subspaces of a face — each coming from a different pose and expression— still exhibit similar variation with respect to illumination. Our approximate illumination compensation exploits this structure.

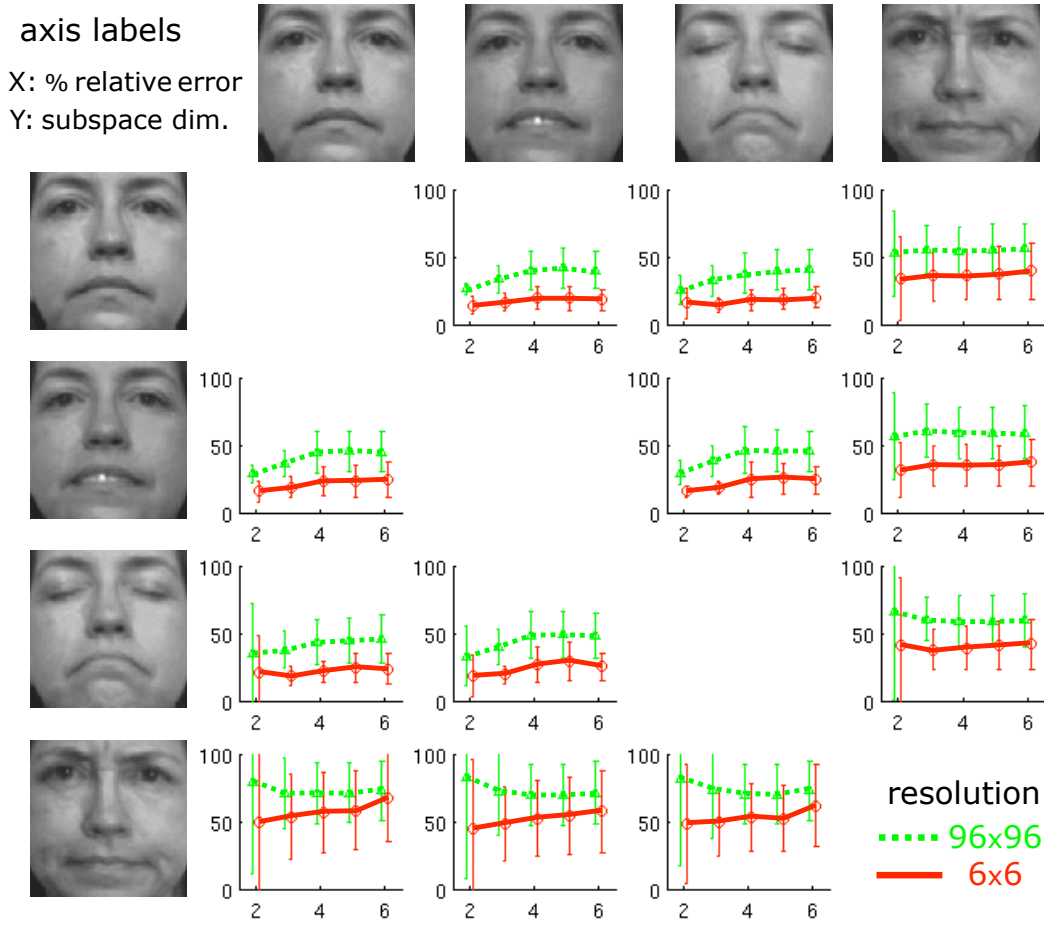

Figure 5.6: In low-resolution, the relative transfer error magnitudes remain acceptably small. This suggests that distinct illumination subspaces of a face exhibit similar variation with respect to illumination.

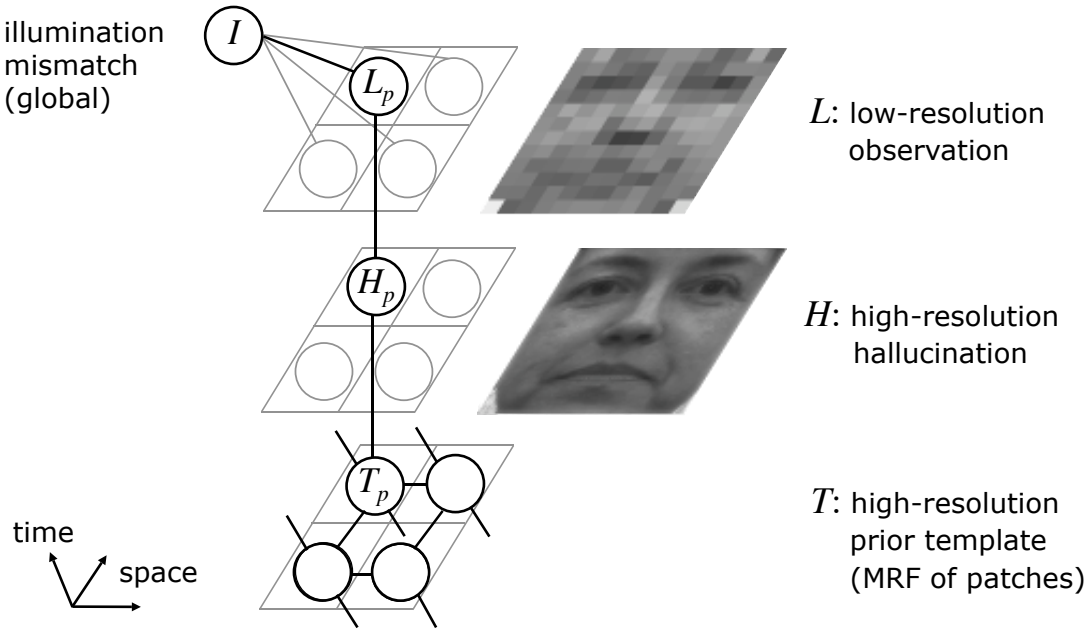

Figure 5.7: We augment the graphical model of Fig. 4.2 with the "illumination mismatch" variable I that can explain the difference between the simulated and observed low-resolution images as a lighting effect. Observe that the facial expression component T and the illumination component I of this model are tightly coupled through the observation L.

5.3 Augmenting the Graphical Model

We augment the graphical model of Fig. 4.2 with a global (per image) illumination mismatch variable I for each frame of the video. As before, the high-resolution template T acts as prior on the hallucination H, which is blurred and downsampled to simulate the formation of the low-resolution image L. The novelty is that the model can now explain the difference between the simulated and observed low-resolution images as a lighting effect. The facial expression component T and the illumination component I are tightly coupled through the observation L.

Inference

Preserving the probabilistic formulation of the video hallucination problem, we marginalize over both the unknown template T and the illumination I:

$$P(H \mid L) = \sum_I \sum_T P(H, I, T \mid L).$$

Following the derivation presented in Section 4.3, we rewrite the posterior as

$$\sum_{I}\sum_{T} P(H \mid I,T,L)\; P(I,T \mid L). \tag{5.1}$$

As we approximate the joint posterior $P(I,T \mid L)$ in (5.1) with a delta-function at (I^*,T^*), finding H_{MAP} turns into the quadratic minimization problem of

$$\|L - I^* - AH\|^2 \;+\; \frac{\sigma_L^2}{\sigma_H^2}\,\|T^* - H\|^2. \tag{5.2}$$

Just as in (4.8), the first term encourages those high-resolution videos H that best *reconstruct* the observation L, this time up to an illumination mismatch I^*.

An Iterative Algorithm

We take an iterative approach to find the maximizer of the posterior $P(I,T \mid L)$. To start, we assume there is no illumination mismatch between the training and testing conditions, and we choose the template to be the average of all examples in the database. We solve for I^* and T^* iteratively using Alg. 2, which typically converges after 5 iterations.

Regularizing the Illumination Mismatch I^*

We regularize the estimate I^* with a low-dimensional illumination (image) subspace built for the subject in the template database. In Alg. 3, the orthogonal basis images of this subspace are denoted by B_b, and the notation $<\cdot,\cdot>$ stands for the inner-product. We project each frame of the input and simulated low-resolution peak template videos (denoted by L_f and AT_f^*, respectively) onto the given subspace. The regularized illumination term is the difference between these projections.

5.4 Qualitative Results

To characterize the illumination subspace, we asked the subject to maintain a neutral face while we captured 6 more snapshots, this time with lighting variation. We blurred and downsampled these images to `6x6` pixels, ran SVD analysis, and adopted the first two

input : observed video L
output: illumination mismatch I^*, peak template T^*

/ initialize I^* with zero (no mismatch) */*

$I^* \leftarrow 0$

/ initialize T^* with the mean of the database (with D entries) */*

for *all video patch locations* p **do**

$$T_p^* \leftarrow \frac{1}{D} \sum_k t_k \text{ such that } s_k = p$$

end

repeat

given T^*, solve for I^* */* see Alg. 3 */*
given I^*, solve for T^* */* see Alg. 1 */*

until *convergence* ;

Algorithm 2: We solve for I^* and T^* iteratively.

input : observed video L, peak template T^*
output: illumination mismatch I^*

for *all video frames* f **do**

$$I_f^* \leftarrow \underbrace{\sum_{b=1}^{2} < B_b, L_f > B_b}_{\text{subspace projection of } L_f} - \underbrace{\sum_{b=1}^{2} < B_b, AT_f^* > B_b}_{\text{subspace projection of } AT_f^*}$$

end

Algorithm 3: For regularization purposes, we compute I^* in a low-dimensional illumination subspace.

basis vectors for our subspace representation. With two illumination coefficients, our model captured more than 50% of the overall variation within the subspace.

To test our algorithm under novel illumination conditions, we recorded videos under dynamically varying lighting, obtained by swinging a diffuse light source in front of the subject. Recall that our model treats illumination as a nuisance parameter and hallucinates faces in the same illumination condition as in the training set. This complicates the quantitative evaluation of hallucinations under novel lighting conditions: even if a facial expression is recovered correctly, the hallucinated pixel intensities can be very different from those of the ground truth. In such cases, signal-level reconstruction metrics such as the MSE would not reflect the accuracy of the recovered faces. For this reason, we only show hallucination snapshots for a qualitative assessment[1].

In Fig. 5.8, we visually compare hallucination results for selected test frames. For brevity, we only include hallucination results with the full spatio-temporal interactions enabled, since this yields the most accurate results. The first column shows the `6x6` pixel input, whereas the last column shows the underlying `96x96` pixel ground truth. Observe how the images of this test are, in general, brighter than those shown in Fig. 4.11.

In the second column of Fig. 5.8, we ignore the lighting effects and attempt to hallucinate without illumination compensation. Since our data-driven template model and feature vectors largely depend on pixel intensities, the illumination mismatch leads to total failure: in addition to exhibiting blocking artifacts, the hallucinated face does not replicate the behavior of the ground truth face.

As we enable our method's lighting compensation mechanism, we recover the illumination mismatch images of the third column of Fig. 5.8. We plot this variable as a 3D surface over the `6x6` low-resolution pixel grid. The slant of this surface suggests that one side of the face received more light than the other. This can be visually confirmed in the low-resolution images: scanning the pixels from left to right, faces indeed get brighter. When our algorithm adjusted the observed video for illumination, it inferred the hallucinations shown in the fourth column. Notice the dramatic improvement in the hallucination quality: the mouth and eye motions are recovered successfully.

[1] Videos are available at http://www.cs.cmu.edu/~dedeoglu/thesis

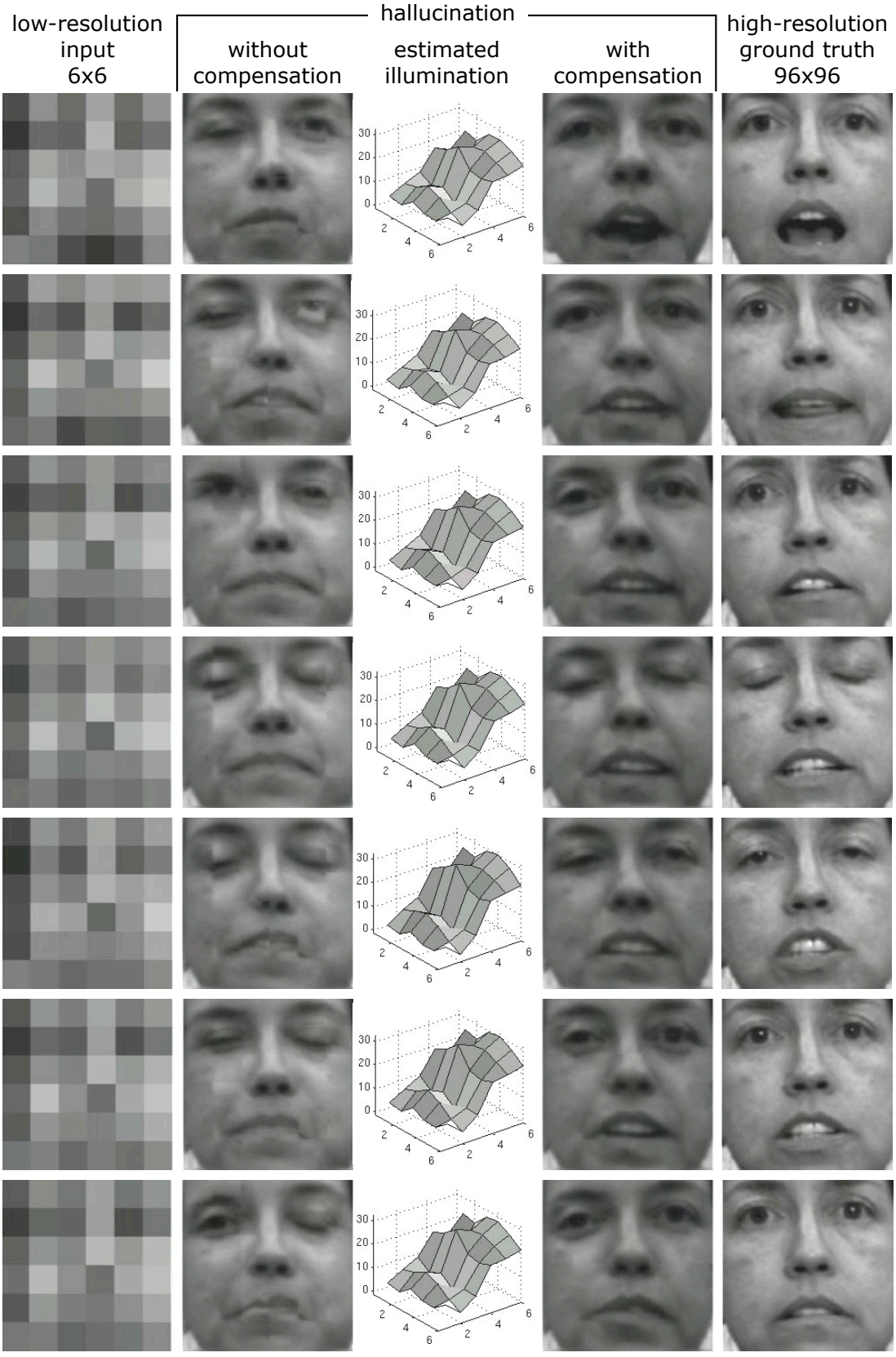

Figure 5.8: Hallucination benefits from illumination compensation (see Section 5.4).

5.5 Multiple and Mixed Illumination Subspaces

For lighting compensation, we exploited the illumination subspace of the subject's face under a neutral expression. Consequently, the compensation vector is exact for only one pose/expression configuration; otherwise it is an approximation. An immediate extension of this approach would be to build a set of illumination models for a sampling of pose and expressions, and to adaptively switch from one to another.

One might also consider a single but *mixed* illumination model that would blend various pose and expressions. Observe that this would not model an illumination subspace per se; instead it would capture a general appearance subspace with significant illumination variation. The fundamental problem with this approach would be the competition between the template and the (mixed) illumination components to *explain* facial expressions. If part of the expression signal is interpreted as illumination artifact and removed from the data, the template model would not be able to recover the underlying expression. Thus, it is desirable to limit the expressive power of the illumination component with lighting effects.

5.6 Conclusion

We proposed an approximate compensation scheme against lighting conditions. Our approach was motivated by the observation that, in low-resolution, distinct pose/expression illumination subspaces of a face exhibited similar variation with respect to illumination. To exploit this structure, we augmented the video model of Chapter 4 with a low-dimensional illumination subspace and solved for its parameters jointly with high-resolution face details. This allowed Face Hallucinations beyond the lighting conditions of the training examples.

Chapter 6

Conclusion

This thesis aimed to recover and to reconstruct subtle signals in degraded images and videos. In particular, we proposed models and inference algorithms to analyze low-resolution videos of human faces. We have demonstrated that a careful exploitation of space (image) and space-time (video) models could yield effective solutions to the problem of face resolution enhancement, or Face "Hallucination".

6.1 Summary of Achievements

Throughout this thesis, we demonstrated accurate restoration of facial details, with person-specific resolution enhancements up to a scaling factor of 16. To highlight the achievements of the proposed algorithms, we include a selection of hallucination examples.

6.1.1 Exploiting an Image Model

Chapter 3 demonstrated the importance of carefully crafted metrics and algorithms in meeting the challenges of resolution degradation in Face Hallucination. The key observation was a resolution-induced asymmetry in model-to-image or image-to-image registration problems: under relative scaling, one must start with the higher-resolution image (or model) and warp it onto the lower-resolution one while incorporating a blur-formation process in the fitting criterion.

We showed that the asymmetry principle is most relevant to Face Hallucination. We adopted the popular AAM as a face model, and showed how the traditional AAM fitting

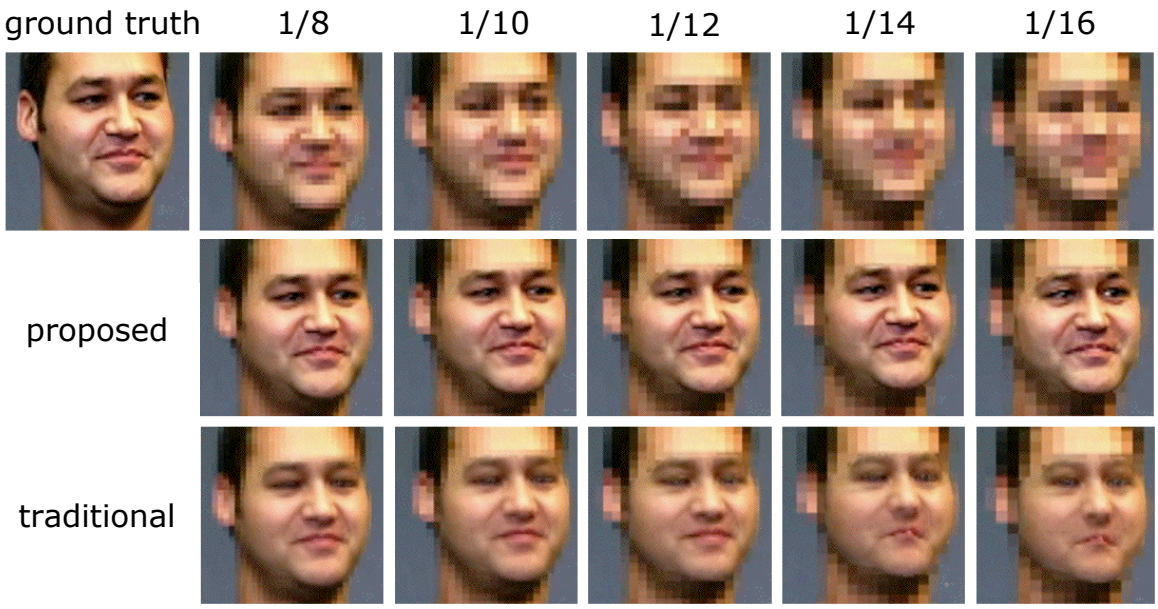

Figure 6.1: Our novel AAM fitting formulation yielded significantly more accurate reconstructions of facial details (middle). State-of-the-art algorithms (bottom) that relied on the traditional formulation were shown to exhibit a systematic bias.

formulation overlooked the asymmetry issue. This caused the fitting accuracy to degrade quickly when the observed faces were smaller than their model. We then formulated a novel fitting criterion that respected the asymmetry, and derived a numerical optimization method for it.

We compared the proposed algorithm against a state-of-the-art method across a variety of resolutions and AAM complexity levels. For a quantitative analysis, we first defined accuracy metrics based on the shape and appearance of parameters of AAMs. We then compared and contrasted the algorithms on various AAMs (both single- and multiperson) across a range of input resolutions. As shown in Fig. 6.1, our novel fitting algorithm (middle) proved to be significantly more accurate in estimating and reconstructing faces.

6.1.2 Exploiting a Video Model

Chapter 4 demonstrated that Face Hallucination can benefit from the spatio-temporal dynamics of faces. To investigate the role of time, we first devised a novel generative model of face videos. This model treated a video as a composition of space-time patches and

Figure 6.2: Exploiting the spatio-temporal dynamics of faces, our video hallucination algorithm (bottom) reconstructed facial expressions that closely resembled the ground truth (middle). Quantitative experiments revealed the smoothing role of temporal dynamics in overcoming 16-fold resolution degradations.

encoded visual phenomena in a non-parametric, example-based fashion. The patch-based representation was also used to define a prior in both space and time.

We ran extensive hallucination experiments and quantified the effect of spatial and temporal models on the hallucination performance. Using both signal reconstruction and objective visual quality metrics, we demonstrated the smoothing role of temporal dynamics in Face Hallucination. As illustrated in Fig. 6.2, our algorithm produced high-resolution expressions (bottom) that closely resembled the ground truth (middle). Our results highlighted the importance of a video's temporal dimension in hallucinating facial expressions correctly.

Chapter 5 proposed a compensation scheme against illumination. This involved augmenting the video model of Chapter 4 with a low-dimensional illumination subspace. Illumination was treated as a nuisance parameter: its effects were estimated and then removed from observed videos. The proposed algorithm permitted Face hallucinations beyond the lighting conditions of the training videos.

6.2 Contributions

The contributions of this thesis can be summarized as follows:

- It has demonstrated that Face Hallucination critically depends on model-fitting metrics: a resolution-induced bias was shown to affect most model-to-image and image-to-image fitting algorithms operating on low-resolution images. It was found that models and observations should be treated *asymmetrically* both to formulate an unbiased objective function and to derive an accurate optimization algorithm. The analysis underlined the inherent trade-off between computational efficiency and estimation accuracy in low-resolution regimes.

- It has proposed a model-fitting algorithm that respected the above-mentioned asymmetry: it adopted the popular Active Appearance Model and derived a novel Face Hallucination and tracking algorithm that proved significantly more accurate than state-of-the-art methods in low-resolution.

- It has demonstrated how Face Hallucination could benefit from facial dynamics: a statistical generative model of face videos was proposed to represent and reason about facial expressions. This model treated videos as compositions of space-time patches, efficiently capturing complex visual phenomena such as eye-blinks and the occlusion or appearance of teeth.

- It has exploited the space-time representation to define a data-driven face prior on a 3-dimensional Markov Random Field. It posed Face Hallucination as a probabilistic inference problem and demonstrated the crucial role of a video's temporal dimension in hallucinating the correct facial behaviors.

- It has proposed an approximate compensation scheme against illumination variation. It augmented the generative video model with a low-dimensional illumination subspace, whose parameters were estimated jointly with high-resolution face details. This made Face Hallucinations beyond the lighting conditions of the training videos possible.

- It has achieved person-specific resolution enhancements up to a scaling factor of 16.

6.3 Limitations and Future Directions

Chapters 3 and 4 have already discussed specific aspects of the proposed image- and video-based approaches to Face Hallucination. To conclude, we comment on some of the limitations and identify directions for future investigation.

6.3.1 Hallucinating Familiar vs. Unfamiliar Subjects

Could we hallucinate the faces of subjects we have never seen before? This was the question addressed in the seminal work of [Baker and Kanade, 2002] and the answer was shown to be positive. Similar results can be found in the recent work of [Liu et al., 2007]. In contrast to the 16-fold *subject-specific* resolution enhancements of this thesis, hallucination for *generic* faces could only be reliably demonstrated up to a scaling factor of 4. This difference in performance is not surprising, since the construction and fitting of generic face models is substantially more difficult compared to person- or group-specific ones [Gross et al., 2005].

In this thesis, we assumed that we could learn a model *a priori* for the subjects whose faces we would be hallucinating. In a practical surveillance scenario, such models could be built for tracking, recognizing or verifying the personnel of a particular facility. The generality of the face model can be an important requirement for certain applications, but this issue remains orthogonal to our contributions.

Hallucination with a Mismatching Model

What would we hallucinate if our face model did not match the test subject? To answer this question, we fit the multiperson AAM of Section 3.4 to low-resolution images of a subject outside the training set. Fig. 6.3 depicts two "ground truth" test frames that are downsampled progressively from left to right, followed by hallucinations generated by the traditional (AAMR-SIM) and the proposed (RAF) fitting algorithms. Observe that, even at higher resolutions, the mismatching model is unable to reconstruct the *appearance* of the underlying face. Nevertheless, there are similarities between the hallucinated and ground truth *expressions*. As the input data becomes lower in resolution, the proposed fitting algorithm (RAF) is able to extract the facial pose, teeth and eyes better than the traditional one (AAMR-SIM). Even though Face Hallucination suffers overall from the mismatching model, there are observable benefits in using an accurate model-fitting algorithm.

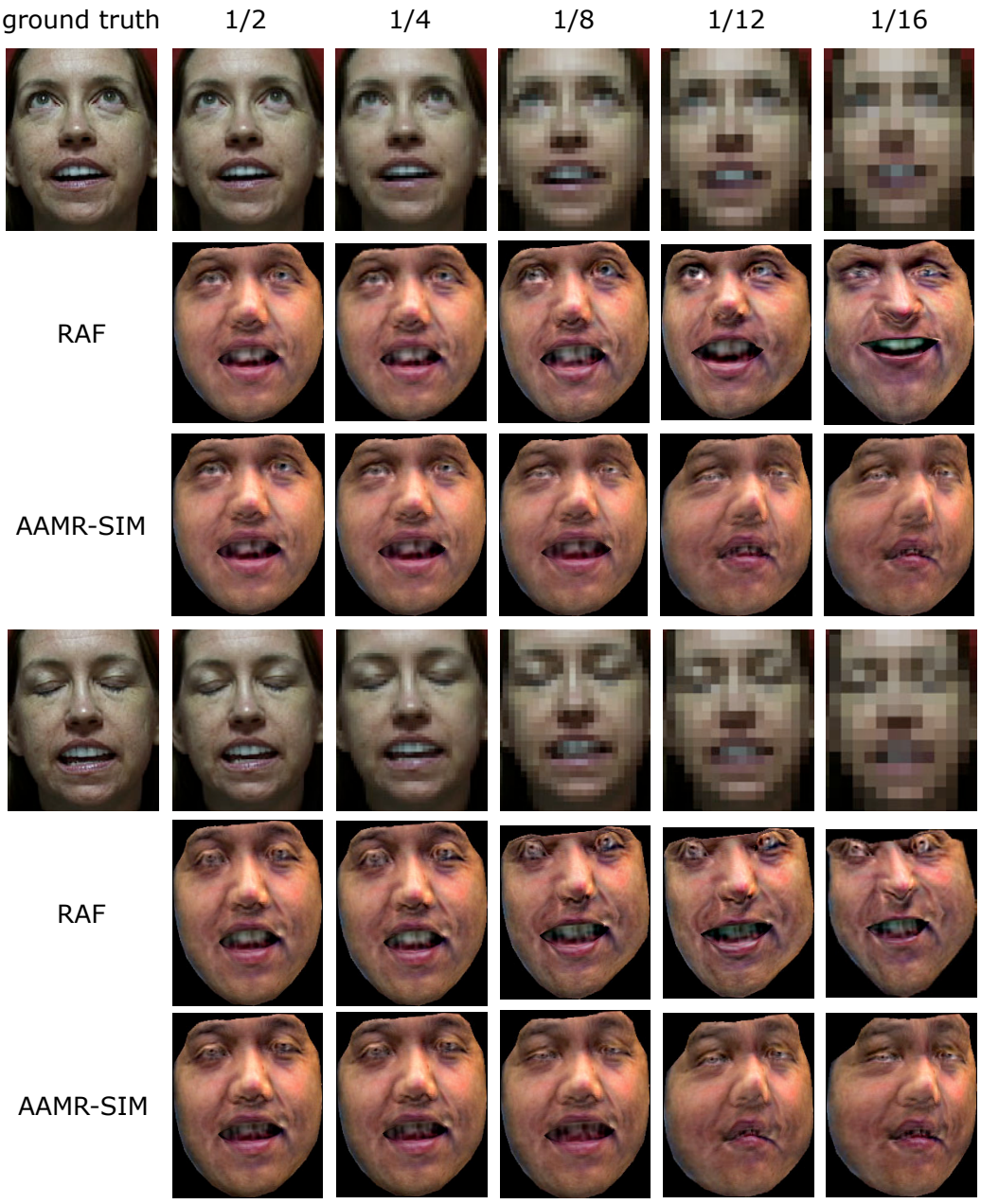

Figure 6.3: What would we hallucinate if our face model did not match the test subject? We fit the multiperson AAM of Section 3.4 to low-resolution observations of a subject outside the training set. Even in high-resolution, the mismatching model is unable to reconstruct the underlying face. Nevertheless, there are similarities between the hallucinated and ground truth expressions. As the input data becomes lower in resolution, the proposed fitting algorithm (RAF) is able to extract the facial pose, teeth and eyes better than the traditional one (AAMR-SIM). Even though Face Hallucination suffers overall from the mismatching model, there are observable benefits in using an accurate model-fitting algorithm.

6.3.2 Ambiguity Analysis: What's the Limit?

Face Hallucination is a learning-based approach to super-resolution. It starts by extracting the object/scene features that have survived the destructive effects of camera blur & quantization, and it uses this information to constrain the space of high-resolution solutions. Since blur destroys information, lower resolution observations become increasingly ambiguous. If an image is blurred to the point of leaving no discriminative information in it, one cannot infer the underlying state of the world and reconstruct it.

Analyzing when the resolution-induced ambiguities arise in images and how fast they grow would reveal the limits of learning-based super-resolution algorithms. With an understanding of these limitations, one might be able to improve upon both models and algorithms. For instance, if one could first gauge the difficulty of a given hallucination problem and estimate the level of visual details that can be reliably recovered, one could choose a face model of appropriate complexity.

6.3.3 Performance Metrics

An open problem in the domain of resolution enhancement is the lack of objective quality metrics for a quantitative assessment of the results. The MSE measure is commonly used in the literature, although it does not always reflect the perceived quality by humans [Girod, 1993]. For instance, an algorithm might be able to super-resolve and enhance a high-contrast edge very well, except for a small error in its position. While the the MSE metric would heavily penalize such reconstructions, the human eye may forgive the geometric imperfection.

Quality metrics can also be tailored for particular applications of Face Hallucination. For instance, one could define metrics based on semantic events such as eyeblinks and mouth deformations. One could disregard a reconstructed eye's appearance but focus on how wide the eyelid is open. Alternatively, one could penalize for mismatches in the timing and duration of eyeblinks. In some applications it may be sufficient for the hallucinated face to show the correct facial expression such as sadness, surprise, and happiness.

Appendix A

Comparing the Forward and Backward Algorithms

Assuming the idealized scenario of Section 3.2.2, let us express the image warp operations of (3.15) using point coordinates

$$\hat{\mathbf{W}}_{\mathbf{21}} = \arg\min_{\mathbf{W_{21}}} \int_{\mathbf{y}\in\mathbf{dom}I_1} \Big[S\big(\hat{\mathbf{W}}_{\mathbf{1S}}(\mathbf{y})\big) - S\Big(\hat{\mathbf{W}}_{\mathbf{2S}}\big(\underbrace{\mathbf{W_{12}}(\mathbf{y})}_{\mathbf{z}}\big)\Big)\Big]^2 \mathbf{dy}. \tag{A.1}$$

We can rewrite the integration in (A.1) in the domain of I_2 by defining $\mathbf{z}=\mathbf{W_{12}}(\mathbf{y})$. Since $\mathbf{dy}=\big|J(\mathbf{W_{21}})\big|\mathbf{dz}$, (A.1) can be written as

$$\begin{aligned}\hat{\mathbf{W}}_{\mathbf{21}} &= \arg\min_{\mathbf{W_{21}}} \int_{\mathbf{z}\in\mathbf{dom}I_2} \Big[S\Big(\hat{\mathbf{W}}_{\mathbf{1S}}\big(\mathbf{W_{21}}(\mathbf{z})\big)\Big) - S\Big(\hat{\mathbf{W}}_{\mathbf{2S}}\big(\underbrace{\mathbf{W_{12}}\big(\mathbf{W_{21}}(\mathbf{z})\big)}_{\mathbf{z}}\big)\Big)\Big]^2 \big|J(\mathbf{W_{21}})\big|\mathbf{dz} \\ &= \arg\min_{\mathbf{W_{21}}} \int_{\mathbf{z}\in\mathbf{dom}I_2} \Big[S\Big(\hat{\mathbf{W}}_{\mathbf{1S}}\big(\mathbf{W_{21}}(\mathbf{z})\big)\Big) - S\big(\hat{\mathbf{W}}_{\mathbf{2S}}(\mathbf{z})\big)\Big]^2 \big|J(\mathbf{W_{21}})\big|\mathbf{dz}. \end{aligned} \tag{A.2}$$

Switching back to image warp notation, (A.2) becomes

$$\hat{\mathbf{W}}_{\mathbf{21}} = \arg\min_{\mathbf{W_{21}}} \int_{\mathbf{z}\in\mathbf{dom}I_2} \Big[warp\big(\underbrace{warp(S;\hat{\mathbf{W}}_{\mathbf{S1}})}_{\hat{I}_1};\mathbf{W}_{21}^{-1}\big)(\mathbf{z}) - \underbrace{warp(S;\hat{\mathbf{W}}_{\mathbf{S2}})}_{\hat{I}_2}(\mathbf{z})\Big]^2 \big|J(\mathbf{W_{21}})\big|\mathbf{dz} \tag{A.3}$$

Recalling $\mathbf{W_{12}} = \mathbf{W_{21}^{-1}}$, we observe that the difference between the *forward* (3.13) and *backward* (A.3) algorithms' objective functions is the extra Jacobian term $|J(\mathbf{W_{21}})|$ in (A.3). Since a general homography's Jacobian varies spatially, this term would normally act as a spatial weighting function and influence the minima of the objective function.

Appendix B

Quantifying the Scaling-Induced Bias

The analysis in Section 3.2 shows that when the scene S is not known, any blur in the imaging system will cause the *forward* and *backward* algorithms to be biased in the presence of relative scaling. Quantifying the blur effect, however, is not trivial because it is ultimately related to image content: while blurring (*i.e.*, low-pass filtering) visually rich and detailed images would produce a significant effect, it would barely alter already smooth images. This motivated us to focus on a particular class of images, namely those of human faces. Accurate registration algorithms are crucial in this domain, because it determines the performance of various tracking, recognition and biometric verification systems. We ran our face-domain experiments on a set of 140 grayscale, frontal face images from the FERET database [Phillips et al., 2000].

To quantify the magnitude of the image registration bias, we generated a variety of synthetic experiments wherein we simulated the image formation process. A real face image, acting as S, was first blurred, then geometrically transformed according to specific warp parameters, and finally resampled to generate images I_1 and I_2. In solving this synthetically generated registration problem, only I_1 and I_2 were used (*i.e.*, unknown scene case).

Given the ground truth warp parameters, we tested whether minimizing our objective function gave accurate estimates of the warp. For simplicity, we limited our investigation to similarity transforms with known scaling parameter (s). This left us with three degrees of freedom, namely, translation (t_x, t_y) and rotation (θ), which we considered independently. Having assigned s and θ their ground truth values, we exhaustively searched for the global best values for translation. Similarly, the global minimum for rotation was sought, with s and (t_x, t_y) set to their correct values. These searches were repeated in the neighborhood of their free parameters' ground truth values, and the magnitude of their biases was recorded.

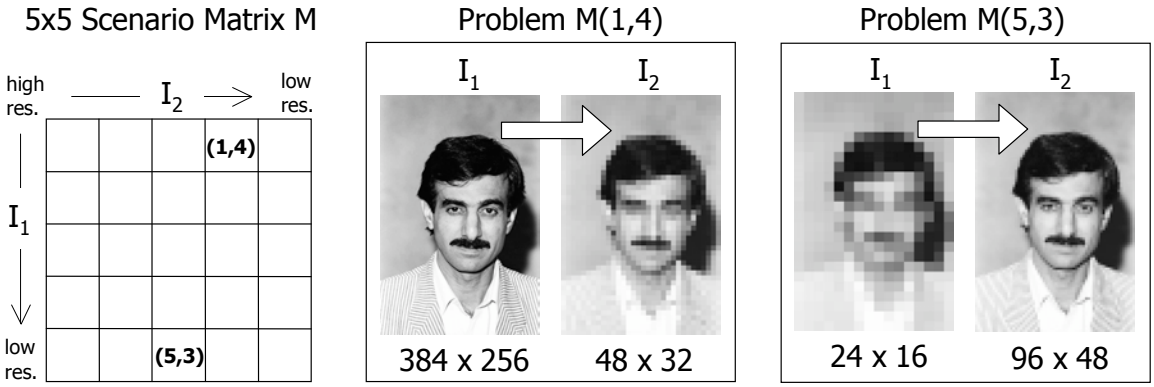

Figure B.1: Our quantitative results are organized in a matrix of registration problem instances (left). Entries of this matrix show the bias statistics of the *forward* algorithm, with blur compensation applied as needed.

B.1 Testing Conditions

Starting from a `384x256` pixel face image, five different scale reduction parameters (by factors of 1, 2, 4, 8, and 16) were used in generating the test images I_1 and I_2. Without loss of generality, rotation and translation parameters were set to zero. As shown in Fig. B.1 (left), this resulted in a `5x5` scenario matrix M of image pairs being registered for each face image. Note that the diagonal elements of this matrix correspond to problem instances where images have the same resolution, whereas off-diagonal elements represent cases where they differ in this respect. Our tests aimed to measure how accurately the ground truth translation and rotation values (*i.e.*, no translation and rotation) could be estimated.

We limited our experiments to the *forward* algorithm, which always warps I_1 onto I_2, regardless of their scale. However, since the full scenario matrix includes all possible pairings, both downscaling and upscaling cases were covered, as exemplified by the instances $M(1,4)$ and $M(5,3)$ in Fig. B.1. Bilinear interpolation was used whenever the source image I_1 of the warp was smaller than the destination image I_2, and no deblurring was attempted. In simulating the imaging process, we used a pillbox PSF whose width in scene pixels equaled the integer downscale factor. Similarly, when I_1 was being downscaled, the extra blurring to be applied to T' in (3.24) was also obtained using a pillbox PSF whose width in I_1 pixels equaled the relative scale factor.

For practical reasons, our "global search" for the best parameter settings was limited to the immediate neighborhood of corresponding ground truth values. For translation, we

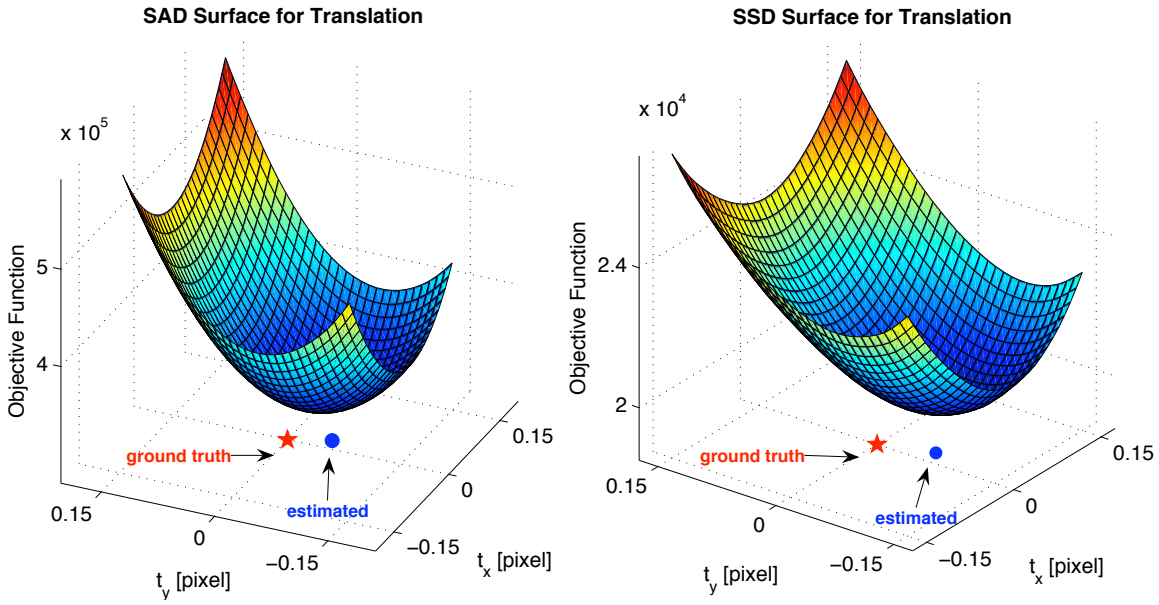

Figure B.2: L_1 (left) and L_2 (right) norm objective functions for translation parameters of problem instance $M(5,3)$ of Fig. B.1. The *forward* algorithm is biased in this case, because its global solution (•) does not coincide with the ground truth (⋆).

sampled the interval of $[-0.20, +0.20]$ pixels in 0.01 pixel increments for both t_x and t_y. Similarly, we sampled the $[-2, +2]$ degree interval in 0.1 degree increments for rotation (around the center of the image). Fig. B.2 shows example surfaces obtained by sampling translation parameters of L_1 and L_2 norm objective functions for the problem $M(5,3)$ of Fig. B.1, where I_1 is lower in resolution than I_2. As predicted, the global minima of these functions do not lie at the origin, confirming a bias due to the problem formulation (*i.e.*, the objective function) itself, rather than the assumed noise or the minimization method.

B.2 Quantitative Results

The *forward* algorithm was investigated for all 25 entries of the scenario matrix, and repeated for 140 face images. Fig. B.3 reports means and standard deviations of the computed bias magnitudes, organized in the same matrix form as the scenario matrix. Note that the translation biases are originally measured in I_2 pixel units, because t_x and t_y are added onto I_1 pixel coordinates *after* scaling. Fig. B.4 displays selected histograms of translation biases.

Translation Bias [pixel]

L1 Norm Obj. Func.

	1	**1/2**	**1/4**	**1/8**	**1/16**
1					
1/2	.01/.01				
1/4	.01/.01	.01/.01			
1/8	.02/.01	.02/.01	.02/.01		
1/16	.03/.02	.04/.02	.04/.02	.04/.02	

L2 Norm Obj. Func.

	1	**1/2**	**1/4**	**1/8**	**1/16**
1					
1/2	.01/.01				
1/4	.01/.01	.02/.01			
1/8	.03/.02	.03/.02	.03/.02		
1/16	.04/.02	.06/.02	.07/.04	.07/.04	

Rotation Bias [degree]

L1 Norm Obj. Func.

	1	**1/2**	**1/4**	**1/8**	**1/16**
1					
1/2	.00/.00				
1/4	.00/.00	.00/.00			
1/8	.00/.02	.00/.00	.00/.00		
1/16	.17/.26	.11/.26	.02/.13	.00/.00	

L2 Norm Obj. Func.

	1	**1/2**	**1/4**	**1/8**	**1/16**
1					
1/2	.00/.00				
1/4	.00/.00	.00/.00			
1/8	.01/.03	.00/.01	.00/.00		
1/16	.14/.19	.11/.21	.04/.14	.01/.07	

Figure B.3: The translation and rotation bias magnitude of the *forward* algorithm organized in the 5x5 scenario matrix form. Rows and columns correspond to scaling factors applied to I_1 and I_2, respectively. Entries are: Mean/Standard Deviation. The diagonal and upper triangle of the matrices are expected, and empirically verified to be zero. For clarity, zero's are not shown in the table. Since the translation parameters are in I_2 pixel units, so are their reported biases. See Fig. B.5 for scale-normalized versions of translation biases.

The lower triangle of the matrices corresponds to cases where I_1 is lower in resolution than I_2, calling for bilinear interpolation of I_1 during the warp. Confirming our analysis, both translation and rotation parameters are found to be biased. For a given column (*i.e.*, fixed I_2 resolution), we observe that the bias in question gets larger as I_1 is degraded in resolution. This is due to the increased mismatch between T and T' as discussed in Section 3.2.3, and the fact that the computed objective function increasingly relies on interpolation.

The diagonal and upper triangle of the matrices represent cases where I_1 is equal or higher in resolution than I_2. Following our analysis, we first compensate for the difference between T and T' by blurring I_1 as needed, and then proceed with the geometric warp. As expected, the bias in these cases is empirically found to be zero. For clarity, these entries are not shown.

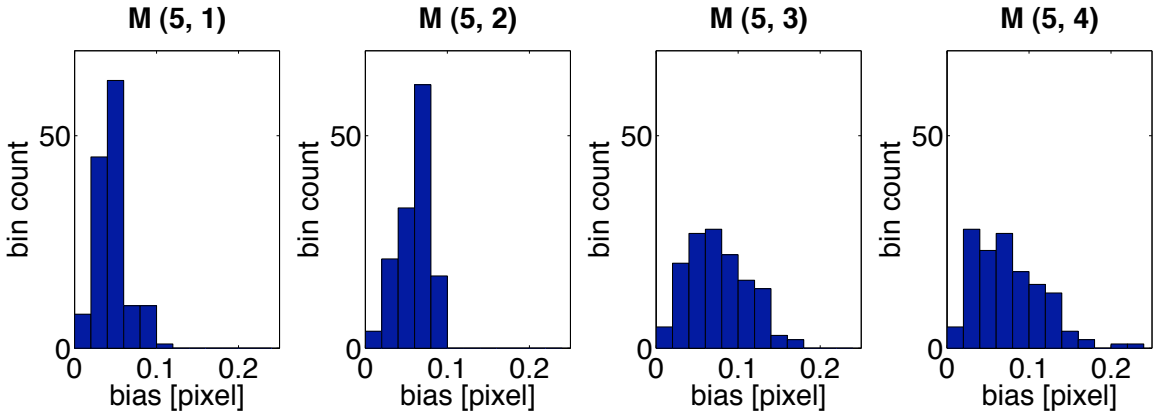

Figure B.4: Translation bias histograms for the L_2 norm objective function, corresponding to the last row (*i.e.*, I_1 scaled by 1/16) of the upper-right scenario matrix M in Fig. B.3.

Translation Bias [pixel]

L1 Norm Obj. Func.

	1	**1/2**	**1/4**	**1/8**	**1/16**
1					
1/2	.01/.01				
1/4	.01/.01	.03/.02			
1/8	.02/.01	.04/.03	.09/.05		
1/16	.03/.02	.08/.04	.17/.09	.36/.19	

L2 Norm Obj. Func.

	1	**1/2**	**1/4**	**1/8**	**1/16**
1					
1/2	.01/.01				
1/4	.01/.01	.03/.02			
1/8	.03/.03	.06/.04	.13/.07		
1/16	.04/.02	.12/.04	.30/.15	.59/.33	

Figure B.5: Scale-normalized translation bias values (*c.f.* Fig. B.3)

B.3 Scale-Normalized Translation Biases

As indicated in Section B.2, the translation bias results of Fig. B.3 are reported in I_2's pixel units. However, since I_2 has a different resolution in every column, the entries of the matrix are not directly comparable. In Fig. B.5, we replicate these bias values in a common (highest-resolution) scale.

Bibliography

Y. Adini, Y. Moses, and S. Ullman. Face recognition: The problem of compensating for changes in illumination direction. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 19(7):721–732, 1997.

P. Anandan. A computational framework and an algorithm for the measurement of visual motion. *International Journal of Computer Vision*, 2:283–310, January 1989.

H. C. Andrews and B. R. Hunt. *Digital Image Restoration*. Prentice-Hall, Inc., Englewood Cliffs, New Jersey, June 1977.

A. Ashburner, J. L. R. Andersson, and K. J. Friston. High-dimensional image registration using symmetric priors. *NeuroImage*, 9:619–628, 1999.

C. B. Atkins. *Classification-based method in optimal image interpolation*. PhD thesis, Purdue University, 1998.

C. B. Atkins, C. A. Bouman, and J. P. Allebach. Tree-based resolution synthesis. In *Proceedings of Image Processing, Image Quality, Image Capture, Systems Conference (PICS)*, pages 405–410, 1999.

T. Bachmann. Identification of spatially quantized tachistoscopic images of faces: How many pixels does it take to carry identity? In V. Bruce, editor, *Face Recognition*, pages 87–103. Lawrence Erlbaum Associates, 1991.

S. Baker and T. Kanade. Super resolution optical flow. Technical Report CMU-RI-TR-99-36, Robotics Institute, Carnegie Mellon University, Pittsburgh, PA, 1999a.

S. Baker and T. Kanade. Hallucinating faces. Technical Report CMU-RI-TR-99-32, Robotics Institute, Carnegie Mellon University, Pittsburgh, PA, September 1999b.

S. Baker and T. Kanade. Limits on super-resolution and how to break them. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 24(9):1167–1183, 2002.

S. Baker and I. Matthews. Equivalence and efficiency of image alignment algorithms. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, volume 1, pages 1090–1097, 2001.

S. Baker and I. Matthews. Lucas-Kanade 20 years on: A unifying framework. *International Journal of Computer Vision*, 56(3):221–255, March 2004.

S. Baker, R. Gross, and I. Matthews. Lucas-Kanade 20 years on: A unifying framework: Part 3. Technical Report CMU-RI-TR-03-35, Robotics Institute, Carnegie Mellon University, Pittsburgh, PA, November 2003.

M. R. Banham and A. K. Katsaggelos. Digital image restoration. *IEEE Signal Processing Magazine*, 14(2):24–41, March 1997.

D. F. Barbe, editor. *Charge-Coupled Devices*. Springer-Verlag, July 1980.

B. Bascle, A. Blake, and A. Zisserman. Motion deblurring and super-resolution from an image sequence. In *Proceedings of the 4th European Conference on Computer Vision (ECCV)*, volume 1064 of *Lecture Notes in Computer Science*, pages 573–582, 1996.

R. Basri and D. W. Jacobs. Lambertian reflectance and linear subspaces. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 25(2):218–233, 2003.

P. N. Belhumeur and D. J. Kriegman. What is the set of images of an object under all possible illumination conditions? *International Journal of Computer Vision*, 28(3):245–260, 1998.

J. R. Bergen, P. Anandan, K. J. Hanna, and R. Hingorani. Hierarchical model-based motion estimation. In *Proceedings of the Second European Conference on Computer Vision (ECCV)*, pages 237–252, 1992.

M. Bertero and P. Boccacci. *Introduction to Inverse Problems in Imaging*. Taylor & Francis, 1998.

J. Besag. On the statistical analysis of dirty pictures (with discussion). *Journal of the Royal Statistical Society B*, 48(3):259–302, 1986.

S. K. Bhatia, V. Lakshminarayanan, A. Samal, and G. V. Welland. Human face perception in degraded images. *Journal of Visual Communication and Image Representation*, 6(3): 280–295, September 1995.

C. M. Bishop, A. Blake, and B. Marthi. Super-resolution enhancement of video. In C. M. Bishop and B. Frey, editors, *Proceedings Artificial Intelligence and Statistics*. Society for Artificial Intelligence and Statistics, 2003.

S. Borman and R. L. Stevenson. Spatial resolution enhancement of low-resolution image sequences a comprehensive review with directions for future research. Technical report, University of Notre Dame, Notre Dame, IN, 1998.

S. Borman and R. L. Stevenson. Simultaneous multi-frame map super-resolution video enhancement using spatio-temporal priors. In *Proceedings of the IEEE International Conference on Image Processing (ICIP)*, volume 3, pages 469–473, 1999.

J. W. Brandt. Analysis of bias in gradient-based optical-flow estimation. In *Proceedings of the 28th Asilomar Conference on Signals, Systems and Computers*, volume 1, pages 721–725, 1994.

J. Bride and P. Meer. Registration via direct methods: A statistical approach. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, volume 1, pages 984–989, December 2001.

L. G. Brown. A survey of image registration techniques. *ACM Computing Surveys*, 24: 325–376, December 1992.

V. Bruce and T. Valentine. When a nod's as good as a wink: The role of dynamic information in facial recognition. In M. Gruneberg, P. Morris, and R. Sykes, editors, *Practical aspects of memory: Current research and issues*, pages 169–174. Wiley, Chichester,UK, 1988.

V. Bruce and A. W. Young. *In the Eye of the Beholder*. Oxford University Press, 2000.

P. Cachier and D. Rey. Symmetrization of the non-rigid registration problem using invertion-invariant energies: Application to multiple sclerosis. In *Proceedings of the International Conference on Medical Image Computing and Computer Assisted Intervention (MICCAI)*, pages 472–481, October 2000.

F. M. Candocia. *A Unified Superresolution Approach for Optical and Synthetic Aperture Radar Images*. PhD thesis, University of Florida, May 1998.

F. M. Candocia and J. C. Principe. Super-resolution of images based on local correlations. *IEEE Transactions on Neural Networks*, 10(2):372–380, March 1999.

D. P. Capel. *Image Mosaicing and Super-resolution*. PhD thesis, University of Oxford, 2001.

D. P. Capel and A. Zisserman. Super-resolution from multiple views using learnt image models. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, volume 2, pages 627–634, December 2001.

D. P. Capel and A. Zisserman. Computer vision applied to super resolution. In *IEEE Signal Processing Magazine*, pages 75–86, May 2003.

G. Casella and R. L. Berger. *Statistical Inference*. Duxbury Press, Belmont, California, 1990.

C. Cédras and M. A. Shah. Motion-based recognition: a survey. *Image and Vision Computing*, 13(2):129–155, March 1995.

B. Chalmond. *Modeling and Inverse Problems in Image Analysis*. Springer-Verlag, New York, 2003.

S. Chaudhuri, editor. *Super-Resolution Imaging*. Kluwer Academic Publisher, Boston, 2001.

P. Cheeseman, B. Kanefsky, R. Kraft, J. Stutz, and R. Hanson. Super-resolved surface reconstruction from multiple images. In G. R. Heidbreder, editor, *Maximum Entropy and Bayesian Methods*, pages 293–308. Kluwer Academic Publishers, Dordrecht, the Netherlands, 1996.

G. E. Christensen. Consistent linear-elastic transformations for image matching. In *Information Processing in Medical Imaging*, volume 1613 of *Lecture Notes in Computer Science*, pages 224–237, 1999.

T. F. Cootes and C. J. Taylor. Constrained active appearance models. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, pages 748–754, Apr. 2001.

T. F. Cootes, G. J. Edwards, and C. J. Taylor. Active appearance models. In *Proceedings of the 5th European Conference on Computer Vision (ECCV)*, volume 1406 of *Lecture Notes in Computer Science*, pages 484–498, 1998.

T. F. Cootes, G. Edwards, and C. Taylor. Active appearance models. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 23(6):681–685, 2001.

N. P. Costen, D. M. Parker, and I. Craw. Spatial content and spatial quantisation effects in face recognition. *Perception*, 23:129–146, 1994.

T. M. Cover and J. A. Thomas. *Elements of Information Theory*. Wiley-Interscience, New York, NY, 1991.

R. Cutler and L. S. Davis. Robust real-time periodic motion detection, analysis, and applications. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 22(8): 781–796, 2000.

C. Daniell and R. M. Matic. Neural networks for coefficient prediction in wavelet image coders. In *Proceedings of the International Work-Conference on Artificial and Natural Neural Networks (IWANN), Engineering Applications of Bio-Inspired Artificial Neural Networks*, volume 2 of *Lecture Notes in Computer Science (1607)*, pages 351–360. Springer, 1999.

J. S. DeBonet and P. A. Viola. A non-parametric multi-scale statistical model for natural images. In *Advances in Neural Information Processing Systems (NIPS)*, volume 10. The MIT Press, 1998.

G. Dedeoğlu, T. Kanade, and J. August. High-zoom video hallucination by exploiting spatio-temporal regularities. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, volume 2, pages 151–158, June 2004.

Y. Dufournaud, C. Schmid, and R. Horaud. Matching images with different resolutions. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1612–1618, 2000.

G. J. Edwards, C. J. Taylor, and T. F. Cootes. Interpreting face images using active appearance models. In *Proceedings of the International Conference on Automatic Face and Gesture Recognition*, pages 300–305, June 1998.

A. A. Efros, A. C. Berg, G. Mori, and J. Malik. Recognizing action at a distance. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, pages 726–733, 2003.

P. Ekman and W. Friesen, editors. *The Facial Action Coding System: A Technique for the Measurement of Facial Movement*. Consulting Psychologists Press, San Fransisco, 1978.

M. Elad and A. Feuer. Restoration of a single superresolution image from several blurred, noisy, and undersampled measured images. *IEEE Transactions on Image Processing*, 6 (12):1646–1658, December 1997.

M. Elad and A. Feuer. Super-resolution reconstruction of image sequences. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 21(9):817–834, 1999.

P. E. Eren, M. I. Sezan, and A. M. Tekalp. Robust, object-based high-resolution image reconstruction from low-resolution video. *IEEE Transactions on Image Processing*, 6 (10):1446–1451, October 1997.

C. Fermueller, D. Shulman, and Y. Aloimonos. The statistics of optical flow. *Computer Vision and Image Understanding*, 82:1–32, 2001.

W. T. Freeman and E. C. Pasztor. Learning low-level vision. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, pages 1182–1189, 1999.

W. T. Freeman, E. C. Pasztor, and O. T. Carmichael. Learning low-level vision. *International Journal of Computer Vision*, 40(1):25–47, 2000.

W. T. Freeman, T. R. Jones, and E. C. Pasztor. Example-based super-resolution. *IEEE Computer Graphics and Applications*, 22:56–65, March/April 2002.

B. Girod. What's wrong with mean-squared error? In *Digital images and human vision*, pages 207–220. MIT Press, Cambridge, MA, 1993. ISBN 0-262-23171-9.

R. Gross, I. Matthews, and S. Baker. Generic vs. person specific active appearance models. *Image and Vision Computing*, 23(11):1080–1093, 2005.

B. K. Gunturk, A. U. Batur, Y. Altunbasak, M. I. Hayes, and R. M. Mersereau. Eigenface-domain super-resolution for face recognition. *IEEE Transactions on Image Processing*, 12:597–606, May 2003.

P. W. Hallinan. A low-dimensional representation of human faces for arbitrarylighting conditions. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 995–999, 1994.

B. B. Hansen and B. S. Morse. Multiscale image registration using scale trace correlation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, volume 2, pages 202–208, 1999.

R. C. Hardie, K. J. Barnard, and E. E. Armstrong. Joint MAP registration and high-resolution image estimation using a sequence of undersampled images. *IEEE Transactions on Image Processing*, 6(12):1621–1633, 1997.

L. D. Harmon and B. Julesz. Masking in visual recognition: Effects of two-dimensional filtered noise. *Science*, 180:1194–1197, June 1973.

R. Hartley and A. Zisserman. *Multiple View Geometry in Computer Vision*. Cambridge University Press, 2000.

B. K. P. Horn. Closed-form solution of absolute orientation using unit quaternions. *Journal of the Optical Society A*, 4(4):629–642, April 1987.

B. K. P. Horn and B. G. Schunck. Determining optical flow. *Artificial Intelligence*, 16: 185203, August 1981.

D. H. Hubel. *Eye, Brain, and Vision*. W H Freeman & Co, 1988.

M. Irani and P. Anandan. About direct methods. In *Proceedings of the International Conference on Vision Algorithms: Theory and Practice*, volume 1883 of *Lecture Notes in Computer Science*, pages 267–277, 2000.

M. Irani and S. Peleg. Improving resolution by image registration. *CVGIP: Graphical Models and Image Processing*, 53:231–239, May 1991.

M. Irani and S. Peleg. Motion analysis for image enhancement: Resolution, occlusion, and transparency. *Journal on Visual Communications and Image Representation*, 4(4): 324–335, 1993.

M. Isard and A. Blake. Condensation – conditional density propagation for visual tracking. *International Journal of Computer Vision*, 29(1):5–28, 1998.

K. Jia and S. Gong. Multi-modal tensor face for simultaneous super-resolution and recognition. *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, 2:1683–1690, 2005. ISSN 1550-5499.

K. Jia and S. Gong. Hallucinating multiple occluded face images of different resolutions. *Pattern Recognition Letters*, 27:1768–1775, November 2006.

Z. Jiang, T.-T. Wong, and H. Bao. Practical super-resolution from dynamic video sequenes. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, volume 2, pages 549–554, June 2003.

K. Kanatani. *Statistical Optimization for Geometric Computation: Theory and Practice*. Elsevier Science, Amsterdam, The Netherlands, 1996.

J. K. Kearney, W. B. Thompson, and D. L. Boley. Optical flow estimation: an error analysis of gradient-based methods with local optimization. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 9(2):229–244, March 1987.

S. P. Kim, N. K. Bose, and H. M. Valenzuela. Recursive reconstruction of high resolution image from noisy undersampled multiframes. *IEEE Transactions on Acoustics, Speech, Signal Processing*, 38:1013–1027, June 1990.

A. Kokaram. *Motion Picture Restoration*. Springer Verlag, 1998.

K. Lander, V. Bruce, and H. Hill. Evaluating the effectiveness of pixelation and blurring on masking the identity of familiar faces. *Applied Cognitive Psychology*, 15(1):101–116, 2001.

S. Z. Li. *Markov Random Field Modeling in Image Analysis*. Springer, Tokyo, 2001.

Z. Lin and H.-Y. Shum. On the fundamental limits of reconstruction-based super-resolution algorithms. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, volume 1, pages 1171–1176, 2001.

Z. Lin and H.-Y. Shum. Fundamental limits of reconstruction-based superresolution algorithms under local translation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 26(1):83–97, 2004.

C. Liu, H.-Y. Shum, and C.-S. Zhang. A two-step approach to hallucinating faces: Global parametric model and local nonparametric model. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, volume 1, pages 192–198, 2001.

C. Liu, H.-Y. Shum, and W. T. Freeman. Face hallucination: Theory and practice. *International Journal of Computer Vision*, in press, 2007.

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)*, pages ?–?, 2006.

B. D. Lucas and T. Kanade. An iterative image registration technique with an application to stereo vision. In *Proceedings of the 7th International Joint Conference on Artificial Intelligence (IJCAI)*, pages 674–679, April 1981.

J. B. A. Maintz and M. A. Viergever. A survey of medical image registration. *Medical Image Analysis*, 2(1):1–36, April 1998.

D. Marr. *Vision: A Computational Investigation into the Human Representation and Processing of Visual Information*. W. H. Freeman, 1982.

I. Matthews and S. Baker. Active appearance models revisited. *International Journal of Computer Vision*, 60:135–164, November 2004.

J. Modersitzki. *Numerical Methods for Image Registration*. Oxford University Press, 2004.

H. H. Nagel and M. Haag. Bias-corrected optical flow estimation for road vehicle tracking. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, pages 1006–1011, January 1998.

A. J. O'Toole, D. Roark, and H. Abdi. Recognizing moving faces: A psychological and neural synthesis. *Trends in Cognitive Sciences*, 6:261–266, 2002.

M. Pantic and L. J. M. Rothkrantz. Automatic analysis of facial expressions: The state of the art. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 22(12): 1424–1445, 2000.

A. Papoulis. Generalized sampling expansion. *IEEE Transactions on Circuits and Systems*, 24:652–654, November 1977.

T. N. Pappas and R. J. Safranek. Perceptual criteria for image quality evaluation. In A. Bovik, editor, *Handbook of Image and Video Processing*, pages 669–684. Academic Press, San Diego, 2000.

J.-S. Park and S.-W. Lee. Resolution enhancement of facial image based on top-down learning. In *Proceedings of the First ACM SIGMM international workshop on Video surveillance (IWVS)*, pages 59–64, 2003.

J.-S. Park and S.-W. Lee. Enhancing low-resolution facial images using error back-projection for human identification at a distance. In *Proceedings of the 17th International Conference on Pattern Recognition (ICPR)*, volume 1, pages 346–349, 2004.

S. C. Park, M. K. Park, and M. G. Kang. Super-resolution image reconstruction: A technical overview. In *IEEE Signal Processing Magazine*, pages 21–36, May 2003.

A. J. Patti, M. I. Sezan, and A. M. Tekalp. Superresolution video reconstruction with arbitrary sampling lattices and nonzero aperture time. *IEEE Transactions on Image Processing*, 6(10):1064–1076, 1997.

P. J. Phillips, H. Moon, P. J. Rauss, and S. Rizvi. The FERET evaluation methodology for face recognition algorithms. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 22(10):1090–1104, 2000.

G. E. Pike, R. I. Kemp, N. A. Towell, and K. C. Phillips. Recognizing moving faces: The relative contribution of motion and perspective view information. *Visual Cognition*, 4: 409–438, December 1997.

R. B. Polana. *Temporal Texture and Activity Recognition*. PhD thesis, University of Rochester, 1994.

D. A. Roark, S. E. Barrett, A. J. O'Toole, and H. Abdi. Learning the moves: The effect of familiarity and facial motion on person recognition across large changes in viewing format. *Perception*, 35:761–773, 2006.

P. Rogelj and S. Kovacic. Symmetric image registration. In *Proceedings of Medical Imaging 2003: Image Processing*, volume 5032. SPIE, February 2003.

A. Samal. Minimum resolution for human face detection and identification. In B. E. Rogowitz, M. H. Brill, and J. P. Allebach, editors, *Proceedings of SPIE Symposium on Electronic Imaging: Human Vision, Visual Processing, and Digital Display II*, volume 1453, pages 81–89, June 1991.

H. Schneiderman and T. Kanade. Object detection using the statistics of parts. *International Journal of Computer Vision*, 56:151–177, February 2004.

R. R. Schultz and R. L. Stevenson. Extraction of high-resolution frames from video sequences. *IEEE Transactions on Image Processing*, 5(6):996–1011, 1996.

M. Shah and R. Jain. *Motion-Based Recognition*. Kluwer Academic Publishers, Dordrecht, 1997.

N. R. Shah and A. Zakhor. Resolution enhancement of color video sequences. *IEEE Transactions on Image Processing*, 8(6):879–885, June 1999.

A. Shashua. *Geometry and Photometry in 3D Visual Recognition*. PhD thesis, Massachusetts Institute of Technology, 1992.

E. Shechtman, Y. Caspi, and M. Irani. Increasing space-time resolution in video. In *Proceedings of the 7th European Conference on Computer Vision (ECCV)*, volume 2350 of *Lecture Notes in Computer Science*, pages 753–768. Springer-Verlag Heidelberg, 2002.

H. R. Sheikh and A. C. Bovik. Image information and visual quality. *IEEE Transactions on Image Processing*, 15(2):430–444, 2006.

P. Sinha, B. J. Balas, Y. Ostrovsky, and R. Russell. *Face Processing: Advanced Modeling and Methods*, chapter Face recognition by humans, pages 257–291. Academic Press, December 2005.

O. Skrinjar and H. Tagare. Symmetric, transitive, geometric deformation and intensity variation invariant nonrigid image registration. In *Proceedings of the IEEE International Symposium on Biomedical Imaging: Macro to Nano*, volume 1, pages 920–923, April 2004.

H. Stark, editor. *Image Recovery: Theory and Application*. Academic Press, 1987.

H. Stark and P. Oskoui. High-resolution image recovery from image-plane arrays, using convex projections. *Journal of Optical Society of America, A*, 6(11):1715–1726, November 1989.

M. B. Stegmann. *Generative Interpretation of Medical Images*. PhD thesis, Informatics and Mathematical Modelling, Technical University of Denmark, 2004.

A. Storkey. Dynamic structure super-resolution. In *Advances in Neural Information Processing Systems (NIPS 2002)*, volume 15. The MIT Press, 2003.

J. Sun, N.-N. Zheng, H. Tao, and H.-Y. Shum. Image hallucination with primal sketch priors. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, volume 2, pages 729–736, 2003.

Y. Tian. Evaluation of face resolution for expression analysis. In *Proceedings of the CVPR Workshop on Face Processing in Video (FPIV)*, 2004.

Y.-L. Tian, T. Kanade, and J. Cohn. Facial expression analysis. In S. Z. Li and A. K. Jain, editors, *Handbook of face recognition*. Springer, 2005.

M. E. Tipping and C. M. Bishop. Bayesian image super-resolution. In *Advances in Neural Information Processing Systems (NIPS 2002)*, volume 15. The MIT Press, 2003.

A. Torralba and P. Sinha. Detecting faces in impoverished images. Technical Report AI Memo 2001-028, MIT Artificial Intelligence Laboratory, 2001.

R. Y. Tsai and T. S. Huang. Multi-frame image restoration and registration. In R. Y. Tsai and T. S. Huang, editors, *Advances in Computer Vision and Image Processing*, pages 317–339. JAI Press, Inc., Greenwich, CT, 1984.

H. Ur and D. Gross. Improved resolution from sub-pixel shifted pictures. *CVGIP: Graphical Models and Image Processing*, 54:181–186, March 1992.

T. Vetter and N. F. Troje. Separation of texture and shape in images of faces for image coding and synthesis. *Journal of the Optical Society of America A-Optics Image Science and Vision*, 14:2152–2161, 1997.

C. A. Vogel. *Computational Methods for Inverse Problems*. Society for Industrial and Applied Mathematics, Philadelphia, PA, 2002.

X. Wang and X. Tang. Hallucinating face by eigentransformation. *IEEE Transactions on Systems, Man and Cybernetics, Part C: Applications and Reviews*, 35:425–434, August 2005.

Z. Wang, A. C. Bovik, H. R. Sheikh, and E. P. Simoncelli. Image quality assessment: from error visibility to structural similarity. *IEEE Transactions on Image Processing*, 13(4): 600–612, 2004.

L. Wasserman. *All of Nonparametric Statistics*. Springer-Verlag New York, Inc., Secaucus, NJ, 2006. ISBN 0387251456.

S. Winkler. Issues in vision modeling for perceptual video quality assessment. *Signal Processing*, 78(2):231–252, 1999.

A. W. Young, editor. *Face and Mind*. Oxford University Press, August 1998.

A. L. Yuille, D. Snow, R. Epstein, and P. N. Belhumeur. Determining generative models of objects under varying illumination: Shape and albedo from multiple images using svd and integrability. *International Journal of Computer Vision*, 35(3):203–222, 1999.

W. Zhao and H. Sawhney. Is super-resolution with optical flow feasible? In *Proceedings of the 7th European Conference on Computer Vision (ECCV)*, volume 2350 of *Lecture Notes in Computer Science*, pages 599–613. Springer-Verlag Heidelberg, 2002.

B. Zitova and J. Flusser. Image registration methods: A survey. *Image and Vision Computing*, 21:977–1000, 2003.