Appearance Modeling for 4D Representations

Vagia Tsiminaki

Communauté
UNIVERSITÉ Grenoble Alpes

THÈSE

Pour obtenir le grade de

DOCTEUR DE LA COMMUNAUTÉ UNIVERSITÉ GRENOBLE ALPES

Spécialité : **Mathématiques, Sciences et Technologies de l'Information**

Arrêté ministériel : 7 août 2006

Présentée par

Vagia Tsiminaki

Thèse dirigée par **Edmond Boyer**
et codirigée par **Jean-Sebastien Franco**

préparée au sein d' **Inria Grenoble**
et de **l'école doctorale MSTII : Mathématiques, Sciences et Technolo-gies de l'Information, Informatique**

Appearance Modelling for 4D Multi-View Representations

Thèse soutenue publiquement le **14 Décembre 2016**,
devant le jury composé de :

Prof. Luce Morin
University of Rennes, Rennes, France, Présidente
Prof. Gabriel Brostow
University of London, London, England, Rapporteur
Prof. Hyewon Seo
University of Strasbourg, Strasbourg, France, Rapporteur
Prof. Edmond Boyer
Inria Grenoble, Montbonnot, France, Directeur de thèse
Prof. Jean-Sébastien Franco
INP Grenoble University, Grenoble, France, Co-Directeur de thèse

Κι αν πτωχική την βρεις, η Ιθάκη δεν σε γέλασε.
Έτσι σοφός που έγινες, με τόση πείρα,
ήδη θα το κατάλαβες οι Ιθάκες τί σημαίνουν.

Ιθάκη, Κωνσταντίνος Π. Καβάφης

And if you find her poor, Ithaka won't have fooled you.
Wise as you will have become, so full of experience,
you will have understood by then what these Ithakas mean.

Ithaka, Constantine P. Cavafy

Στους γονείς και στην αδελφή μου
Γιάννη, Καλλιόπη και Κατερίνα

To my parents and my sister
Giannis, Kalliopi and Katerina

Acknowledgements

Looking back at the beginning of this journey I feel the need to mention and thank all these people, who made it such a marvelous experience.

First and foremost, I would like to thank my supervisors, Prof. Edmond Boyer and Prof. Jean-Sébastien Franco, for introducing me to the world of 4D modelling and giving me the opportunity to conduct research in such an interesting field. They provided me with an excellent working environment. I am grateful to them for all the valuable scientific guidance, for the constructive criticism on my work and for the fruitful discussions throughout these years. Most of all, I need to thank them for being my mentors and for giving me the freedom and space to take initiatives through this research study.

I am sincerely honored by Professors Luce Morin, Gabriel Brostow and Hyewon Seo who have accepted to serve as committee members for my thesis defense. Their feedback was important for the final version of this work and their questions were a real source of inspiration for future research ideas.

I would also like to thank all (former and present) members of Morpheo team, Franck Hétroy-Wheeler, Julien Pansiot, Lionel Revéret and Stefanie Wuhrer for their support, their helpful comments and their fruitful scientific discussions. I need to particularly thank Adnane Boukhayma for the excellent collaboration and for sharing the same enthusiasm about brainstorming ideas. Many thanks go to Benjamin Allain, Benjamin Aupetit, Laurence Boissieux, Mickaël Heudre, Vincent Leroy, Thomas Pasquier, Pauline Provine, Romain Rombourg, Li Wang for all the scientific discussions and for all the solutions to technical problems throughout these years. I also need to acknowledge the support of my office mates Abdelaziz Djelouah, Mohammad Rouhani, Aurela Shehu as well as Victoria Fernández Abrevaya, Jinlong Yang. They were always open to discuss ideas, to share the stress, agonies, laughters and every day lab life. A big thank goes to our lab's (former and present) assistants, Laurence Gudyka, Elodie Toihein and Nathalie Gillot, for solving all administrative issues.

This journey would not have been possible without the support of many friends. I would like to thank my friends, Serafeim Perdiki and Alexandro Terzi, who supported me from the very first beginning when I was planning this new step in my life.

Many thanks also go to my French, international friends and of course the Greek "mafia" in Grenoble, who all helped me enjoy these years. I need to particularly thank Benjamin Aupetit, Thomas Pasquier for their introduction to the French cuisine, to Mickaël Heudre for being the best organizer of the weekly croissant day in the lab as well as to Julien Pansiot and Pauline Provine for the bet of running a semi-marathon. It was the beginning of a new challenge in my life.

Acknowledgements

Special thanks go to my friends Maria Amperiadou, Zoi Aftzoglou and Lena Psara, who although dispersed throughout the world stand by me all these years.

Last but not least, I need to thank my sister Katerina Tsiminaki for always reminding me the real values in life, and supporting me practically and psychologically, as well as my parents, Gianni Tsiminaki and Kalliopi Mpilli, who showed me how to keep Ithaka always in my mind.

Grenoble, 14 December 2016 Vagia-Tsiminaki

Abstract

Capturing spatio-temporal models (4D modelling) from real world imagery has received a growing interest during the last years urged by the increasing demands of real-world applications and the tremendous amount of easily accessible image data. The general objective is to produce realistic representations of the world from captured video sequences. Although geometric modelling has already reached a high level of maturity, the appearance aspect has not been fully explored. The current thesis addresses the problem of appearance modelling for realistic spatio-temporal representations. We propose a view-independent, high resolution appearance representation that encodes the visual variability of objects under various movements.

First, we introduce a common appearance space to express all the available visual information from the captured images. In this space we define the representation of the global appearance of the subject. We then introduce a linear image formation model to simulate the capturing process and to express the multi-camera observations as different realizations of the common appearance. Identifying that the principle of Super-Resolution technique governs also our multi-view scenario, we extend the image generative model to accommodate it. In our work, we use Bayesian inference to solve for the super-resolved common appearance.

Second, we propose a temporally coherent appearance representation. We extend the image formation model to generate images of the subject captured in a small time interval. Our starting point is the observation that the appearance of the subject does not change dramatically in a predefined small time interval and the visual information from each view and each frame corresponds to the same appearance representation. We use Bayesian inference to exploit the redundant as well as the hidden non-redundant visual information across time, in order to obtain an appearance representation with fine details.

Third, we leverage the interdependency of geometry and photometry and use it to estimate appearance and geometry in a joint manner. We show that by jointly estimating both, we are able to enhance the geometry globally that in turn leads to a significant appearance improvement.

Finally, to further encode the dynamic appearance variability of objects that undergo several movements, we cast the appearance modelling as a dimensionality reduction problem. We propose a view-independent representation which builds on PCA and decomposes the underlying appearance variability into Eigen textures and Eigen warps. The proposed framework is shown to accurately reproduce appearances with compact representations and to resolve appearance interpolation and completion tasks.

Keywords: appearance modelling, multi-view 4D representations, 4D video, Super-Resolution, texture mapping, super-resolved texture map, joint geometry-appearance estimation, photometry-geometry decoupling, appearance variability, dimensionality reduction, Eigen Appearance Maps

Résumé

Ces dernières années ont vu l'émergence de la capture des modèles spatio-temporels (modélisation 4D) à partir d'images réelles, avec de nombreuses applications dans les domaines de post-production pour le cinéma, la science des sports, les études sociales, le divertissement, l'industrie de la publicité. A partir de plusieurs séquences vidéos, enregistrées à partir de points de vue variés, la modélisation 4D à partir de vidéos utilise des modèles spatio-temporels pour extraire des informations sur la géométrie et l'apparence de scènes réelles, permettant de les enregistrer et de les reproduire. Cette thèse traite du problème de la modélisation d'apparence.

La disponibilité des donnée d'images offre de grands potentiels pour les reconstructions haute fidélité, mais nécessite des méthodes plus élaborées. En outre, les applications du monde réel nécessitent des rendus rapides et des flux réduits de données. Mais l'obtention de représentations d'apparence compactes, indépendantes du point de vue, et à grande résolution est toujours un problème ouvert.

Pour obtenir ces caractéristiques, nous exprimons l'information visuelle de l'objet capturé dans un espace de texture commun. Les observations multi-caméra sont considérées comme des réalisations de l'apparence commune et un modèle linéaire est introduit pour matérialiser cette relation. Le modèle linéaire d'apparence proposé permet une première étude du problème de l'estimation d'apparence dans le cas multi-vue et expose les sources variées de bruit et les limitations intrinsèques du modèle.

Basé sur ces observations, et afin d'exploiter l'information visuelle de la manière la plus efficace, nous améliorons la méthode en y intégrant un modèle de super-résolution 2D. Le modèle simule le procédé de capture d'image avec une concaténation d'opérations linéaires, générant les observation d'image des différents points de vue et permettant d'exploiter la redondance. Le problème de super-résolution multi-vue résultant est résolu par inférence bayésienne et une représentation haute-résolution d'apparence est fournie permettant de reproduire la texture de l'objet capturé avec grand détail.

La composante temporelle est intégrée par la suite au modèle pour permettre d'y recouper l'information visuelle commune sous-jacente. En considérant des petits intervalles de temps ou l'apparence de l'objet ne change pas drastiquement, une représentation super-résolue cohérente temporellement est introduite. Elle explique l'ensemble des images de l'objet capturé dans cet intervalle. Grâce à l'inférence statistique Bayésienne, l'apparence construite permet des rendus avec une grande précision à partir de point de vue nouveau et à des instants différent dans l'intervalle de temps prédéfini.

Pour améliorer l'estimation d'apparence d'avantage, l'inter-dépendance de la géométrie et de la photométrie est étudiée et exploitée. Les modélisations de la géométrie et de l'apparence sont

unifiées dans le framework de super-résolution permettant une amélioration géométrique globale, ce qui donne à son tour une amélioration importante de l'apparence.

Finalement pour encoder la variabilité de l'apparence dynamique des objets subissant plusieurs mouvements, une représentation indépendante du point de vue s'appuyant sur l'analyse en composantes principales est introduite. Cette représentation décompose la variabilité sous-jacente d'apparence en texture propres et déformations propres. La méthode proposée permet de reproduire les apparences de manière précise avec des représentation compactes. Il permet également l'interpolation et la complétion des apparences.

Cette étude montre que la représentation compacte, indépendante du point de vue, et super-résolue proposée permet de confronter les nouvelles réalités du problème de modélisation d'apparence. Elle représente un contribution vers des représentations d'apparence 4D haute-qualité et ouvre de nouvelles directions de recherche dans ce domaine.

Contents

List of Figures

1 Introduction

Recent years have seen the emergence of capturing spatio-temporal models (4D modelling) from real world imagery. These spatio-temporal models (4D models) comprise geometric and appearance information allowing for static as well as dynamic scenes to be recorded and reused and they have thus taken center stage in applications which require real 3D content for analysis, free viewpoint and animation purposes (see Figure 1.1).

To achieve highly realistic representations, both geometry and appearance need to be replicated with great details. To that goal, efficient shape representations have been proposed to model geometric information of static subjects. Temporal coherent shape representations, which evolve and deform over time, have been introduced to model moving subjects. In dynamic scenarios, where multiple subjects perform several motions, shape spaces that encode both pose and subject variabilities have also been studied. While geometry has been largely covered, appearance has received less attention.

In this dissertation, we focus on the appearance aspect of 4D modelling. In particular, we address the problem of retrieving visual information through real world imagery and proposing a high quality appearance representation.

1.1 Motivation and Challenges

The emergence of multi-view capturing systems and the increasing demands of real-world applications that require 3D contents, have introduced a new reality. The ability to record datasets of subjects undergoing several movements, yields a tremendous amount of image data. On one hand, this easily accessible visual information yields high expectation for the quality of appearance. More visual information should arguably drive to more accurate appearance modelling. On the other hand, it contributes to a higher complexity of the conditions of the problem. More measurements correspond to more sources of noise and increase the sensitivity of the estimation process. Furthermore, demands of real-world applications on compactness, speed and quality expand continuously and the proper trade-off needs to be introduced. In overall new conditions in the area of appearance modelling have been established and there is arguably room for further investigation.

In this new context, state-of-the-art works appear to have significant shortcomings. Traditionally, view-dependent texture mapping methods (Debevec et al. [1996], Buehler et al. [2001],

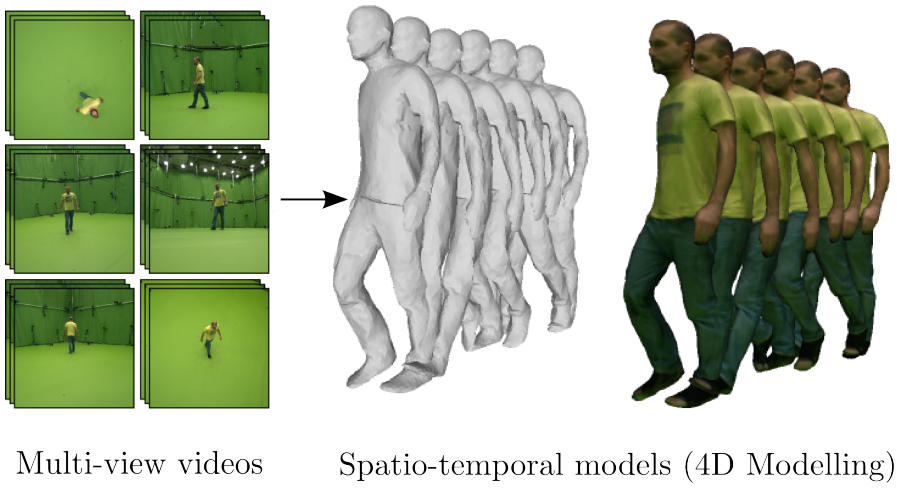

Figure 1.1 – Capturing spatio-temporal models from real world imagery. For realistic representations both geometry and appearance need to be replicated with great details.

Cobzaş et al. [2002], Eisemann et al. [2008], Takai et al. [2010], Jagersand et al. [2010]) aim to assign appearance value to the 3D surface by averaging the color contributions based on viewpoint. These methods are mainly limited by viewpoint selection and are not initially designed to store intrinsic appearance details. To eliminate viewpoint dependency and the associated shortcomings, multi-view texture estimation methods (Burt and Adelson [1983], Lensch et al. [2001], Theobalt et al. [2007], Lempitsky and Ivanov [2007], Allene et al. [2008], Janko and Pons [2009]) build an image-based texture atlas to store appearance information in a more compact way. Although these methods succeed to provide compact and view-independent representations, there is still a lack in quality with respect to the redundancy with which the appearance of subjects is observed. It is only recently that some works (Tung [2008], Goldluecke and Cremers [2009b,a]) aspire to increase visual quality by adopting image enhancement methods and in particular Super-Resolution techniques. Although these methods aim to take full advantage of the available visual information, the related geometric information is not globally integrated, which should intuitively prove beneficial. A high resolution, compact appearance representation is thus still missing. The demand becomes even more prominent when we consider the temporal component. In dynamic scenarios, where the subject undergoes several movements, the appearance variability increases. Despite the fact that recent works (Casas et al. [2014], Volino et al. [2014], Casas et al. [2015]) attempt to exploit the temporal and spatial visual redundancy, they primarily aim at animation synthesis using mesh data and do not propose a global appearance model for sequences. The leap to more efficient appearance representations is thus still a largely open problem.

This direction is indeed a great challenge. The available video sequences are down-sampled, degraded observations of the appearance and a naive blending method can lead to blurry results with larger misalignments. Existing inaccuracies in the geometric reconstruction, in the synchronization of cameras and in the camera calibration increase even more the complexity. Additionally, considering large time intervals the underlying appearance variability increases due

to lighting conditions and due to motion changes. A proper model is thus required to account for all these highly challenging conditions.

In this thesis, we elaborate on these issues and aspire to bridge some of the gaps. Our research goal is to make the leap to appearance representations that will take full advantage of the massive amount of visual information and achieve a proper trade-off among storage size, speed and quality.

1.2 Thesis Outline

To tackle the problem of appearance modelling, we consider that a subject is captured in several movements and from multiple viewpoints by cameras whose position, orientation and intrinsic parameters are known and that the geometric representation is also given. The goal is to optimally leverage the available video sequences and provide a compact, view-independent, high resolution appearance representation.

We first introduce the common appearance space, where the unknown appearance is defined and we then propose a linear image formation model to simulate the capturing process. This generative model aims to explain the observed captured images and it is progressively improved to account for existing inaccuracies in the multi-view setup. The task of computing the appearance is cast as Bayesian inference. We further tackle the dynamic appearance variability modelling of objects undergoing several movements by casting it as a dimensionality reduction.

A brief outline of the chapters follows:

- Chapter 2 - State of the art.

  This chapter reviews previous techniques on appearance modelling and aims to identify the main limitations of the literature. View-dependent texturing methods are examined as a first category that computes the appearance at a given view-point by blending image information from the neighboring views. Multi-view texture estimation methods are then presented as an alternative to eliminate viewpoint dependency. To attack the problem of low visual quality, works that apply on appearance modelling image enhancement techniques and in particular Super-Resolution techniques and unify this step with the additional geometric modelling, are also reviewed. To optimally handle the dynamic appearance variability, methods on efficient representations of the dynamic nature of appearance are also presented.

- Chapter 3 - Linear Image Formation Model.

  This chapter provides a view-independent, compact appearance representation and a linear image formation model to express the multi-camera observations as different realizations of the common global appearance. Solving for the unknown appearance is cast as a problem of computing the contribution of each captured image into the common space and examining the central tendency of all contributions.

- Chapter 4 - Super-Resolved Appearance Model.

  This chapter presents a novel Super-Resolution appearance model to overcome the inherent

limitations of the previous simplified model. In particular, we propose a compact representation which builds on the linear appearance model and successfully enhances it with the Super-Resolution (SR) fundamentals integrated in the multi-view context. Quantitative and qualitative evaluation on several datasets under various scenarios demonstrate the high visual quality of the estimated appearance.

- Chapter 5 - Temporally Coherent Super-Resolved Appearance Model.

  This chapter proposes a temporally coherent super-resolved appearance model, which generalizes the previous multi-view appearance Super-Resolution work to the temporal domain and regains higher quality. The appearance of a subject is assumed to remain the same in a predefined time interval, where video sequences at multiple viewpoints are captured. Exploiting the appearance information of the subject that is not only shared among the views but also is fused in time in order to compute a detailed appearance is the fundamental of the proposed model. Experimental validation demonstrates the additional quality that is regained through the temporal case.

- Chapter 6 - Geometry and Appearance through a Super-Resolution Framework.

  This chapter introduces a Super-Resolution framework to accommodate both geometric and appearance modelling. Driven by the fact that jointly estimating appearance and geometry should prove mutually beneficial and driven also by the success of the already presented model to compensate locally geometric inaccuracies, we extend the image formation model to integrate geometry as an additional variable. Qualitative experimental evaluation demonstrates that it is possible to achieve higher visual quality of appearance by jointly optimizing both parameters.

- Chapter 7 - Application: Dynamic Appearance Representations.

  This chapter accounts for larger time intervals and it proposes a view-independent representation which builds on PCA and decomposes the underlying appearance variability into Eigen textures and Eigen warps. In dynamic scenarios where the subject undergoes several movements, the appearance can change dramatically. We thus study the evolution of the appearance in the time domain to reveal the intra correlation of the observations. Appearance variations due to viewpoint, illumination changes and also due to object pose and dynamics are encoded through the Eigen textures while the visual variability due to geometric modelling imprecisions is encoded through the Eigen warps. The framework succeeds to provide accurate appearances and resolve interpolation as well as completion tasks.

1.3 Contributions

The main contributions of this thesis are:

- **Compact, view-independent and super-resolved appearance representation.**

  We propose a compact representation of the appearance and a Super-Resolution framework to retrieve fine visual details. The proposed model exploits the non redundant appearance

information of video sequences at multiple viewpoints and provides a unique, view-independent, high resolution appearance. It allows for the object to be rendered not only at given viewpoints but also at new ones with greater details. Its compact representation leads to a significant decrease of storage requirements as well to an increase of transmission speed.

- **Temporally coherent, view-independent and super-resolved appearance representation.**

  We introduce small time intervals, where the appearance of the object does not change dramatically and we propose a model to exploit the non redundant information fused over time. It allows to extend the previous benefits to the temporal domain. The object can be rendered with higher accuracy at new viewpoints and at time instances in the predefined time interval, which also translates into a decrease of storage.

- **Geometric enhancement in conjunction with super-resolved appearance.**

  We unify geometric and appearance modelling into a Super-Resolution framework. This allows for a global geometric enhancement which further leads to a significant improvement of the super-resolved appearance.

- **Compressive and view-independent appearance representation for significant temporal appearance variability.**

  We provide a representation to encode the appearance variability of a subject undergoing several movements. It reduces the high dimensionality of the problem allowing for compression and it successfully resolves interpolation in synthesizing new poses and completion tasks for original poses.

2 State of the art

Modelling appearance and rendering images at novel view points have received a lot of attention from both communities of computer vision and graphics. Although image based modelling and rendering use mainly images as primitives, previous surveys have suggested to characterize a technique based on how image centric or geometric centric it is, yielding the image-geometry continuum of image-based representations (Lengyel [1998], Shum and Kang [1999]). According to it, rendering methods are classified into three categories: 1) rendering with no geometry; 2) rendering with implicit geometry; 3) rendering with explicit geometry. In the current thesis, we consider that a geometric representation is available and we thus make explicit use of this modality in conjunction with the captured images.

For the sake of a complete overview of the existing methods however, representatives of each category are briefly discussed. At one extreme of the continuum, there are methods that use only the captured images as observed discrete samples to characterize the plenoptic function (Adelson and Bergen [1991]). In the middle of the continuum lie techniques (Chen and Williams [1993], Laveau and Faugeras [1994], Seitz and Dyer [1996]) which use implicit geometry in the form of feature correspondences between the images and new views are generated by manipulating these correspondences. At the other end of the rendering spectrum, techniques use directly 3D information either in the form of depth or in the form of 3D coordinates. An image is rendered from any nearby viewpoint by projecting the pixels to the 3D locations and then re-projecting them onto the new viewpoint. 3D warping techniques (Mark et al. [1997]) and layered depth image rendering (Shade et al. [1998]) use depth information while texture mapping techniques use directly 3D coordinates. In a multi-view setup, efficient shape representations (multi-view reconstruction algorithms Seitz et al. [2001, 2006], Furukawa and Ponce [2010]) have been proposed to obtain such 3D models with fine geometric details and 3D reconstruction algorithms have been introduced in the dynamic case to provide a temporal coherent model which evolves and deforms over time.

Although several works have been devised towards this direction of appearance modelling with explicit use of geometric information, there is still space for further investigation. A new reality has been created with requirements that cannot be met anymore by the current state-of-the-art. First, the ability to record datasets of subjects leads to a plethora of easily accessible captured images. Second, there is an increasing demand of applications on high realism, low cost and high performance. In the following of this chapter we review previous studies to report limitations

and identify the main key variables that need to be targeted for their alleviation. In particular, in Section 2.1 view-dependent texturing methods are presented which compute at a given viewpoint the appearance of the object. Multi-view texture estimation methods are discussed in Section 2.2 as the next step to overcome the shortcomings of the previous methods. Aiming at high quality results, we review in Section 2.3 methods which adopt image enhancement techniques and in particular Super-Resolution techniques in the case of multi-view setup. We then study in Section 2.4 how geometry and appearance are jointly modelled in the context of appearance modelling. Finally, to consider more challenging dynamic scenarios we review in Section 2.5 works that aim to exploit and encode temporal variability.

2.1 View-Dependent Appearance Modelling

View-dependent texture mapping methods assign texture value to the 3D surface by averaging the color contributions based on the viewpoint relative to the sampled viewpoints or applying more elaborate weighting schemes.

Debevec et al. [1996] reproject and blend view contributions according to visibility and viewpoint-to-surface angle. More recently, view-dependent techniques have been generalized to model and approximate the plenoptic function for the scene object, capturing view dependent shading effects based on relative angular position, resolution and field of view (Buehler et al. [2001]) but this requires many dense views. Since these blending techniques rely on geometry, any inaccuracy of the reconstructed 3D model results in misalignments. These ghosting effects are further compensated by additional steps introduced in numerous techniques. Eisemann et al. [2008] combine an interpolation step and an optical flow-based warping refinement step to correct for every local misalignment. Images are back-projected on the geometric proxy and then projected at a desired viewpoint, between which, optical flow is computed to correct local displacements. A weighting and visibility scheme is furthermore introduced to account for the occlusions and thus to eliminate the related artifacts Fig. 2.1. Based on the same principle of compensating for local displacements, Takai et al. [2010] optimize the 3D mesh with respect to shape and texture consistency. As a second step, textures are deformed given the refined mesh and they are blended according to a weighting scheme to generate new images at virtual points. An interesting idea to capture the intensity variability and model new poses of an object is presented in Cobzaş et al. [2002]. A coarse geometry is firstly reconstructed and used to subdivide the scene into texture patches, whose intensity variability is captured using a spatial basis. By modulating the texture basis, new patches are synthesized and the generated texture is then back-projected into the geometry. Similar in spirit is their later work (Jagersand et al. [2010]), where they propose to represent the appearance variabilities caused by geometric inaccuracies, by parallax and illumination variations not on surface patches but in the image domain. A differential texture basis is built to encode geometric and photometric variations. Each image is composed from modulating vector, applied on this basis. Images at new view points are then synthesized by interpolating the modulation vectors from the nearest sample views and being applied on the basis.

All the aforementioned methods study how to generate images at given view points by making

Figure 2.1 – Comparison of standard linear interpolation using the Unstructured Lumigraph weighting scheme (left) and the Floating Textures approach (right) for similar input images and visual hull geometry proxy. Ghosting along the collar and blurring of the shirt's front, noticeable in linear interpolation, are eliminated on-the-fly by Floating Textures (Eisemann et al. [2008]).

use of the existing appearance information from the closest views. Weighting schemes and more elaborate techniques are proposed to overcome the ghosting effects that are caused by geometric inaccuracies. Although the provided appearance is of good quality, important features are not fully exploited. First, in the large captured image datasets, there are appearance properties which are view independent and are not encoded in the proposed models (the fact that appearance consists most often of smooth areas, which are interrupted by strong edges). Second, the view-dependent texture mapping methods do not leverage the encoded non redundant visual information to super-resolve visual quality, despite the fact that they make use of the appearance redundancy through naive and more elaborate weighting schemes among views. Third, they do not easily extend to the time domain for deformable objects in terms of storage and they do not exploit the additional appearance non-redundancy in favor of the quality.

2.2 Multi-View Texture Estimation

The first limitation of the above category, which is mainly related to the view point dependency and the lack of a compact way of representing the appearance information is partially eliminated by multi-view texture estimation approaches. To store intrinsic details of the acquired object and later render them, numerous methods build an image-based texture atlas, which models appearance information by blending for each texel contributions from each view.

These techniques do not come without their own difficulties. Ghosting effects are still inevitable and realignment takes often place to compensate for them (Lensch et al. [2001], Theobalt et al. [2007]). In the work of Theobalt et al. [2007] input video sequences are firstly warp-corrected to account for photo-inconsistencies due to inexact body geometry and then they are projected into the texture domain to generate the multi-view video textures. Aiming

to compute a distortion-free texture map Lempitsky and Ivanov [2007] introduce the idea of mosaicing texture fragments into the texture. Instead of considering a common reference frame and constructing parametrization for objects to blend texture fragments they define the texture as a mosaic of unique-view contributions. Through this representation, they avoid the quality degradation due to distortions of geometrically complex parameterizations. The resulted seams among the fragments are handled through the seam levelling procedure, which takes place as a second step. Upgrading this idea of partitioning the surface into patches so as to optimize color continuity and visual details at boundaries, Allene et al. [2008] propose a two step approach. The first step follows the work of Lempitsky and Ivanov [2007] using graph cut optimization to partition the surface into patches. And the second step is inspired by the work of Burt and Adelson [1983], which proves that multi-band blending is effective for 2D image mosaicing. Extending this idea, they introduce Laplacian pyramids of the input images as the multi-band decomposition, which they approximate by differences of Gaussians. The color at particular pixel is then computed as a combination of the defined bands.

Similarly, this strategy was extended to the temporal domain in the work of Janko and Pons [2009]. Through this direction which is closer to our proposed method, additional appearance information available in time is further exploited in favor of the quality. The dynamic scene is represented as a mesh with fixed connectivity and with constant reflectance properties through time and the spatio-temporal surface of the scene is partitioned into patches, associated to input images. In particular, a labeling function assigns to each face of the animated mesh input images by optimizing the visual quality while minimizing the visibility of seams. This way a high quality spatio-temporal texture atlas is computed, which allows the synthesis of novel views with an optimal visual quality and with minimally visible color discontinuities and flickering artifacts. Through this spatio-temporal representation, on one side the amount of texture data is decreased resulting in an enhancement of the portability and the rendering efficiency and from the other side appearance redundancy is exploited to recover from grazing views, highlights and occlusions. In Figure 2.2, rendered images at novel views demonstrate that the additional appearance information in temporal domain succeeds to cover temporarily hidden regions and thus results in smaller untextured regions. The proposed 'multi-frame optimized' method succeeds to decrease the visibility discontinuities by introducing an energy function which globally optimizes the surface partition. Due to this global energy optimization function the proposed method is robust to any geometric inaccuracy, which often appears in the multi-view setup and which by naively assembling texture atlases from input images leads to strong misalignments (see 'multi-frames' column of Figure 2.2).

Although the general approach is very interesting with remarkable results, there are inherent limitations which need to be further studied. First, the proposed method is based on the assumption of constant reflectance, which in turn builds on the fact that fine geometric representation is computed. When such a detailed geometric reconstruction is not available, which can be often the case, the assumption is violated and appearance variations are averaged out. Second, although the available appearance information is used, it is not optimally exploited to model visual details.

In overall, through the multi-view texture estimation approaches, intrinsic appearance details of objects are explored and stored to later render them. Although the extension to the temporal

Figure 2.2 – Comparison of static ('single-frame' and 'single-frame optimized') and temporal case ('multi-frame' and 'multi-frame optimized'). The 'single-frame' maps each face to the highest quality input image and the 'single-frame optimized' achieves a trade-off between visual quality and visibility of seams similarly to (Allene et al. [2008], Lempitsky and Ivanov [2007]). Analog to the static one, the 'multi-frame' method maps each face to the highest quality input image among viewpoints and time frames, whereas the 'multi-frame optimized' accomplishes visual quality and visibility of seams according to the method of Janko and Pons [2009].

domain results in larger amount of available texture information and thus in more realistic renderings, visual redundancy is not exploited to super-resolve the quality. In the current thesis, we propose an improved, unified model to deal with geometric variability due to reconstruction error and small deformation across time for multi-view super-resolution. We do not make a simple use of the available visual information but we exploit it in such a way so as to unfold the hidden non-redundant information and thus regain quality.

2.3 Super-Resolution

Such high quality appearance results are achieved through Super-Resolution methods, from which we get inspiration to introduce our appearance model and estimation algorithm. Whilst the Super-Resolution techniques are still not broadly applied in a multi-view context, the problem has been extensively studied in the monocular case.

The aim of these techniques is to recover additional details lost in the acquisition process and to this goal multiple observations, which differ in sub-pixel precision, need to be available. These techniques define the image formation model as a geometric warping, blurring and sub-sampling process of the initial high-resolution image (Baker and Kanade [2002]). In Section 4.2 a detailed review of the 2D Super-Resolution is presented to exhibit the underlying main principles and common techniques and thus to sustain our proposed model. In particular, the key-points which drive us to propose our Super-Resolution appearance model are the same goal of high visual quality, the availability of multiple observations, the fundamental technical aspect of SR to express the image generation process as a concatenation of operations and the fact that we can introduce view-independent appearance information through the model. Notably, super-resolving multiple videos of a moving subject was examined in a performance capture context, but only for the input viewpoints (Tung [2008]). The authors combine Super-Resolution on multi-view images and on single-view video sequences with dynamic 3D shape reconstruction through a MRF formulation.

Only a handful of particularly relevant works examine how to super-resolve fine appearance details from viewpoint redundancy at a single time frame (Goldluecke and Cremers [2009b,a], Goldluecke et al. [2013]). In Goldluecke and Cremers [2009b], a Super-Resolution image formation model is presented, to express the captured images as observations of the unknown appearance. Through a conformal atlas for the unknown geometry, the desired texture map is solved via total variation deblurring on planar 2D texture space. In Figure 2.3 fine visual details are successfully recovered compared to input image and to simple averaging schemes.

In this thesis, we propose a model which adopts the Super-Resolution principles and extends it in the multi-view case. As opposed to previous work of Tung [2008], which enhances input views directly, our model builds one common super-resolved appearance, available to be re-used in the future. Through this representation, great visual details are regained and in addition view-independent intrinsic appearance properties are expressed. Compared to Goldluecke and Cremers [2009b], our method accounts for every source of noise uniformly through a linear representation of the image generation process and thus exhibits high quality visual results.

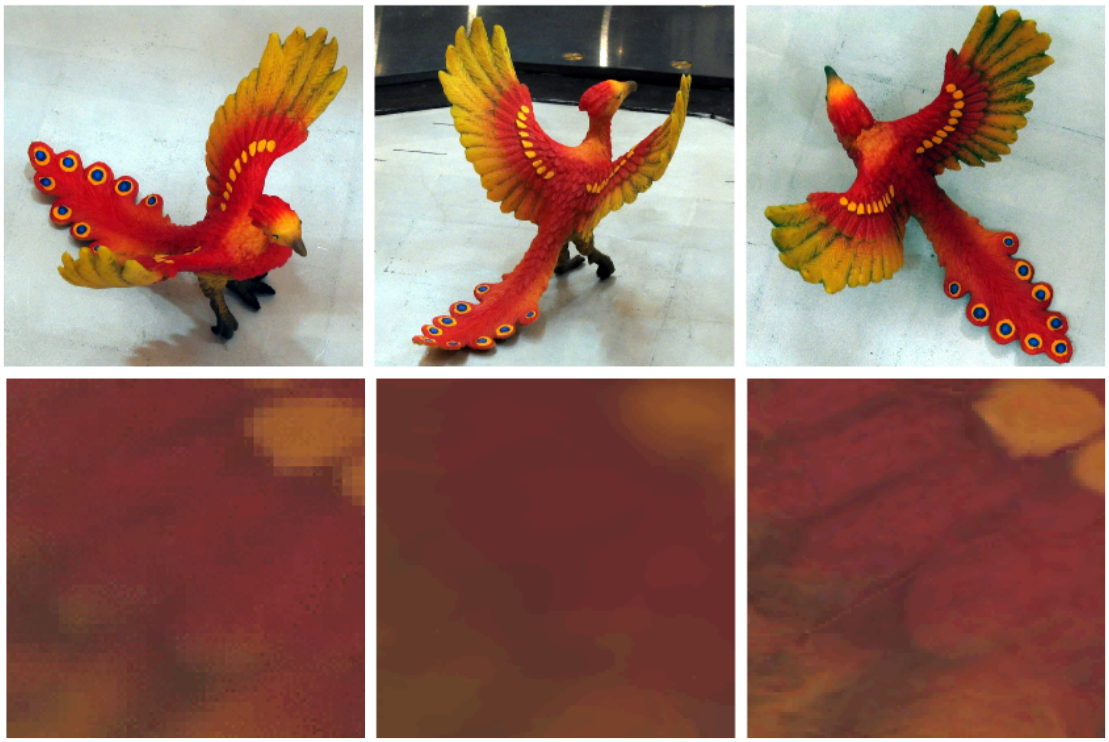

Figure 2.3 – First row: Input images for a multi-view reconstruction. Second row: Close-up of one of the low-resolution input images. Rendered model with blurry texture initialized by weighted averaging of input images. High-quality texture optimized with the proposed super-resolution approach (Goldluecke and Cremers [2009b]).

2.4 Joint Appearance and Geometry Modelling

Investigating how appearance and geometry as separate modalities can be combined and exploited has drawn the attention of several researchers. Firstly, we present works, which model the photometric aspect in conjunction to the geometric one and we discuss methods which acknowledge the fact that fusing observations from different sensors can boost the accuracy of both geometry and appearance. In this context of simultaneously modelling geometry and appearance, several works have also inserted the notion of Super-Resolution. This direction is of particular interest to our study and we further examine it.

Advocating that geometric and photometric aspects are intertwined, state of the art works have aspired to model geometry in conjunction to appearance. In particular the work of Delaunoy and Pollefeys [2014] shows that notable improvements in geometry and appearance are achieved by minimizing a photometric error with respect to shape and calibration parameters. In the same context but with a different approach, Aubry et al. [2011] manifests the necessity of decoupling photometry and geometry in the estimation of camera parameters. The authors propose an iterative scheme, where through a relaxation step they firstly solve for the geometric error by minimizing a photometric error and then solve for the camera parameters which best explain the

computed geometric error. At the end of their iterative scheme, the refined camera parameters lead to significant improvement of the appearance as well as of the 3D model. Although these techniques provide a notable enhancement of the appearance, the regain of visual quality comes as a side-effect of the geometric refinement which appears in the form of shape or of calibration parameters. Focusing more on the visual aspect, in the work of Pujades et al. [2014] geometry is integrated in estimation through an error term in the objective function. They compute the appearance at novel views, by minimizing an energy function whose first data term comprises both sensor and geometric error and whose second term acts as regularization term. The key component of their method is the fact that geometry is integrated in the estimation process by accounting for the error in the estimated geometric proxy.

The fact that both sources of information boost each other's accuracy manifests in the emergence of studies based on data fusion. Although in this thesis the input observations are visible images, it is interesting to mention few of the works which combine multiple sensors to achieve high quality reconstructions. In the work of Morris et al. [2001] combining laser altimetry and visible images into a single high resolution surface model, the authors succeed to reconstruct accurate geometry and albedo. In the work of Cabezas et al. [2014] the authors propose an integrated probabilistic model for multi-model fusion of imagery, LIDAR data and GPS measurements to reconstruct large aerial areas. Through a joint probabilistic framework both geometry and appearance are inferred given a smaller number of input observations and enabling at the same time reconstruction of fine details.

To boost even more the quality of the results in the context of simultaneous geometric and visual modelling, studies apply the idea of Super-Resolution to 3D visual reconstruction. In such 3D settings, Super-Solution techniques are introduced for a first time in the seminal work of Berthod et al. [1994], where the albedo as well as the depth of a scene are reconstructed from a set of low resolution images. The authors proposed an iterative scheme to reconstruct high-resolution depth maps in the first step and high-resolution images in the second step. In their later work of Shekarforoush et al. [1995] they explicitly refer to the 3D super-resolution algorithm, where they simultaneously solve for depth and albedo. The focus of these previous works is super-resolving depth maps and as such there is less attention on the advantages of using this information in conjunction to the super-resolution techniques to improve image quality.

This notion of multi-objective super-resolution technique is introduced by the work of Rajan et al. [2003], which addresses the fact that super-resolution and scene geometry could be approached simultaneously and hence they propose a technique to extract 3D information from the low resolution observations so as to super-resolve them. More recently, Bhavsar and Rajagopalan [2010] acknowledge the fact that image Super-Resolution and high-resolution depth estimation are intertwined and they introduce an integrated approach to estimate jointly these quantities. Through a Markov Random Field (MRF) formulation, the authors provide the MAP point estimates which they compute through the equivalent energy minimization. They follow an iterative strategy to estimate firstly the depth via graph cuts and then image via iterated condition modes (ICM). In line with this direction, Park et al. [2012] propose also a unified framework to solve for both problems of multi-view stereo and super-resolution. They introduce an energy function to impose consistency at the same viewpoint between high and low resolution image as well among

Figure 2.4 – From left to right. Input mesh and frame. Super-resolution 3D video reconstruction. (Tung [2008]).

high resolution images at different view points. They adopt an iterative scheme to alternate between the high resolution estimated and super-resolved depth maps using a tree-reweighted message passing algorithm (TRW-S). In their later work of Seok Lee and Mu Lee [2013] the authors restrict the conditions more using only a single camera and they focus on the simultaneous estimation of depth and image. Kimura et al. [2007] utilize multiple low-resolution stereo images to estimate high resolution 3D information and images. The authors incorporate into a unified cost function both subproblems and they provide a Maximum a Posteriori (MAP) framework. In the work of Tung [2008] an iterative coarse to fine approach is presented to combine both super-resolution surface texture and shape refinement in a single MRF framework. In Figure 2.4 it is shown that both the reconstructed mesh and the super-resolved frame gain in details.

Interesting but from a different perspective are the works, which achieve high quality results by applying the Super-Resolution principles on the geometry. Although it is not well-aligned with the scope of this study we present these methods to serve a complete overview of the literature. In the works of Cheeseman et al. [1996] the authors super-resolve the geometry through a Bayesian framework. The surface is represented by a discrete uniform grid and surface properties are defined at each point. To reconstruct this surface grid, the authors exploit the differences between the values of the observed images. Due to the misalignments of the input observations which correspond to additional information, the surface can be computed at a finer resolution than that of the observed images. In the work of Morris et al. [1999] the surface is parameterized by a set of heights and a set of albedos. Triangulation of these heights which form the surface as well as lighting conditions and camera parameters are known and the surface model parameters are estimated in a finer resolution than the one of input images though a Bayesian framework. Restricting the application domain and specializing on 3D human faces, recent works apply Super-Resolution on triangle-meshed human face models resulting in high resolution meshes. In Berretti et al. [2012] low resolution 3D scans are used as input observations and Super-Resolution is materialized through three main processing modules,namely the face detector, the face registration and the face sampler. At the end of the pipeline high resolution 3D models are reconstructed.

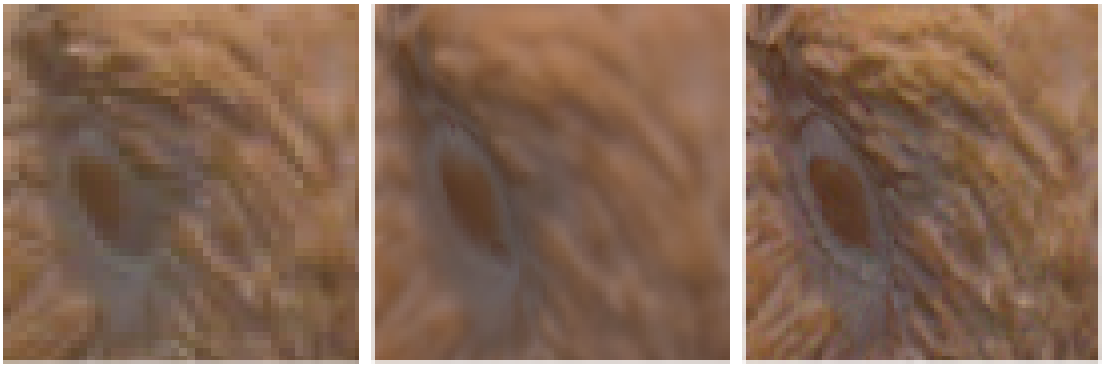

Figure 2.5 – From left to right: Close-up of one of the low-resolution input images. Rendered model with texture initialized by weighted averaging of input images. High-quality texture computed through the framework that optimizes for texture, displacement map on the surface and calibration parameters proposed in Goldluecke et al. [2013].

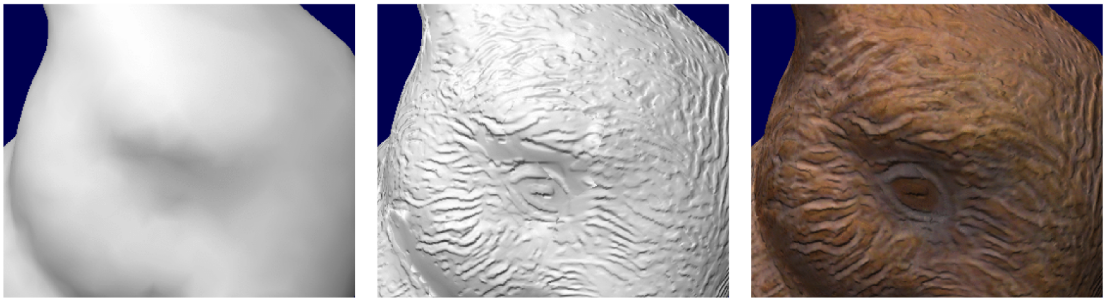

Figure 2.6 – From left to right: Rendering with Gouraud shading. The underlying geometry has low geometric detail. Normal map lighting showing improved geometric detail from the estimated displacement map. Rendering with Super-Resolution texture and normal map lighting proposed in Goldluecke et al. [2013].

The works of Goldluecke and Cremers [2009a] and Goldluecke et al. [2013] resemble our proposed method in the sense that the surface is parametrized in order to express the appearance in the form of texture atlas. In the work of Goldluecke and Cremers [2009a] geometry is integrated into the framework in the form of a displacement map defined on the normals of the surface and it is estimated jointly with the appearance. In their latter paper Goldluecke et al. [2013] the geometric aspect is further extended. In addition to the normal displacement and the appearance the authors optimize for camera calibration parameters. In Figure 2.5 great details are recovered. The refinement of the geometry through their proposed method is visible in Figure 2.6.

Although it has been acknowledged in the literature that geometry and appearance estimation problems are interrelated, it is still unclear how the modelling of both modalities can interfere with each other in a general Super-Resolution framework. In this thesis we approach the problem of appearance estimation through a Super-Resolved appearance model by explicitly using the geometric information in the beginning and by introducing further the geometric modelling as

additional step. As opposed to previous work of Goldluecke et al. [2013] through our compact appearance model we allow for an update, which best explains the image observations. Through our image generation model we succeed to decouple both modalities, study each of them and accommodate them in a simple manner.

2.5 Dynamic Appearance Representations

Time component plays an essential role in the problem of appearance modelling. Considering large time intervals, the appearance of the object of interest can change due to lighting conditions and due to motion changes. To capture this dynamic nature of the appearance in a compact way so as to be later reused becomes really challenging. A central aspect is how to represent it, while achieving a proper trade-off between storage size and quality. On one side, the view-dependent estimation methods presented in Section 2.1, which store and resample from each initial video frame, eventually with additional alignments to avoid ghosting effects, are memory costly. On the other side, the multi-view texture estimation methods discussed in Section 2.2, which compute single appearance texture atlases, reduce storage but they may compromise quality. To overcome the resolution and sampling limitations, approaches which use Super-Resolution techniques discussed in Section 2.3, integrating also the geometric modelling in Section 2.4 have been devised. In the previous works however, appearance variability is not explicitly studied with the danger to be eliminated. Visual redundancy with which the appearance of subjects is observed, across several sequences of the same subject performing different actions or motions, may not be fully exploited.

Towards the direction of both temporal and spatial exploitation, recent works (Casas et al. [2014]) introduce the 4D video texture representation, which resamples the captured video into a set of texture map layers according to surface visibility. Each layer stores the visual information for each mesh face in order of visibility and resampling into this domain is realized through optimization to preserve spatio-temporal coherence. To further compensate for misalignments, appeared in the combination of multiple video sequences, an additional alignment step takes place. Optic flow correction is applied on the rendered images from different video sequences to compensate for the ghosting artifacts and to maintain the visual quality. In the work of Volino et al. [2014] the idea of optic flow alignment is also introduced, in the form of precomputed surface-based optical flow. Similarly with Casas et al. [2014], the authors make use of the 4D video texture representation and they define the problem of dynamic surface appearance representation as a resampling from the layered texture maps through the same labeling formulation. The key difference is the pre-computed and stored optical flow to accelerate the rendering process. The geometry is rendered at given viewpoint, textured with the visual information from other views and optical flow is then computed to express the required alignment due to geometric and calibration inaccuracies. This idea of pre-computed and stored optical flow is also explored in the later work of Casas et al. [2015], where the authors interestingly extend the framework of Casas et al. [2014] by this additional feature for more accelerated renderings. Most of these works were primarily aimed at animation synthesis using mesh data, and do not propose a global appearance model for sequences.

In this thesis, we study the appearance variability of a dynamic subject, observed over one or several temporal sequences in the context of spatio-temporal domain. Drawing inspiration from state of the art and extending concepts initially explored for faces and static objects (Turk and Pentland [1991b], Blanz and Vetter [1999], Cootes et al. [2001], Nishino et al. [2001], Cobzas and Jägersand [2002]) we cast the temporal extension of the appearance modelling as a dimensionality reduction problem. In the work of Turk and Pentland [1991b], face variability of a population was represented by eigenfaces for recognition purposes. The concept was broadened to built a 3D generative model of human faces both in the geometry and texture domains, using the fact that the appearance and the geometry of faces are well suited to learning their variability as linear subspaces (Blanz and Vetter [1999]). Cootes et al. [2001] perform the linear PCA analysis of appearance and geometry landmarks jointly in their active appearance model. Nishino et al. [2001] instead use such linear subspaces to encode the appearance variability of static objects under light and viewpoint changes at the polygon level. A differential texture basis in the work of Cobzas and Jägersand [2002] is built to encode geometric and photometric variations. Each image is composed from modulating vector, applied on this basis. Images at new view points are then synthesized by interpolating the modulation vectors from the nearest sample views and being applied on the basis.

In this thesis we provide a method which can be seen as a generalization of the popular works of Nishino et al. [2001] and Cobzas and Jägersand [2002] to the case of fully dynamic subjects with several viewpoints and motions. Through the proposed compact appearance representation, we enhance our insight about the linearities and the non-linearities of the appearance in the temporal domain and we extend our model to encode all the underlying variabilities as Eigen Texture combinations. As opposed to the work of Nishino et al. [2001], which uses linear subspaces to encode the appearance variability of static objects under light and viewpoint changes at the polygon level, we use linear subspaces for full body appearance and over multiple sequences. Compared to the work of Cobzas and Jägersand [2002], which also builds a local basis to capture geometric and photometric variations for each divided region of the image, the proposed dynamic appearance representation unifies this concept in the common texture space, which in turns results in a view-independent and thus more compact representation.

2.6 Conclusion

The massive amount of easily accessible visual information and the continuously increasing demands of real world applications on high photorealism, low cost, high performance lead to a reality with new requirements in the context of appearance modelling, which cannot be met by the state-of-the-art. This chapter examines previous works and reveals the current gaps which we aspire to bridge throughout this study.

As a first strategy to the problem of retrieving and rendering the appearance of objects appears to be the category of view-dependent estimation methods. These techniques aim to provide renderings at novel viewpoints by averaging the color contributions or by adapting more elaborate weighting schemes with additional alignments to avoid ghosting effects. Despite their sufficient quality, they do not allow to exploit important features. First, view-independent appearance

properties are not encoded in the proposed models. Second, appearance redundancy is not optimally explored in the sense of exploiting the hidden non redundant visual information to super-resolve the quality. Third, they can be memory costly considering a temporal extension for deformable objects.

Multi-view appearance methods on the other side build image-based texture atlases by blending contributions from all views for each texel. Through these approaches some of the previous shortcomings are surpassed. In particular the main intrinsic appearance details of objects are now expressed in the common appearance space and thus explored and stored to later render them. By providing a single appearance, these methods succeed to reduce storage and thus in terms of flexibility to temporal extension they appear to be more convenient. However, the available visual information is not extensively explored to reveal new hidden appearance information. The obvious redundancy with which the appearance of subjects is observed, across temporal frames, different viewpoints of the same scene, and often several sequences of the same subject performing different actions or motions is not fully exploited.

Driven by the visual redundancy which should intuitively enhance the quality of the result, methods which extend Super-Resolution techniques in the case of multi-view setup are also examined. These works however do not appropriately consider the previously mentioned key aspects of compactness and storage. As such there is still space for further investigation.

To examine how geometry can be integrated in the problem of appearance modelling, previous works which succeed to jointly model both modalities are also discussed. Although these works acknowledge that both geometry and appearance can prove beneficial, there is not a well established study on how we can unify geometric and appearance modeling in a Super-Resolution framework in the context of multi-view scenario.

Finally, we review works focused on dynamic scenarios. Although several efforts have been done in the literature to express the dynamic nature of the appearance there is still the need for a view-independent appearance representation.

To summarize, the leap to a compact, view-independent, high quality appearance representation in the context of multi-view scenario that is able to represent large dynamic variability is still an open problem.

3 Linear Image Formation Model

3.1 Introduction and Motivations

The ability to record datasets of subjects in multi-view setups bolsters the need for appearance representations, which make optimal use of the amount of image information. Although several strategies have been developed to model the appearance under different assumptions, more compact representations are still required.

In this chapter, we address the problem of appearance modelling by introducing a common appearance space, where all the visual information encoded in the captured images is expressed. Under the assumption that both the geometry and the visual information of the captured object can be perfectly expressed into the appearance space, we consider the multi-camera observations as different realizations of one common appearance and we define a linear model to exhibit directly this relation. Images are thus defined as linear combinations of that unique appearance and the problem of solving for it is reduced to a problem of computing the contribution of each image into the common space and studying their central tendency by examining the most appropriate measure among the arithmetic mean, weighted arithmetic mean and median. The contribution of our proposed approach is threefold: first, as opposed to view-dependent texturing approaches (Cobza et al. [2002], Takai et al. [2010]), our model allows to capture view-independent appearance properties as well as to exploit viewpoint appearance redundancy. Second, compared to multi-view texturing approaches (Lempitsky and Ivanov [2007]), the compactness and linearity of our proposed observation model allow through naive estimation methods to explore in a first, general level, the noisy conditions of the multi-view scenario and to examine the limitation of the simplified assumptions. Third, its linear form opens the avenue for integrating image processing tasks, where similar image-wide linear operators exist.

The linear relation between the common appearance and the multiple observations is defined in Section 3.2. In Section 3.3 the contribution of each image to the defined appearance space is presented as texture realization and in Section 3.4 the problem of estimating the unique appearance given the previous texture realizations is addressed. In particular arithmetic mean, weighted arithmetic mean and median are used as estimation techniques to reveal the complexity of the appearance modelling problem in multi-view setup. To enhance our understanding about the general conditions, in Section 3.5 a first attempt of modelling the noise is presented and the estimation methods are reviewed on that new basis. The limitations of the proposed model due to

the simplified assumptions are thus revealed to direct our study towards more advanced modelling techniques. The main contributions of the proposed model are then summarized in Section 3.6.

3.2 Observation Model

In order to provide a compact representation of the appearance, we consider that all the images, captured at multiple viewpoints are different realizations of one unique appearance. We thus introduce a space (texture space), where we define that common appearance (texture) and we express the relation between this and the multiple observations. We assume that perfect mappings exist to express the geometry and the visual information of the captured object into the appearance space and we introduce a linear operator, which accommodates both the aforementioned mappings and the projection from the texture space into the image space. Through the linear observation model images are defined as linear combinations of the unique appearance.

Inserting the common texture space allows to exploit the view-independent appearance properties, that are encoded in the captured images. As opposed to view-dependent texturing approaches (Cobza et al. [2002], Takai et al. [2010]), our observation model exploits the redundancy among the multiple observations. The linear observation model is a compact representation, which contrary to multi-view texturing approaches (Lempitsky and Ivanov [2007]), allows to explore, in a first, general level, the noisy conditions of the multi-view scenario with respect to our assumptions. In addition to that, it opens the avenue for integrating image processing tasks, where similar image-wide linear operators exist (image enhancement techniques, Super-Resolution Tom et al. [2002]).

To define our linear observation model, the captured images as well as the geometry of the object are used. A geometric model is reconstructed, tracked or refined to be as close as possible to reality and is mapped into a two dimensional parametric domain to define the appearance space as a vector space. The visual information of the object is also expressed in that space and its unique appearance is represented by a vector. The problem of appearance modelling is thus reduced to a problem of defining the relation between the input geometry, the input multiple captured images and the unknown common appearance (see Figure 3.1). In 3.2.1, we firstly study the geometric aspect and introduce the parametric space to map the geometry. In 3.2.2, we leverage this space and we express in it the visual information to define the common appearance space. We then introduce the linear observation model to simulate the generation of images given the common appearance.

3.2.1 Appearance Space as 2D Parameter Domain

We firstly study how the underlined geometry is encoded in the linear observation model and how it defines the space where the visual information of the object is projected. Initially, the surface manifold seems to be a natural choice, on which we could define the appearance function to map to each vertex a color information. Following this direction however has an impact on the resolution of the common appearance, since the domain of the function depends directly on the discretization of the surface. In addition to that, manipulating such a function and expressing prior information is not computationally easy and intuitive. A two dimensional space instead, would be

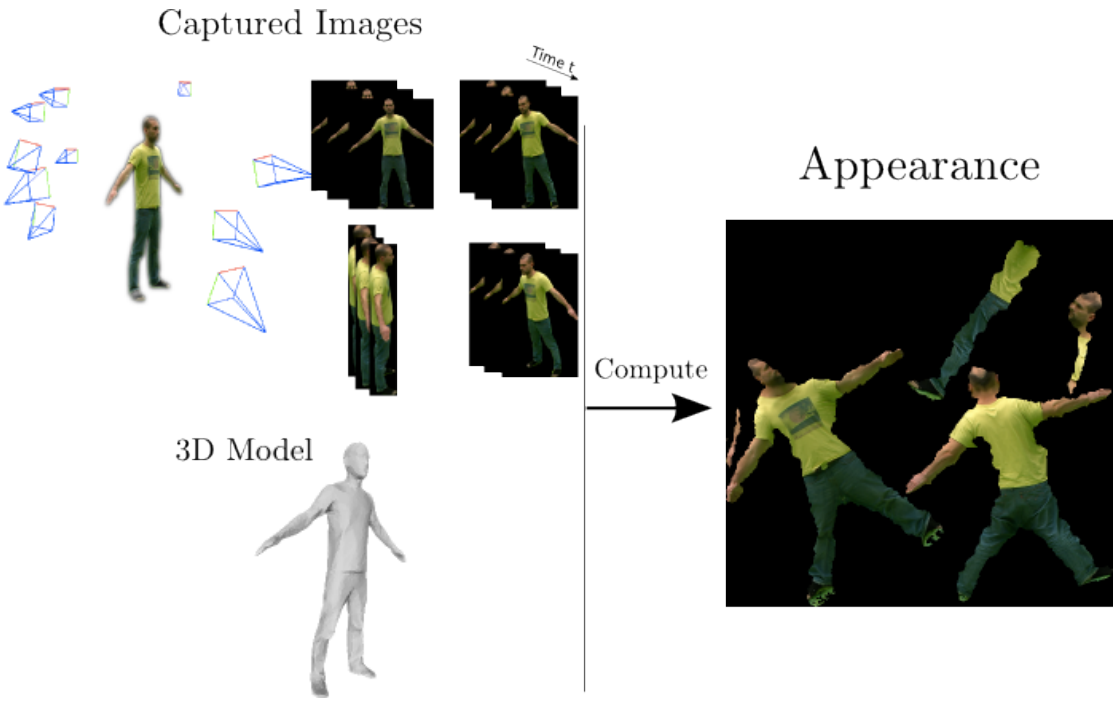

Figure 3.1 – The object of interest is captured at multiple viewpoints, time instances and the corresponding 3D model is computed. Given as input the observed images and the reconstructed geometry the goal is to compute the common appearance that is defined as a texture layer in RGB color space.

a more convenient choice for defining the appearance function of the object. The discretization of that space can be independent on the one of the surface and any constraint can be easily inserted. To introduce such a space, the geometry needs to be expressed through a parametric representation, which maps a 2D domain to the surface. The appearance function is then defined in the same 2D parameter domain. We call the 2D domain of that function appearance space (texture space) and each of its entities texel. The problem of solving for the appearance is thus reduced to a two dimensional problem.

Such a parametrization though is not without its own difficulties, since certain properties need to be handled properly. The surface parameterization needs to be as distortion free as possible in order to accurately model the appearance. By that we mean, that the ratio of a defined patch in the 2D domain to its mapped surface area needs to be similar everywhere. This property of isotropic scaling is preserved by imposing a conformal mapping. To obtain this mapping, we use off-the-shelf algorithm Sheffer et al. [2005], which yields large charts with relatively few components. This non continuous behavior is due to potential cuts and non-zero genus topology of the objects of interest. In these cases, conformal mappings are not continuous and have a support region with several connected components (or charts) in the 2D parameter domain.

Let $\phi_1 \ldots, \phi_k$ be the k conformal mappings of the charts $C_1, \ldots, C_k \subset \mathbb{R}^2$ to the surfaces patches $P_1, \ldots, P_k \subset S \subset \mathbb{R}^3$, $\phi_j := C_j \rightarrow P_j$. The charts $C_1, \ldots, C_k$, are pairwise disjoint and the union of

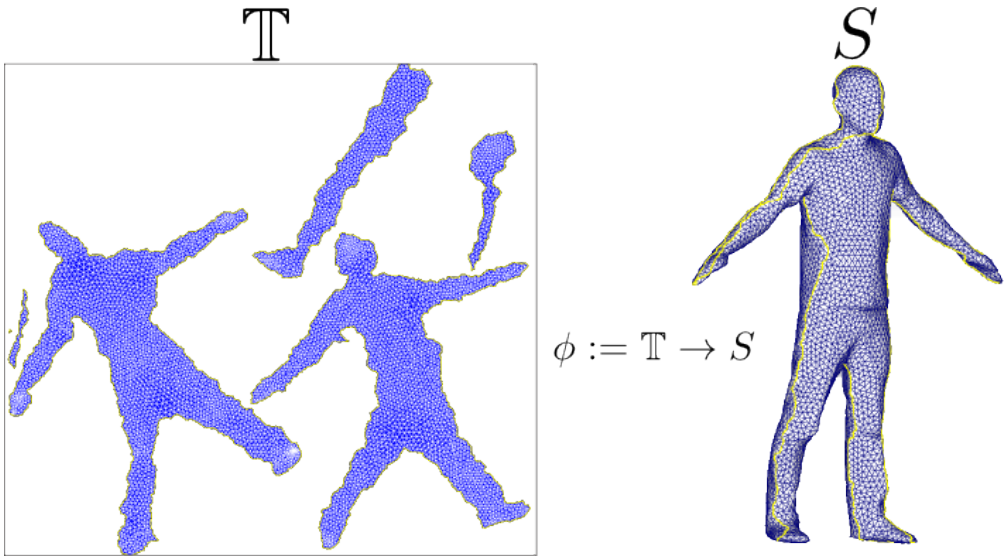

Figure 3.2 – Through the conformal mapping ϕ the appearance space (texture space) $\mathbb{T}$ is mapped to the surface space S. The yellow line on the surface notes the necessary cut in order for the conformal property of the mapping to be met.

the surface patches, $\cup_{j=1}^{k} P_j$, covers the surface S. Appearance space (or texture space) is the vector space, defined as the union of the charts:

$$\mathbb{T} := \bigcup_{j=1}^{k} C_j \subset \mathbb{R}^2. \tag{3.1}$$

To simplify notation, the global conformal mapping is defined as $\phi := \mathbb{T} \rightarrow S$ and is illustrated in Figure 3.2.

3.2.2 Image Formation Model

We now introduce the linear image formation model to relate the unknown appearance with the captured images. The geometric information has already been encoded in the 2D texture space and the next remaining step is to express in that space the visual information. We firstly define the appearance function and then we introduce the linear operator to simulate the image generation process.

The appearance of an object is described as the way it looks under certain lighting conditions. Each vertex of the surface has a characteristic value, which we express through an appearance function defined on the surface. We furthermore extend this relation in the 2D parameter domain (appearance space) and we introduce an equivalent appearance function to assign to each texel the corresponding visual value. We define this appearance function as a composition of the conformal mapping of the surface and the appearance function defined on the surface. Intuitively, this means that each texel is firstly mapped to the surface and then it is mapped to a color value. We define the texture map (appearance map) as the union of these color maps.

Let $T := S \rightarrow \mathbb{R}$ be the appearance function, which assigns to the surface an appearance value.

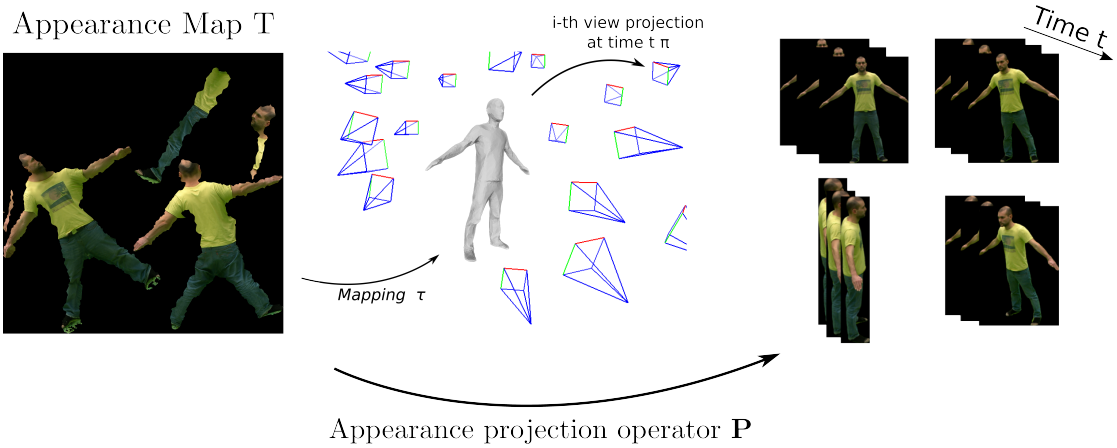

Figure 3.3 – We introduce the sparse matrix $\mathbf{P}_i$ to express the image vector I_i at view i as the linear transformation of the common appearance map T. The image is expressed as $\mathrm{I}_i = \mathbf{P}_i \mathrm{T}$.

For the sake of simplicity of notation, gray scale values will be considered. The extension to other color spaces is realized by introducing appearance functions for each channel independently. Let $\tau := \mathbb{T} \rightarrow \mathbb{R}$ be the equivalent appearance function defined in the texture space $\mathbb{T}$. This function assigns to each texel of texture space a color (appearance value) and it equals to the composition of the conformal mapping of the surface and the appearance function defined on the surface, $\tau := T \circ \phi$. In the case the texture space is not continuous and there are k conformal mappings, for each chart C_j, a corresponding appearance function is defined as $\tau_j := T \circ \phi_j$. Texture map (or appearance map) is the vector, defined as the union of the image of the appearance functions:

$$\mathrm{T} := \cup_{j=1}^{k} \tau_j(C_j). \tag{3.2}$$

Now that we have defined the texture map as a vectorized form we can proceed to the image formation model to simulate the generation of the images. In particular, representing them in vectorized form, we can express each pixel as a linear combination of the texels of the appearance map. This is materialized through the composition of operations, which maps the texel from texture space to the surface and then to the image space. Occasionally the density of projected texels is insufficient (*i.e.* in high curvature regions of the surface) for a pixel to be mapped to any texel. In this situation, the underlying surface appearance perceived by this pixel can be approximated by interpolation of neighboring texels. To treat also the case where the density of projected texels is sufficient in a uniform and continuous way, this interpolation scheme is applied on every pixel. The continuity of this scheme ensures that no artificial discontinuity is created as a result of a discrepancy in the treatment of these cases.

Let $\mathscr{M}$ be the given pose of the surface, $\{\mathrm{I}_i\}$ be the set of input images, where $i \in \{1, \cdots, n_i\}$ is the camera index. We note $\{\pi_i\}$ the known 3×4 projection matrix for each view i. The texel at location k of the texture space $\mathbb{T}$ is mapped to the geometric point $\phi(k)$ of the model $\mathscr{M}$, it is then projected on the image space at the pixel $\pi_i \circ \phi(k)$. The value of the pixel is equal to the appearance value, to which the texel is mapped through the appearance function $T \circ \phi(k)$. Let $\mathbb{I}$

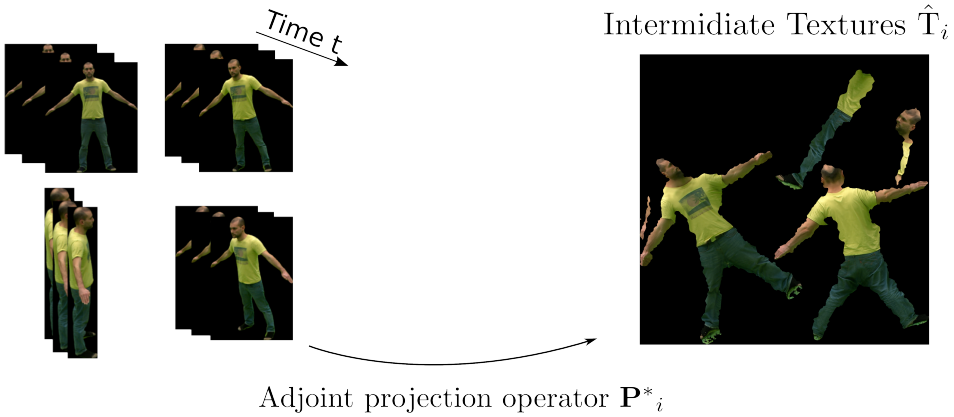

Figure 3.4 – We introduce the adjoint operator $\mathbf{P}^*{}_i$ to express the contribution of the image vector I_i at view i to the texture space T. The image I_i is back-projected into the texture space through $\hat{\mathrm{T}}_i = \mathbf{P}^*{}_i \mathrm{I}_i$.

be the vector space , where the images are defined. We note $\mathbf{P}$ the linear projection operator to map the vector space $\mathbb{T}$ to the vector space $\mathbb{I}$. In particular, we introduce the sparse matrix $\mathbf{P}_i$ to represent the image vector I_i as the linear transformation of the common appearance map (see Figure 3.3). The linear projection mapping is written as follows:

$$\mathrm{I}_i = \mathbf{P}_i \mathrm{T}. \tag{3.3}$$

Both the appearance and the image are expressed as vectors with dimension $w_{\mathrm{I}_i} \cdot h_{\mathrm{I}_i} \times 1$ and $w_{\mathrm{T}} \cdot h_{\mathrm{T}} \times 1$ respectively. The matrix $\mathbf{P}_i$ is of size $w_{\mathrm{I}_i} \times h_{\mathrm{I}_i}$ rows and $w_{\mathrm{T}} \times h_{\mathrm{T}}$ columns.

3.3 Image Contribution to Appearance Space

According to the previous model, each captured image is defined as an observation of the unique appearance of the object. Now that we have defined the direct model from texture space to image space, we go back to the original texture estimation problem. To compute thus how each image contributes to the common appearance space the opposite direction needs to be examined. We need to back-project each input image into the common texture space and to express there the captured color value. To that goal the adjoint of the linear projection operator need to be defined. Since the observation operator is a linear mapping from the texture space to the image space the adjoint operator equals to the transpose matrix (conjugate-transpose boils down to transpose for non complex entities in the matrix). Let $\mathbf{P}^*{}_i$ be the adjoint operator for the image I_i, the contribution of that image is then expressed as the realization of texture map defined $\hat{\mathrm{T}}_i = \mathbf{P}^*{}_i \mathrm{I}_i$ (see Figure 3.4).

Through this back-projection, each image contributes to the appearance space with a specific color value. Several appearance candidates correspond to each texel and as a result appearance redundancy can be easily exploited as opposed to view-dependent texturing approaches (Cobza et al. [2002], Takai et al. [2010]).

3.4 Texture Estimation

The estimation of the unique texture map, is thus reduced to the problem of exploiting in the most optimal way the appearance redundancy. The several texture realizations behave as samples of distributions and their central tendency can serve as solution for the unique texture map. In statistics, the most common measures of a population are the arithmetic mean, the weighted arithmetic mean and the median, which depending on the conditions can reflect the central tendency. To develop thus an intuition of the general problem, we examine the use of naive estimation algorithms for texture retrieval by leveraging these measures. We explore in a first, general level, the noisy conditions of the multi-view scenario and the inherent limitations of our proposed model.

To achieve the most precise statistical analysis, we aim to increase the size of the set of texel candidates as much as possible. A large set of color candidates for each texel reflects in a more precise way its statistical behavior. To that goal the appearance of the subject needs to be captured with the most possible coverage. To maximize the possible coverage, a careful physical arrangement of the cameras in a multi-view capturing system and an appropriate increase of the field of view of each of them need to take place. Even when these specific configurations are met, the complexity of the geometry or the pose of the subject can introduce non visible regions e.g the sole of the feet when the captured subject is a human with a standing pose. Due to these occlusions, there are texels in the texture space without any sample and modelling the appearance in this case becomes intrinsically difficult. To accommodate these cases, additional information about the expected appearance needs to be introduced and it will be detailed in Chapter 4. For the sake of simplicity, we account for these cases by carefully selecting the view points in the capturing process. TOMAS dataset is used, which consists of 68 input views each captured at resolution 2048×2048 pixels per frame. For each captured image, the corresponding texture realization is computed according to 3.3 and the texture map is estimated using arithmetic mean, median and weighted arithmetic mean. To examine the general conditions and to choose the most appropriate measure, images are rendered using the estimated texture at each case according to 3.2.2 and they are compared to the input images.

3.4.1 Arithmetic Mean of Image Contributions

Considering each appearance sample equally important in the estimation of the texture map, we apply the arithmetic mean. The texels of the texture map are equal to the sum of all corresponding samples divided by their number. Let $\{\mathrm{I}_i\}$ be the set of images and $\{\hat{\mathrm{T}}_i\}$ be the computed texture realizations. The texture map which corresponds to the arithmetic mean is:

$$\hat{\mathrm{T}}_{\mathrm{mean}} = \frac{1}{Z(N)} \sum_{i=1}^{n_i} \hat{\mathrm{T}}_i, \tag{3.4}$$

where $Z(N)$ the normalization factor reflecting the number of samples for each texel of the texture space. As was previously mentioned, it is low bounded by 1 imposing that there is at least one camera seeing that surface patch and upper bounded by the maximum number of cameras.

In Figure 3.5 the input image at view i is compared to the image rendered with the estimated

Figure 3.5 – From left to right: Computed image with the arithmetic mean texture map through the observation model $\hat{\mathrm{I}}_i = \mathbf{P}_i \hat{\mathrm{T}}_{\mathrm{mean}}$. Close ups on the T-Shirt and face of input and computed image.

texture map. The over-smoothing nature of the arithmetic mean is demonstrated in the close ups, where the T-Shirt of the computed image is blurry and the details of the letters are missing. A possible explanation for such result is that the initial assumption about equal contribution of each texel does not hold and an elaborate filtering should be proposed.

3.4.2 Weighted Arithmetic Mean of Image Contributions

Towards the direction of a more careful consideration of the texture samples, the weighted arithmetic mean appears to be a good candidate. However, texture samples should not contribute equally in the texture estimation, since they originate from images whose reliability also differs. And by that, we mean that the appearance value of surface patches perceived by cameras, which are fronto-parallel facing the patches, is more reliable compared to the corresponding appearance value perceived by cameras at grazing views. In order to express the different reliability of the texture samples, a weighting scheme is introduced to define the weighted arithmetic mean of the texture map:

$$\hat{\mathrm{T}}_{\substack{\mathrm{weighted} \\ \mathrm{mean}}} = \sum_{i=1}^{n_i} \frac{1}{Z(\mathrm{D}_i)} \mathrm{D}_i, \hat{\mathrm{T}}_i \tag{3.5}$$

where D_i is a diagonal matrix and $Z(\mathrm{D}_i)$ a normalization function of D_i. When acquiring appearance in the 3D case it is well known that contributions of each pixel q need to be modulated according to the angle θ_q between viewing vector and local surface normal Debevec et al. [1996]. We develop this idea and we introduce the function $\mathrm{d}'(\theta_q) = e^{-s\tan\theta_q}$ as a faster approximation of a normal distribution over the angles of the perceived surface. We then express the weights assigned to every pixel of the image space into the texture space by back projecting them and we set each diagonal element of $\mathrm{D}_i = \mathbf{P^*}_i \mathrm{D}'_i$. The $Z(\mathrm{D}_i)$ is the normalization factor, which equals to

Figure 3.6 – From left to right: Computed image with the weighted arithmetic mean texture map through the observation model $\hat{\mathrm{I}}_i = \mathbf{P}_i \hat{\mathrm{T}}_{\substack{\text{weighted} \\ \text{mean}}}$. Close ups on the T-Shirt and face of input and computed image. Due to the weighting scheme appearance candidates from cameras who are not fronto-parallel to the object have a smaller weight in the computation of the appearance and thus previous artifacts on the T-shirt are eliminated.

the sum of weights of the samples for each texel of the texture space. We use $s = 7\pi/16$ over all experiments. This weight is more conservative than the $\cos\theta_q$ weight usually used for blending in multi-view texturing techniques Debevec et al. [1996], and yields improved results in our experiments, as it downgrades unreliable contributions from surface points at a grazing angle.

We present in Figure 3.6 the corresponding close ups and we demonstrate that indeed previous artifacts in the case of simple arithmetic mean disappear. Through the weighting scheme image contributions have different importance and the visual discrepancy on the patches of the T-Shirt is eliminated. However, compared to the initial input image, the computed one is still blurry and details are not retrieved. Given these observations, the idea of averaging becomes questionable and examining a more robust measure is the next attempt.

3.4.3 Median of Image Contributions

The median serves as a more robust measure to outliers and it can thus anticipate for the over smoothing behavior of both previous cases. In the context of the texture estimation, these outliers appear as measurement errors due to inaccurate capturing system and also as numerical errors in the computation of the linear observation operator. To compute the median texture map, all the texture samples are sorted in ascending order and the middle texture sample is chosen. In the case, when an even number of samples is available the median corresponds to the average of the

Figure 3.7 – From left to right: Computed image with the median texture map through the observation model $\hat{\mathrm{I}}_i = \mathbf{P}_i \hat{\mathrm{T}}_{\text{median}}$. Close ups on the T-Shirt and face of input and computed image. The median as a more robust metric results in less ghosting effects.

two middle values:

$$\text{Let } \{\mathrm{T}_1, \ldots, \mathrm{T}_{n_i/2}, \mathrm{T}_{(n_i+1)/2}, \ldots, \mathrm{T}_{n_i}\} \text{ be the ascending order,}$$

$$\hat{\mathrm{T}}_{\text{median}} = \begin{cases} \mathrm{T}_{(n_i+1)/2} & \text{if } n_i \text{ is odd,} \\ \frac{\mathrm{T}_{n_i/2} + \mathrm{T}_{(n_i+1)/2}}{2} & \text{if } n_i \text{ is even.} \end{cases} \tag{3.6}$$

Sharper details and less blurry areas are retrieved through the median (see Figure 3.7) However, blocky effects appear, which visualize the lack of a smoothing scheme to account for appearance information in a predefined vicinity. On top of that, the fact that not all the samples take part in the calculation of the texture map could mean that valuable visual information is neglected.

Through the previous measures, the complex nature of the problem is revealed. In Figure 3.8, close ups of the rendered images using the estimated texture in each case demonstrate the overall low quality. This naturally drives us to question the simplified assumptions of the model and to investigate further the problem of appearance modelling.

3.5 Discussion

To enhance our understanding about the noisy environment and to reason for the low visual quality of the estimated texture maps, both steps in the previous estimation methods need to be analyzed. First, the common appearance is computed by back projecting all the visual information into the texture space through the adjoint operator of the observation model. Second, this estimated texture map is used through the same observation model to generate the images at the same viewpoints. In the ideal case of a noiseless environment, the computed images would have

approximately the same quality to the one of the input images. However, the experimental validation shows that instead, they are poor approximations of the input images and thus noise is probably accumulated in the whole observation process. The initial assumptions that the geometry and the visual information can be perfectly encoded in the linear observation operator are simplified assumptions and a more elaborate way to express this noise should be provided.

As a first attempt, we could express this noise through an additive term in the image observation process and reexamine the measures used for the texture estimation on a new basis. Let n_i be the noise while capturing the image I_i. The observation model then writes as follows:

$$\mathrm{I}_i = \mathbf{P}_i \mathrm{T} + n_i. \tag{3.7}$$

Rephrasing them into the context of the new formalization of the observation model, useful information about the conditions can be extracted. Expressing the problem in a probabilistic framework the mean of the texture realizations is equivalent to the maximum likelihood solution of the texture map under the assumption that the noise follows Gaussian distribution with zero mean value and variance equal to 1. In the same vein, the weighted mean corresponds to the maximum likelihood solution under the assumption that the noise follows Gaussian distribution with zero mean value and variance analog to the weight. The median corresponds to the maximum likelihood solution under the assumption that the noise follows Laplace distribution. The experimental validation of the texture estimation methods shows that the quality of the rendered images using the estimated textures is still low compared to the input images. A specific conclusion about the noise cannot thus be easily drawn and a more elaborate observation model should be provided.

Representing the noise as an additive term does not describe the general condition of the observation process and there is still the need to study in details the parameters of the problem as potential noise sources. The linear observation operator encodes both geometric information and information about the capturing process. The geometric aspect of the problem, which is implicitly inserted in the operator through the conformal mapping and the projection matrices cannot be neglected. Imprecisions of the geometric reconstructions and inaccurate projection matrices are adding noise, which is not simulated by a simple additive term. The image capturing process on the other hand has its own specificities. It is common in the literature Tom et al. [2002] to simulate the capturing process by a concatenation of linear operations, each of which contributes to a certain degree to the accumulation of noise. Further study should thus be carried out to explore the possibilities of advance modelling techniques, a direction towards which, the linearity of the current observation model plays an additional support.

3.6 Conclusions

In this chapter a common appearance space is introduced to encode the visual information of the object. Under the simplified assumptions that both geometry and appearance can be expressed in that space, we consider the multi-camera observations as different realizations of one common appearance and we define a linear observation model to exhibit directly this relation. Through the common texture space, appearance redundancy as well as view-independent appearance properties

are exploited as opposed to view-dependent texturing approaches. Through the compact and linear observation model image contributions to the appearance space are easily computed, which in turns allow for a global texture estimation. Through the simple texture estimation techniques, we succeed to explore the complexity of the appearance modelling problem and the need for a more elaborate study of the noise sources. The linearity of the proposed observation model drives the study towards image enhancement tasks, where similar image-wide linear operators exist, e.g super-resolution techniques and noise sources are handled in a more sophisticated manner. Representing in a more realistic way the image generation from the common appearance and exploiting the linearity through image enhancement techniques are the main concepts that are assembled in the next chapter to provide a Super-Resolution scheme.

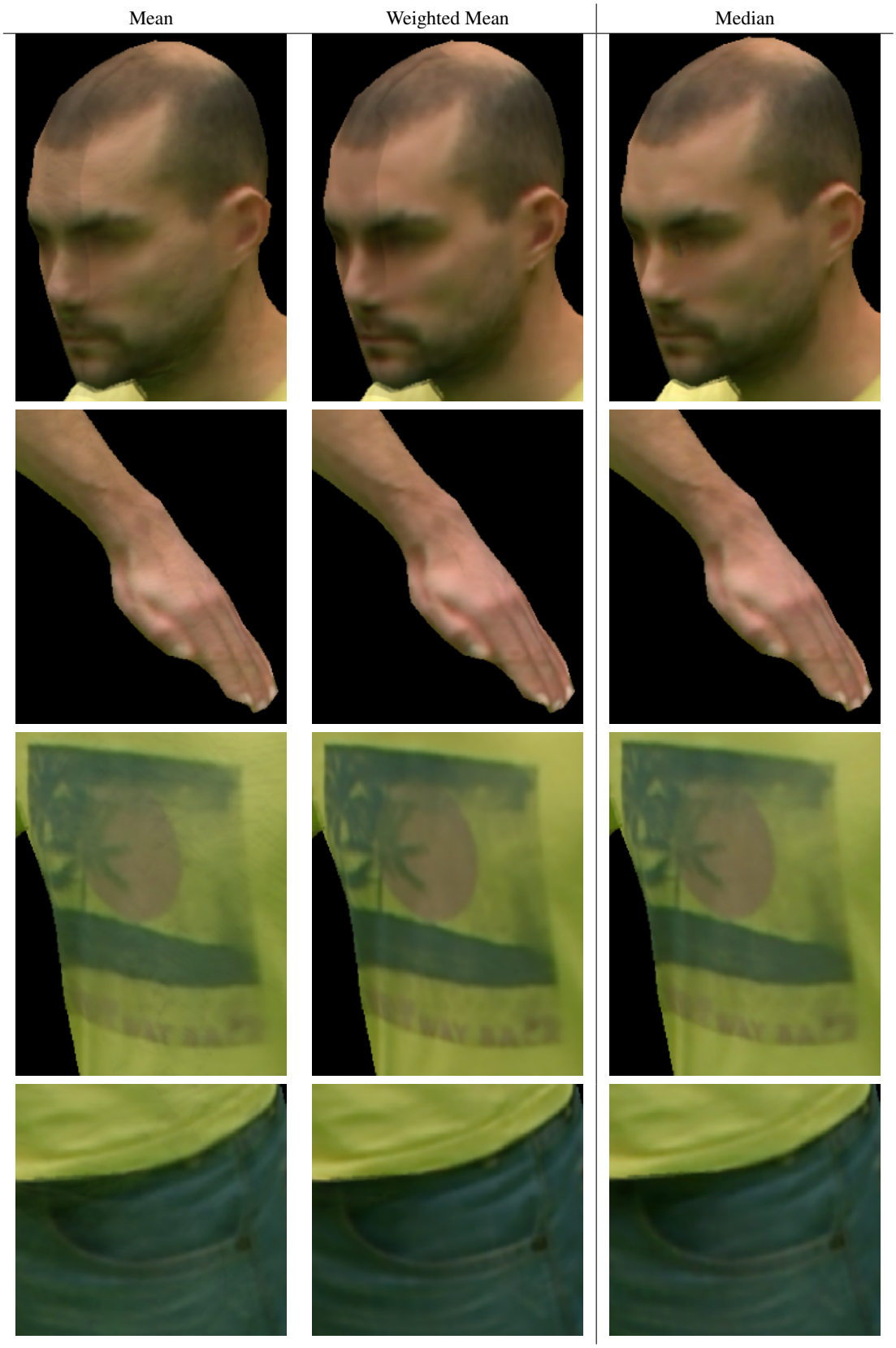

Figure 3.8 – Median serves as a more robust measure. It anticipates for the over smoothing behavior compared to the cases of arithmetic mean and weighted arithmetic mean. Blocky effects are however still apparent to visualize the lack of smoothing scheme to account for appearance information in a predefined vicinity.

4 Super-Resolved Appearance Model

4.1 Introduction and Motivations

In Chapter 3 a linear generation model has been provided to express the input observations in a common texture space. Through this representation naive blending techniques have been revisited and the inherent limitations of the model have been revealed.

In this Chapter we propose a novel model which overcomes previous shortcomings and serves a Super-Resolved appearance. It is a compact representation which builds on the previously presented linear appearance model and successfully enhances it with the Super-Resolution (SR) fundamentals integrated in the multi-view context. Our proposed method successfully leverages the following common features of Super-Resolution techniques (SR). First, SR techniques aim to recover additional details lost in the acquisition process. We also target to high visual quality of the appearance. Second, the underlying principle of the SR is the availability of multiple observations, which differ in sub-pixel precision. These sub-pixel shifts exist also in the multi-view case. Third, the fundamental technical aspect of SR to express the image generation process as a concatenation of operations allows to integrate it in our previous linear appearance model. Although SR techniques gain a lot of attention in the recent literature and have been extensively studied in the monocular case, few works exist in the multi-view case. As opposed to previous work Tung [2008], which enhances input views directly, our proposed model builds one common super-resolved appearance available to be re-used in the future. Compared to Goldluecke and Cremers [2009a], our method accounts for every source of noise uniformly through a linear representation of the image generation process and thus exhibits higher visual results.

In the first Section 4.2, we present our reasoning behind our proposed Super-Resolution (SR) model in the multi-view case. We examine the definition of SR, the principle of SR to be applied and its fundamentals from technical aspect. Summarizing the main features of SR, we then introduce in Section 4.3 the concept of Super-Resolution in the multi-view case and we present the model, which materializes the SR. Solving then for the super-resolved appearance is recast into a form of statistical Bayesian inference problem in Section 4.4. We exhibit a two-stage iterative algorithm which utilizes as initialization the weighted arithmetic mean for the appearance and converges to the unique super-resolved appearance. In Section 4.5 the performance of the proposed method is demonstrated through quantitative and qualitative evaluation using several datasets and under different scenarios. In last Sections 4.6 and 4.7 we summarize the contributions

of our Super-Resolved appearance model and we introduce the new research directions, which the findings of our proposed model opens.

4.2 2D Monocular Super-Resolution

The limitations of the previous model disclosed the need to study in more details the general conditions of the framework. Our intuition that the availability of multiple observations can give better visual results directed our study towards image enhancement methods. It is a well established research area with the image interpolation and the image restoration as naive representatives. Among the existing methods, 2D Super-Resolution (SR) gains a lot of attention in the recent literature and it inspires our current study.

In this section monocular Super-Resolution is introduced as a method to acquire an image with high visual quality. The terminology of visual quality is firstly explained and the goal of SR is defined in 4.2.1 through a comparison with the image interpolation and the image restoration method. In 4.2.2 the underlying principle of SR technique is presented and in 4.2.3 its technical aspect is reviewed.

The goal of the current thesis is to obtain an appearance with high visual quality and by that we mean an appearance which reflects the reality. Resolution is another term to quantify visual quality and it has been used in the literature to describe several concepts.

- Spatial resolution

  The spatial resolution refers to the density and the spacing among the pixels, elements of which the digital image consists. It is measured in pixels per unit area. The higher it is, the greater the number of pixels, the smaller their size is and thus the more details and color transitions is possible to be captured.

- Brightness resolution

  It defines the quantization of the light energy that is absorbed by the device sensor and thus the quantization level of the system. It is related to the number of brightness levels that can be recorded.

- Temporal resolution

  It refers to the number of frames that can be captured per second by the system which justifies the frame rate as another term which is used instead. The higher the temporal resolution the sharper motions can be captured.

In the current thesis, resolution refers to spatial one. In the context of visual quality enhancement, methods, which increase the spatial resolution are of particular interest to us.

4.2.1 Super-Resolution vs Image Interpolation and Image Restoration

Interpolation and restoration methods belong to the class of image enhancement methods, which we are interested in. Image resizing is realized through interpolation methods and there are several algorithms (Schoenberg [1972], Crochiere and Rabiner [1981], Unser et al. [1995]) which

yield better quality images overcoming the blocky effects from simpler approaches. Increasing however the number of pixels of the image does not translate into an increase of visual quality. Image interpolation methods are inherently limited by the information that is present on the image and they are not able to compute the visual details that were lost in the capturing process. Aiming for higher quality visual results, image restoration methods could be another choice. It is a well-established area (Andrews and Hunt [1977], Katsaggelos [1991]) whose goal is to compute high quality images. These methods recover degraded images without changing their size.

A superset of the aforementioned methods is considered to be the family of Super-Resolution (SR) algorithms, which has been developed recently (Chaudhuri [2001] and Milanfar [2010]). These methods increase the resolution of the image like interpolation methods and they de-blur it like restoration methods. Their additional advantage is that SR recovers details, which were not visible before. The super-resolved image includes new, previously hidden information. Retrieving such level of visual quality is the scope of our study and thus Super-Resolution method becomes appealing to further investigate.

4.2.2 Super-Resolution Principle

A first assessment of our proposed idea, to apply SR in the multi-view case, is realized through the study of the principle of monocular SR. In SR literature, an image is super-resolved when multiple observations of the same scene which differ in sub-pixel precision are available. Recovering visual quality could become indeed more challenging when observations with such sub-pixel differences are available; aliasing and blurring are stronger. However, due to these misalignments there is non-redundant information, which can be retrieved. If observations were shifted by integer units, then each image would contain the same information and could be replicated from the others, a scenario that would be of no benefit for Super-Resolution purposes. SR techniques treat these misalignments as a re-sampling problem, which converts an irregular sampled image to a uniformly sampled one. The works of Katsaggelos et al., Huang and Tsay [1984] and Tom et al. [2002] explain how sub-pixel shifts allow Super-Resolution.

Such a scenario of observations with sub-pixel shifts exists in the multi-view case. Captured images form multiple viewpoints are available and thus common appearance information is shared among them. In Chapter 3 all the observations are expressed in one texture space, and thus the common appearance results in even stronger sub-texel shifts. Extending the multi-view case scenario and including the time component by considering video sequences, the appearance information is fused also in time and yields even stronger aliasing effects, a concept that will be detailed in Chapter 5. Thus, the basic premise of Super-Resolution techniques governs also the multi-view scenario, an observation that naturally provokes a more detailed study of the family of Super-Resolution algorithms.

4.2.3 Techniques for Super-Resolution

To examine the feasibility of introducing such a Super-Resolved appearance model, common techniques in the 2D Super-Resolution literature are reviewed. A complete overview of the categorizations is provided in the works of Borman and Stevenson [1998], Jing and Kai-Kuang [2011] as well as a detailed presentation in the works of Chaudhuri [2001] and Milanfar [2010].

The fundamental idea of every SR approach is to simulate the image generation process. In 4.2.3.1 the steps of the image acquisition are discussed and in 4.2.3.2 the possible ways of modelling the image generation process are presented. Reconstruction based approaches and example based approaches, referred also as learning based in the literature Jing and Kai-Kuang [2011] are introduced in 4.2.3.3 as two techniques to solve for the unknown super-resolved image.

4.2.3.1 Image Acquisition System

To simulate the image generation in the most realistic way, SR techniques decompose the whole image acquisition process. While recording a scene, the signal traverses the optical system, where it is blurred and captured by the CCD which performs light integration at every photosite. At the output of the system the images are noisy, blurry and of low resolution. This chain of steps is simulated by the 2D Super-Resolution literature (Baker and Kanade [2002], Fransens et al. [2007]) through a concatenation of operations.

Modeling the image generation by a stack of operations is of particular interest to us. It allows to integrate in the acquisition process the additional operation which relates the unknown appearance with the image (explained in Chapter 3). To explore how monocular SR techniques simulate this acquisition process comes as a next step to our study.

4.2.3.2 Spatial vs Frequency Domain Approaches

Both spatial or frequency domain can be used to simulate the image generation. The first frequency related method is credited to Tsai and Huang [1984] who transforms the low resolution images in the Discrete Fourier Domain (DFT) and derives a system of equations which combines the aliased DFT coefficients of the observed low resolution images and the one of the unknown high resolution image. The advantage of expressing the problem in frequency domain is to explicitly exploit the aliasing that exists in the low resolution observations. On top of that, it facilitates the theoretical study about the reason that Super-Resolution is possible. However these features are not of particular benefit to our framework.

In the current thesis, we define the problem and develop the reconstruction algorithm in the spatial domain, where many approaches have been developed and draw a lot of attention (Baker and Kanade [2002], Yamaguchi et al. [2010], Tian and Ma [2010], Liu and Sun [2011]). The motivation behind our choice is twofold (Borman and Stevenson [1998], Park et al. [2003], Jing and Kai-Kuang [2011]). First, spatial domain allows for a larger choice of motion representations in the image acquisition process. The assumption of uniformly spaced samples in the Discrete Fourier Domain restricts the motion representation to be translational. This requirement is not anymore valid in spatial domain, where there is the possibility to account for more complex motions, a concept that applies to the complexity of the multi-view scenario and will be discussed later in Section 4.3. Second, the possibility to insert a-priori spatial information on the image acquisition process provides a more intuitive way to express prior knowledge about the parameters of the problem.

4.2.3.3 Reconstruction vs Example based Approaches

Solving for the super-resolved image is treated as an inverse problem, where knowledge about the relation between the parameters of the model is required. This knowledge can be imposed on the model through examples defining the example based approaches or by reconstruction defining the reconstruction based approaches.

Example based approaches according to Yang and Huang [2010], named also as learning based approaches in the works of Lin et al. [2007], Jing and Kai-Kuang [2011] are techniques that utilize examples, from which prior knowledge of the parameters of the problem is learned. In these methods, a database with pairs of low resolution images and their corresponding super-resolved images is used to train every parameter of the model, which captures their relationship. In the seminal work of Freeman et al. a Markov network is proposed to model the SR problem as an inference problem. The likelihood of observing the low resolution image given the high resolution one and the prior distribution of the high resolution image are defined via image patches, whose probabilities are learnt from training images and are approximated by mixture of Gaussians. In the work of Baker and Kanade [2002] examples are formed by a pyramid derivative set of features instead of raw data directly. Regularization is then expressed by demanding a proximity between the spatial derivative of the unknown image to those of the examples. Overall, example based approaches can be efficient when sufficient observations are available and especially when narrow families of images are studied (Capel and Zisserman [2001]).

The lack of training dataset in the multi-view case scenario directs our study towards the other category of reconstruction based approaches. These approaches develop the relation between the low resolution images and the super-resolution image. The image formation model is well identified as a geometric warping, blurring and sub-sampling process of the initial high-resolution image. The priors on the unknown high resolution image as well as on other necessary parameters of the model are directly imposed. Addressing the reconstruction problem as a stochastic process by representing the unknown parameters as stochastic variables and assigning probability distributions to them has been developed in the literature. In particular, Bayesian models have been proposed (Lin and Shum [2004], Fransens et al. [2007], Liu and Sun [2011]) to explicit the priors and the existing dependencies. The problem of solving for the unknown Super-Resolution image is cast as Bayesian inference and Maximum Likelihood (ML) estimator is provided in the work of Pickup et al. whereas Maximum a Posteriori (MAP) estimator in the work of Liu and Sun [2011].

We follow this reconstruction approach to exploit mainly the flexibility to model the noise characteristics and also the a-priori information about the solution.

4.3 Super-Resolution in Multi-View Scenario

The main arguments to propose our Super-Resolved appearance model are revealed through the review of the monocular Super-Resolution method: first, the common goal for high visual quality of SR. Second, the same principle of sub-pixel shifts that governs the multi-view case. Third, the fundamental technical aspect to simulate the image capturing process, which allows to insert the appearance as an additional parameter in the whole chain.

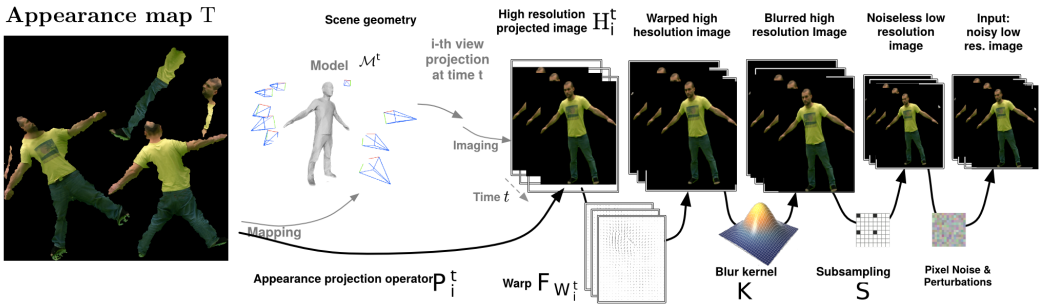

Figure 4.1 – Summary of image formation model and problem notation.

In this Section we introduce and materialize the concept of Super-Resolution in the multi-view context. In particular, we present a representation of the image generative model in the multi-view scenario, which allows to super-resolve the unknown appearance. It is a novel compact way of explaining how the image observations are generated by the unknown appearance (see Figure 4.1). Such a framework is significantly different from 2D Super-Resolution. In a multi-view scenario, the intrinsic appearance of a single 3D object is only partially visible in each view, and observed only after being perspective projected, distorted by 3D geometry, and self-occluded. The 3D geometry itself is subject to reconstruction error and thus uncertain.

Applying Super-Resolution techniques in a multi-view context has only recently started to be addressed. As opposed to previous work Tung [2008] which super-resolves input images directly, our proposed model super-resolves the different input contributions into one single coherent texture space. Using such a compact representation resembles the work of Goldluecke and Cremers [2009a], where the authors also introduce an image formation model to simulate the image generation from the common texture. Our proposed model is also based on the idea of expressing the image observation model given a common appearance. The major difference compared to the work of Goldluecke and Cremers [2009a] is the uniform way of treating the sources of variability.

To formulate the image generation process in the case of multi-view scenario, we follow the same steps of the technical analysis presented in the monocular Super-Resolution 4.2.3. The 3D nature of the problem introduces more degrees of freedom and thus additional sources of noise need to be accounted for. In 4.3.1 the geometric noise, which originates from the 3D nature of the problem, is examined and in 4.3.2 the model, which describes the image generation from the unknown appearance, is presented.

4.3.1 Geometric Noise as Additional Source of Variability

To simulate the image acquisition process in the multi-view context we leverage the corresponding one presented in the monocular Super-Resolution and the analysis of projecting the appearance into the image space presented in Chapter 3. The first two steps in Figure 4.1 describe the mapping of the high resolution appearance to the geometry and the projection of the textured geometry into a high resolution image space, whereas the following steps demonstrate the degradation

levels, which the image undergoes through the capturing process. The reconstructed model and the camera projection matrices, as the new parameters in the whole pipeline, act as additional sources of noise. We define them as two different forms of geometric noise:

- Reconstruction Noise

  The 3D geometry itself is subject to reconstruction error and thus uncertain. These reconstruction inaccuracies result in assigning wrong appearance information, retrieved from the input images, to the 3D model. The wrong assignment occurs through two steps. The reconstruction noise is firstly propagated through the mapping of the geometry into the 2D parametric space and then through the appearance function, which relates a point of that space with appearance information.

- Calibration Noise

  Imperfections in the calibration of cameras in a multi-view setup can also occur. This kind of noise has also the form of geometric noise since it introduces inaccurate assignments between the textured 3D geometry and the high resolution image space.

Occurrence of both forms of geometric noise is often unavoidable resulting in stronger misalignments. Imprecise geometry is mapped to a parametric space, to which inexact appearance information is assigned through the inaccurate camera matrices. All these misalignments appear in the texture space and due to these non-integer shifts, there is non-redundant appearance information, which is exploited to result in the super-resolved texture.

4.3.2 Image Formation Model

A parametric representation of the image formation process in the multi-view context is presented to explain how the low resolution observations are generated by the unknown high-resolution appearance and to accommodate all the levels of degradation (see Figure 4.1).

The first part of the modelling corresponds to the process, where the unknown high resolution appearance is mapped to the 3D geometry and then it is projected into the image space. Similar analysis of Chapter 3 is followed and the same notations are used. The main differences of this modelling are the introduction of the targeted high resolution and the reflection of the noise related to the 3D nature. As in every SR technique, the resolution of the super-resolved appearance is larger than the one of the input captured images up to a predefined factor. The same high resolution is intentionally preserved in the image generation to allow to keep track on every misalignment in high accuracy. Due to geometric noise, misalignments are introduced in the pipeline and they are materialized in the projection operator. Visually, they are expressed as sub-pixel shifts on the images, where they can be simulated by warp fields. In order to represent them in details, high accurate warp fields are required, which translates to fine resolution image spaces where these warps are defined. Overall, at that point of modelling, a high resolution image is generated according to the linear observation model introduced in Chapter 3 but updated to account for the noise sources.

The second part of the modelling resembles the image acquisition process used in 2D Super-Resolution literature (Baker and Kanade [2002], Fransens et al. [2007]), discussed in 4.2.3.1. It

expresses the capturing process by describing how the high resolution image is downgraded while traverses the optical system. A Point Spread Function (PSF) with the form of a Gaussian blur kernel is introduced at a first stage to describe the blurring that takes place. Another step of image sub-sampling is added to simulate the down-scaling the image undergoes while it is captured by CCD sensors. At the end of the whole pipeline, the down-sampled blurring images are further downgraded by the thermal sensor noise and color filtering, which are expressed through additive term.

Let T be the unknown super-resolved appearance with dimension $w_{\mathrm{T}} \times h_{\mathrm{T}}$, $\{\mathrm{I}_i\}$ the set of input captured images with dimension $w_{\mathrm{I}_i} \times h_{\mathrm{I}_i}$ and $\{\mathrm{H}_i\}$ the set of high resolution images corresponding to the set of input images. Following the same notation defined in 3.2.2, let $\hat{\mathbf{P}}_i$ be the noisy projection operator for the high resolution image at view point i. In particular, due to reconstruction inaccuracies, a texel at location k of the texture space $\mathbb{T}$ is mapped to the noisy geometric point $\hat{\phi}(k)$ of the model $\mathscr{M}$, it is then inaccurately projected, due to imperfect camera calibrations, on the image H_i at the pixel $\hat{\pi}_i \circ \hat{\phi}(k)$. The value of the pixel is equal to the appearance value, to which the texel is mapped through the appearance function $T \circ \hat{\phi}(k)$. Thus, in the presence of both forms of geometric noise, wrong assignments between the appearance T and the high resolution image H_i take place and the high resolution image is expressed as:

$$\mathrm{H}_i = \hat{\mathbf{P}}_i \mathrm{T}. \tag{4.1}$$

The sub-pixel shifts in the high resolution image are further simulated by warp fields over the noiseless high resolution image, which are generated by the noiseless projection operator. Let $\mathbf{P}_i$ this operator and $\mathbf{F}_{\mathrm{W}_i}$ the linear warping matrix operator, which maps target pixels from a linear combination of source pixels. The high resolution image is written as follows:

$$\mathrm{H}_i = \hat{\mathbf{P}}_i \mathrm{T} = \mathbf{F}_{\mathrm{W}_i} \mathbf{P}_i \mathrm{T}. \tag{4.2}$$

Let $\mathbf{K}$ be the blur operator to describe the light integration at photosite sensor and let $\mathbf{S}$ be the sub-sampling operator, which are both applied to the high resolution image H_i. The thermal sensor noise, color filtering noise are expressed in the capturing process through an additive term noted as N_i. The low resolution image I_i at the output of the capturing system in view point i is simulated by:

$$\mathrm{I}_i = \mathbf{SK}\mathrm{H}_i + \mathrm{N}_i. \tag{4.3}$$

The full image formation model in the multi-view context, which describes how the low resolution image is generated by the high resolution appearance (texture) is written as follows:

$$\begin{aligned} \mathrm{I}_i &= \mathbf{SKF}_{\mathrm{W}_i} \mathbf{P}_i \mathrm{T} + \mathrm{N}_i \\ \mathrm{I}_i &= \mathbf{A}_i \mathrm{T} + \mathrm{N}_i, \end{aligned} \tag{4.4}$$

where, under the assumption that every parameter is written in lexicographic order, the sparse linear operator $\mathbf{A}_i = \mathbf{SKF}_{\mathrm{W}_i} \mathbf{P}_i$ for each view $\{i\}$ has $w_{\mathrm{I}_i} \times h_{\mathrm{I}_i}$ rows and $w_{\mathrm{T}} \times h_{\mathrm{T}}$ columns and

the noise term is of size $w_{\mathrm{I}_i} \cdot h_{\mathrm{I}_i} \times 1$. Through this formulation, pixels are a linear combination of texels of T, and the problem of solving for the unknown appearance T is mathematically expressed as inverse problem. Given the set of low resolution images, as the results of measuring the appearance of a 3D subject, the actual value of that appearance needs to be inferred.

4.4 Probabilistic Model

To integrate in the imaging pipeline all the uncertainties as well as the prior knowledge of the high resolution texture in a natural way, we address the image reconstruction process as a stochastic process. We represent the unknown parameters of appearance and warp fields as stochastic variables and assign probability distributions to them. The forward process, described in equation 4.4 is a generative probabilistic model and the inverse problem is cast into a stochastic framework through Bayesian inference.

The corresponding graphical model is presented in 4.4.1, where the generative model is expressed in terms of random fields. In 4.4.2 the prior knowledge of the parameters of the model is expressed through analytic probability distributions and in 4.4.3 we cast the problem into a Bayesian framework where Maximum a Posteriori solution (MAP) of the appearance is computed through a two-stage iterative algorithm.

4.4.1 Graphical Model

In the forward process of image generation, expressed in equation 4.4, the high resolution texture T, the warp field W_i, the noise N_i and the observed low resolution images I_i are considered to be samples of random fields. The projection operator $\mathbf{P}_i$, the sub-sampling operator **S** and the Gaussian blur kernel operator **K** (PSF) are assumed to be known. The image formation model can then be described by the graphical model in Figure 4.2. Each random variable is represented by a node, the dependency among them is expressed by edges and is quantitatively defined by conditional probability distributions. The structure of the graph and the conditional probability distributions specify the joint probability distribution over all the random variables as follows:

$$p(\{\mathrm{I}_i\}, \{\mathrm{W}_i\}, \mathrm{T}) = p(\{\mathrm{I}_i\}|\{\mathrm{W}_i\}, \mathrm{T})\, p(\{\mathrm{W}_i\}, \mathrm{T}). \tag{4.5}$$

The conditional probability $p(\{\mathrm{I}_i\}|\{\mathrm{W}_i\}, \mathrm{T})$ expresses how probable is for the set of low resolution images $\{\mathrm{I}_i\}$ to be generated by the unknown texture T and the set of warping operators $\{\mathrm{W}_i\}$. The total joint probability distribution $p(\{\mathrm{W}_i\}, \mathrm{T})$ can further be factorized:

$$p(\{\mathrm{W}_i\}, \mathrm{T}) = p(\{\mathrm{W}_i\}|\mathrm{T})\, p(\mathrm{T}). \tag{4.6}$$

And since the set of warping fields $\{\mathrm{W}_i\}$ is independent of the high resolution texture T $p(\{\mathrm{W}_i\}|\mathrm{T}) = p(\{\mathrm{W}_i\})$ and the joint probability boils down to:

$$p(\{\mathrm{I}_i\}, \{\mathrm{W}_i\}, \mathrm{T}) = p(\{\mathrm{I}_i\}|\{\mathrm{W}_i\}, \mathrm{T})\, p(\{\mathrm{W}_i\})\, p(\mathrm{T}), \tag{4.7}$$

where $p(\{\mathrm{W}_i\})$ is the probability of the set of warping operators and the $p(\mathrm{T})$ the probability distribution of the high resolution texture. The generation of image I_i is independent of the

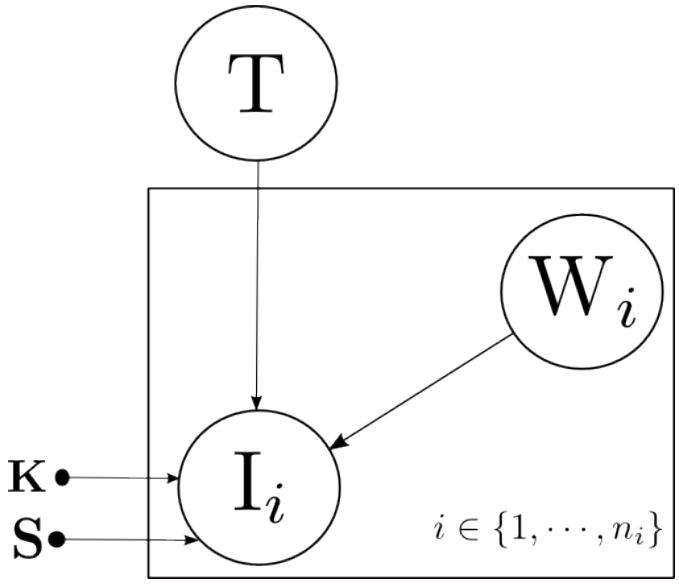

Figure 4.2 – The plate represents n_i nodes for the variable of image and n_i for the variable of warp field. Only one example for each of them is shown explicitly. The smaller bullets express the deterministic parameters of the sub-sampling operator **S** and the Gaussian blur kernel operator **K** (PSF).

generation at any other view point i and it only depends on the corresponding warping operator $\mathbf{F}_{\mathrm{W}_i}$. Thus the probability of the set can be factorized into the following:

$$p(\{\mathrm{I}_i\}, \{\mathrm{W}_i\}, \mathrm{T}) = \prod_i p(\mathrm{I}_i|\mathrm{W}_i, \mathrm{T}) \prod_i p(\mathrm{W}_i)\, p(\mathrm{T}). \tag{4.8}$$

4.4.2 Prior Information

The joint probability distribution of the graphical model is further developed through the analytic representation of the involved distributions. The dependency among the parameters of the graph is quantitatively defined at reconstruction stage in order to reflect the nature of the image generation problem in realistic way and at the same time to facilitate the tractability of the inference.

In 4.4.2.1 the observation model is presented to describe the distribution, which assigns a probability to the observed low resolution image as the output of the image generation process given the high resolution texture and warping operator. The texture model in 4.4.2.2 and the motion model in 4.4.2.3 express the statistical behavior of the texture and warping field.

4.4.2.1 Observation Model

Due to the image formation model of equation 4.4 the joint probability distribution $p(\mathrm{I}_i|\mathrm{W}_i, \mathrm{T})$ is related to the noise N. Zero mean independent white gaussian noise as well as zero mean independent laplacian noise are used in the 2D Super-Resolution literature and the exact model between these candidates is usually selected through the generalized likelihood ratio test (GLRT)

Farsiu et al. [2003].

In the current thesis, zero mean white independent Gaussian noise is assumed to be present in the image formation model, which results in the conditional distribution of the low resolution images given by:

$$p(\mathrm{I}_i|\mathrm{W}_i,\mathrm{T}) = \frac{1}{Z(\mathrm{D}_i)} e^{-(\mathrm{I}_i-\mathbf{A}_i\mathrm{T})^\top \mathrm{D}_i(\mathrm{I}_i-\mathbf{A}_i\mathrm{T})}, \tag{4.9}$$

where D_i is a diagonal covariance matrix introduced to allow different noise characteristics per pixel q, and $Z(\mathrm{D}_i)$ a normalization function of D_i. In 2D super-resolution models, a single variance per input image is usually used, with the i.i.d. noise assumption Fransens et al. [2007]. However when acquiring appearance in the 3D case it is well known that contributions need to be modulated according to the angle θ_q between viewing vector and local surface normal Debevec et al. [1996]. This can be purposely identified in the generative model, where each diagonal element $\mathrm{d}(\theta_q)$ of D_i materializes the breadth of the underlying Gaussian predictive model and thus the confidence in the pixel. We set this value as a robust, conservative function of θ_q which we assume fixed for the purpose of estimation, under small warp perturbations. Note that this is a valid assumption since visibility and grazing angles are generally stable, as we assume given the full poses of the model $\mathscr{M}$ for all frames. We set $\mathrm{d}(\theta_q) = \frac{1}{C} e^{-s\tan\theta_q}$ as a faster approximation of a normal distribution over the angles of the perceived surface, and use C to normalize these weight contributions to 1 among all pixels that see a common texel x to obtain homogeneous weights among pixels in the data term $\sum_{i,q=\tau\circ\pi_i(x)} \mathrm{d}(\theta_q) = 1$. This weight is more conservative than the $\cos\theta_q$ weight usually used for blending in multi-view texturing techniques Debevec et al. [1996], and yields improved results in our experiments, as it downgrades unreliable contributions from surface points at a grazing angle.

Assuming statistical independence of the noise among the low resolution images, a valid assumption due to the independence of the capturing processes among viewpoints, the conditional probability of the set of the low resolution images $\{\mathrm{I}_i\}$ given the high resolution texture T and set of the warping matrices $\{W\}$ is expressed as:

$$\begin{aligned} p(\{\mathrm{I}_i\}|\{\mathrm{W}_i\},\mathrm{T}) &= \prod_{i,t} p(\mathrm{I}_i|\mathrm{W}_i,\mathrm{T}) \\ &= \prod_{i,t} \frac{1}{Z(\mathrm{D}_i)} e^{\sum_{i,t} -(\mathrm{I}_i-\mathbf{A}_i\mathrm{T})^\top \mathrm{D}_i(\mathrm{I}_i-\mathbf{A}_i\mathrm{T})}. \end{aligned} \tag{4.10}$$

4.4.2.2 Texture Model

The distribution $p(\mathrm{T})$ reflects the prior knowledge about the high resolution texture. We assume that the unknown appearance is a natural image.

The characteristic of natural images compared to noisy ones are the specific structures they exhibit. These features are expressed in spatial domain as edges (Takeda et al. [2007]), self similarities (Glasner et al. [2009], Zhang et al. [2010]) as well as strong properties from statistical point of view (Zontak and Irani [2011]-Roth and Black [2009]). To simulate this statistical behavior, studies on the response of the human visual system are carried out (Field [1987]) to

show that the receptive fields of cortex work as localized-oriented-bandpass filters. To model this process, Gaussian distribution of the derivative filter applied on the natural image is often used and it writes as follows:

$$p(I) \propto exp(-\theta||\nabla I||_2^2), \tag{4.11}$$

where θ corresponds to the inverse of variance and ∇I denotes the gradients of the natural image I. This kind of prior knowledge is used in the area of inverse problems, referred also as Tikhonov regularization and its quadratic form drives to closed form solutions. Although it is simple, its quadratic form leads to over smooth image estimates, where edges are over-penalized and thus other more robust models are required.

Generalized Laplacian distributions are used too in a wide range of applications such as image de-noising (Weiss and Freeman [2007], Roth and Black [2009]), de-blurring, and monocular Super-Resolution (Kim and Kwon [2010]) to compensate for the over-smoothing of the Gaussianity assumption. Furthermore, studies on these derivative filters show, that their histograms are exponential with high forth momentum (kurtosis), a statistical behavior, which is well described by generalized Laplacian distributions. To the same family of non-quadratic natural image prior belong Huber functions (Pickup et al. [2007]), bilateral TV priors (Farsiu et al. [2004]). In the current thesis, the natural texture is expected to have smooth regions, interspersed by occasional strong edges. This sparsity of large derivatives is promoted through the usage of the total variation (TV), defined as:

$$p(\mathrm{T}) \propto exp(-\theta TV(\mathrm{T})). \tag{4.12}$$

$TV(\mathrm{T})$ is the Isotropic Total Variation Energy, which is the L1 norm of the vector field ∇T, the sum of the L2 norms.

$$\begin{aligned} \forall f \in \mathbb{R}^2 \, TV_I(f) &= \sum_{i}^{n1,n2} \sqrt{\frac{\partial f[n1,n2]}{\partial x}^2 + \frac{\partial f[n1,n2]}{\partial y}^2} \\ TV_I(\mathrm{T}) &= ||\nabla \mathrm{T}||_1. \end{aligned} \tag{4.13}$$

4.4.2.3 Motion Model

The motion model describes the nature of the warps, which account for the reconstruction and calibration inaccuracies and thus it depends on the conditions of the capturing setup. In the simplest case, considering a planar scene and orthographic projection, the transformation between the camera and the scene is a simple affine matrix with 6 degrees of freedom. For more realistic prespective projection scenario the motion between planar scenes is approximated by quadratic terms; perspective transform or homography with 8 degrees of freedom.

However, in a multi-view setup, the complexity increases; the scene is not planar, motion parallax and occlusion effects are apparent. The previous parametric models cannot capture the motion in this case and more complex models need to be established. Dense optical flow fields allow for parts of the same scene to move to different directions under a more general global assumption and the constraint of a common global transformation, implied through the previous

models, is not anymore valid.

To handle the multi-view complexity, we also introduce optical flow field and its corresponding probabilistic representation to model its statistical behavior. In particular, the nature of the field is imposed by the nature of the geometric noise, which in general exhibits a sparsity. To model these occasional transitions of the flow field W_i, Gibbs probability distribution with potential function the TV norm is proposed, which ensures the necessary sparsity of variations.

$$p(\mathrm{W}_i) = \frac{1}{Z_{\mathrm{W}_i}(\gamma)} e^{-\nu\left(\|\nabla u_i^t\| + \|\nabla v_i^t\|\right)}, \tag{4.14}$$

where ∇ is the gradient operator, u_i^t and v_i^t, the x- and y- components of the warp W_i and $Z_{\mathrm{W}_i}(\gamma)$ denotes the normalization constant.

Under the assumption of statistical independence of the capturing processes among viewpoints, the probability of the set of warp fields $\{\mathrm{W}_i\}$ is expressed as:

$$\begin{aligned} p(\{\mathrm{W}_i\}) &= \prod_{i,t} p(\mathrm{W}_i) \\ &= \prod_{i,t} \frac{1}{Z_{\mathrm{W}_i}(\gamma)} e^{-\nu\left(\|\nabla u_i^t\| + \|\nabla v_i^t\|\right)}. \end{aligned} \tag{4.15}$$

4.4.3 Bayesian Framework

Casting the problem into Bayesian framework and using the probability distribution of the observation model in 4.4.2, the Maximum a Posteriori solution is presented through our proposed two-stage iterative framework. Prior knowledge is imposed through the explicit representations of the texture and motion model, which are used as regularization parameters to restrict the space of solutions.

4.4.3.1 Maximum A Posteriori Solution

In order to constraint the space of solutions and to obtain a unique texture, the prior probability distributions of the texture model and the motion model are used through the maximization of the posterior probability distribution to regularize the problem.

The maximum a posteriori estimators are written as follows:

$$\begin{aligned} \{\mathrm{T}^*_{\mathrm{MAP}}, \{\mathrm{W}^*_{\mathrm{iMAP}}\}\} &= \arg\max_{\mathrm{T},\{\mathrm{W}_i\}} p(\mathrm{T}, \{\mathrm{W}_i\} | \{\mathrm{I}_i\}) \\ &= \arg\max_{\mathrm{T},\{\mathrm{W}_i\}} \frac{p(\{\mathrm{I}_i\} | \{\mathrm{W}_i\}, \mathrm{T}) p(\{\mathrm{W}_i\}, \mathrm{T})}{p(\{\mathrm{I}_i\})} \\ &= \arg\max_{T,\{\mathrm{W}_i\}} \prod_i p(\mathrm{I}_i | \mathrm{T}, \mathrm{W}_i) P(\mathrm{T}) \prod_i p(\mathrm{W}_i). \end{aligned} \tag{4.16}$$

This is equivalent to minimizing the minus log of 4.16, which is expressed

$$\{\mathrm{T}^*_{\mathrm{MAP}}, \{\mathrm{W}^*_{\mathrm{iMAP}}\}\} \quad = \quad \mathrm{arg\,min}_{\mathrm{T},\{\mathrm{W}_i\}} - log \prod_i p(\mathrm{I}_i|\mathrm{T},\mathrm{W}_i) - log\, p(\mathrm{T}) - log \prod_i p(\mathrm{W}_i). \tag{4.17}$$

Inserting the analytic formulations of the distribution of the likelihood 4.10 , of the prior knowledge for the texture 4.13 and for the motion field 4.15 the equation 4.17 is written as:

$$\{\mathrm{T}^*_{\mathrm{MAP}}, \{\mathrm{W}^*_{\mathrm{iMAP}}\}\} = \mathrm{arg\,min}_{\mathrm{T},\{\mathrm{W}_i\}} \sum_i (\mathrm{I}_i - \mathbf{A}_i\mathrm{T})^\top \mathrm{D}_i(\mathrm{I}_i - \mathbf{A}_i\mathrm{T}) + \theta TV_I(\mathrm{T}) + \sum_i \nu(\|\nabla u_i\| + \|\nabla v_i\|). \tag{4.18}$$

Solving the above minimization problem is challenging since it requires a careful algorithmic approach due to the non differentiable behavior with respect to the texture and the motion field. A coordinate-descent optimization procedure is followed, where the function is minimized with respect to the motion field given the texture and then it is minimized with respect to the texture given the computed motion field.

- **Warp Estimates.**

  Each W_i for an input view $\{i\}$ is independently estimated. Minimizing the negative log of 4.18, and dropping all terms independent of W_i yields:

$$\begin{aligned} \mathrm{W}^*_{\mathrm{iMAP}} = \mathrm{arg\,min}_{\mathrm{W}_i}\ & \nu(\|\nabla u_i\| + \|\nabla v_i\|) \\ & + (\mathrm{I}_i - \mathbf{SKF}_{\mathrm{W}_i}\mathbf{P}_i\mathrm{T})^\top \mathrm{D}_i(\mathrm{I}_i - \mathbf{SKF}_{\mathrm{W}_i}\mathbf{P}_i\mathrm{T}), \end{aligned} \tag{4.19}$$

  which can be interpreted as a modified optical flow equation with a TV-regularizer, where the data term is re-weighted by D_i. The intuition here is that the minimization favors the TV prior of sparse variation over the data term for untrustworthy pixels according to D_i, and puts more relative importance on trying to follow data on reliable pixels. We opt for a similar strategy to Liu and Sun [2011] for solving this equation, and initialize the estimation of W_i with the result of a standard optical flow method Liu [2009], applied between H_i and an upsampled I_i at each iteration.

- **Appearance Map.**

  Minimizing the negative log of 4.18 and dropping all terms independent of T yields:

$$\mathrm{T}^*_{\mathrm{MAP}} = \mathrm{arg\,min}_{\mathrm{T}} \sum_i (\mathrm{I}_i - \mathbf{A}_i\mathrm{T})^\top \mathrm{D}_i(\mathrm{I}_i - \mathbf{A}_i\mathrm{T}) + \lambda\|\nabla\mathrm{T}\|, \tag{4.20}$$

  where the data term develops to a weighted sum of per-pixel L_2-norms. Although not specifically using a robust data term norm here as opposed to some works, we nevertheless obtain excellent results enforcing robustness through the constant covariance matrix D_i, as will be shown. Optimizing a L_2 data term with a TV-regularizer has been specifically studied Chambolle [2004], yielding a family of forward-backward splitting solvers whose

implementation are available off-the-shelf Combettes and Pesquet [2011]. Let us note $f_d(\mathrm{T})$ and $f_{\mathrm{TV}}(\mathrm{T})$ the data and the TV-term. Forward-backward splitting is an iterative algorithm for estimating T, alternating between computing a gradient update step and projection $\mathrm{prox}_{\gamma, f_{\mathrm{TV}}}$ which computes an implicit subgradient descent step for the TV-norm.

$$\mathrm{T}_{n+1} = \mathrm{prox}_{\gamma, f_{\mathrm{TV}}}\left(\mathrm{T}_n - \gamma \nabla f_d(\mathrm{T}_n)\right), \tag{4.21}$$

where γ is a step-size parameter. Our re-weighted functional (4.20) only implies a modification of the gradient update with respect to the standard case, with $\nabla f_d(\mathrm{T}_n) = 2\mathbf{F}_{\mathrm{W}_i}{}^{\top}\mathrm{D}_i(\mathbf{F}_{\mathrm{W}_i}\,\mathrm{T}_n - \mathrm{I}_i)$.

These alternating minimization steps will continue until the algorithm converges or until a maximum number of iterations is reached. By jointly solving for both texture and warps we avoid local minima and we successfully solve for the unique super-resolved appearance.

4.5 Experimental Evaluation

To validate the estimation quality of our method, we apply the estimation pipeline on different datasets and under several scenarios, we then regenerate images at the given viewpoints with the computed super-resolved appearance and we compute reconstruction errors. Objective evaluation however of the visual quality is inherently difficult in Super-Resolution frameworks. There is still an active area of research to study the relation between image quality and human perception and to introduce metrics that correlate well with the human performance. We use the mean square error (MSE) and the structural similarity index (SSIM) as a more perceptual metric and we compute them on image space. All of the MSE and SSIM estimates are computed in the texture domain and only among actual silhouette pixels in the image domain.

We study in particular the robustness of our model to several up-sampling factors, to different noise levels originated from both the capturing process as well as from reconstruction and calibration inaccuracies. We quantitatively evaluate the aforementioned aspects using synthetic datasets and we show qualitative results comparing our proposed algorithm to the state of the art. We assess the performance in more challenging setups.

We exhibit results with a MATLAB prototype implementation, and run experiments on a 16-core 2.4 GHz PC with 32GB RAM[1]. Our current implementation is mainly mono-thread, with the exception of the optical flow which we launch in 10 separate threads. To initialize the algorithm, we first use a small C++/OpenGL program to render visibility maps from texture to image space, then initialize the texture map with a simple weighted average of visible inputs. The visibility maps are also used to generate each projection matrix operator $\mathbf{P}$. We use the Optical Flow package from Liu *et al.* Liu [2009] for per-iteration optical flow initialization, and the UNLocBOX package Combettes and Pesquet [2011] for the texture re-estimation in the loop. We use a threshold on the relative norm of the objective function (4.20) as stopping criterion, and observe convergence in 30 to 70 iterations for a given λ.

[1] See video results at `http://hal.inria.fr/hal-00977755`

Parameter values. We set the Gaussian variances with $\sigma_p = 0.25$ and $\sigma_k = 0.1$, respectively for the projection weight and PSF kernel $\mathbf{K}$, for all datasets. Although these could be optimized alongside other parameters, we observe low sensitivity to these parameters when set in the $[0.1, 1]$ range. Higher values introduce over-blurring, while lower values tend to reveal the underlying discretization of the texture map ($\sigma_p < 0.1$) or the input image ($\sigma_k < 0.1$). The convergence parameters γ and λ are fixed according to the dataset.

4.5.1 Synthetic Dataset

Experiments on synthetic dataset are carried out and they are organized into three categories. The first one validates quantitatively the robustness of our Super-Resolution appearance estimation pipeline to different magnification numbers and its out-performance compared to simple interpolation technique. The second class of experiments simulate the generation of noise related to capturing process and demonstrate the response of our algorithm to different noise levels. Third, the geometric noise is studied and the appearance estimation is validated under different scenarios.

A synthetic TORUS dataset is chosen to be generated for the purpose of these experiments. The choice of such a simple geometric object is explained by the fact that there is one conformal mapping for the whole object and no cuts are needed. The texture map is thus continuous and the Super-Resolution appearance estimation is tested in a simplified scenario. A single texture is warped around the torus and images are ray-traced from 48 different locations, distributed evenly. The resolution of the rendered images is 512×512 pixels and the resolution of the original and the reconstructed texture is 1024×1024 pixels.

4.5.1.1 Magnification Number

To test how the Super-Resolution appearance estimation responds to different image resolutions, the rendered images are down-sampled by a factor of 2 and the algorithm is applied on the original and the down-sampled ones. The initialization of the texture is computed as a weighted arithmetic mean of all the image contributions in the texture space according to 3.4.2 In this case, the convergence parameters are fixed to $\gamma = 0.5$ and $\lambda = 5.10^{-3}$ for all experiments, using a second and third round of iterations with $\lambda = 5.10^{-4}$ and $\lambda = 5.10^{-5}$ to down-weigh TV-regularization and thus reveal higher frequency detail. The performance of the algorithm is validated for magnification number 2 and 4. The estimated reconstructed textures as well as the initial textures are used to generate at each view $\{i\}$, according to the sparse linear operator $\mathbf{A}_i = \mathbf{SKF}_{\mathrm{W}_i}\mathbf{P}_i$, image of resolution 512×512 pixels and 256×256 pixels accordingly. The experiment is repeated for 3 textures, which exhibit different characteristics.

In Figures 4.3 and 4.4, image at a specific view, generated with the initial texture and the reconstructed is compared with the input ground truth image. The initialization leads to blurry results and visual details are lost, whereas the quality of super-resolved texture is comparable to the ground truth. For each comparison mean value of mean squared error and of the structural similarity metrics as well as the standard deviation are also recorded. While the magnification factor increases, the reconstruction of the texture becomes more sensitive, which is also justified through the increasing MSE, and decreasing SSIM but still the super-resolved reconstructed

texture is still comparable to the ground truth.

To demonstrate the difference between Super-Resolution with simple image enhancement methods, discussed in 4.2.1, similar experiments with the previous one are carried out. The synthetic TORUS dataset is used and in addition, images with resolution 1024×1024 pixels are rendered given the ground textures. Super-Resolution appearance estimation is applied for each case of magnification number 2 and 4 and the reconstructed textures are used to compute images of size 1024×1024 pixels. Input images are then up-sampled through bilinear interpolation and both are compared with ground truth rendered ones.

In Figures 4.5 and 4.6 results are compared with up-sampled bilinear interpolated input images. While the difference between the results from bilinear interpolation and Super-Resolution is not so indistinguishable, the recored metrics show that SR leads to better estimations. Indeed, the close ups, show that Super-Resolution estimation leads to crisp texture maps and the computed images are more focused.

4.5.1.2 Noise Level

The second category of experiments demonstrates how the Super-Resolution estimation responds to the noise originated from the capturing process. This kind of noise includes thermal noise, color filtering noise and it is expressed as an additive term in equation 4.4 which follows Gaussian distribution with zero mean value. To test how the how the system performs under the various levels of this kind of noise, we use the synthetic TORUS dataset, with images of size 512×512 rendered with one chosen texture. We then add to the synthetic ground truth images white Gaussian noise with standard deviation σ_n and we apply for each case the super-resolution algorithm.

When $\sigma_n = 0$ the quantization is the only source of error, which has been studied in the previous experiments and is neglected at that point. Figure 4.7 shows that SR appearance estimation successfully compensates the noise. While the standard deviation increases and the input images are more noisy the general performance decreases, which explains the decrease in the metrics. Note that in that case convergence parameters are carefully selected in order for the algorithm to compensate for the noise and regain details. To achieve faster convergence in both cases we fix $\gamma = 0.5$. In the case where $\sigma_n = 0.01$, the parameter for the regularization term is set $\lambda = 5.10^{-2}$ using a second round with $\lambda = 5.10^{-3}$. When the noise is stronger $\sigma_n = 0.05$ higher weight needs to be given the TV regularized in order to achieve a reconstruction with stronger TV behavior. Interestingly, this could also be expressed as, the higher the noise is, the less the observations should be trusted. This however does not limit the performance of the proposed method, since knowledge about the noise level can always be retrieved from the input given images.

4.5.1.3 Geometric Noise Level

The third category of experiments assesses the SR estimation with respect to geometric noise. Both reconstruction and calibration noise are two forms of geometric noise, which have been discussed in 4.3.1. To justify our statement that the proposed super-resolution model compensates for this kind of noise, experiments are carried out using the TORUS dataset with the images

of 512×512 pixels. We argue that the noise in the calibration parameters is expressed as misalignment through the projection and thus is of the same nature as the geometric one, on which we will focus. The geometric noise is then simulated by adding on the mesh multivariate white Gaussian noise with standard deviation σ_G. When $\sigma_G = 0$ it boils down to the case of ideal geometry scenario, that was studied previously. While we increase the standard deviation we generate noisy instances of the torus. SR algorithm is then applied on the ground truth input images and the noisy mesh. The computed images with the reconstructed appearance are then compared to the input ground truth ones. Figure 4.8 demonstrates the ability of the proposed method to overcome this noise and to result in high visually textures and images accordingly.

4.5.2 Comparison with State of the Art

We compare our model with the latest state of the art multi-view texture super-resolution technique of Goldluecke et al. [2013]. The comparison is performed on the common applicability domain, *i.e.* static images, as shown in Figures 4.9 and 4.10. The authors provide a public dataset for three objects BEETHOVEN, BUNNY and BIRD, and kindly provided additional data on request, so we could reproduce the experiment in the closest possible setup. This included a high resolution output of their algorithm for the viewpoint originally reported per dataset in Goldluecke et al. [2013], to which we compare our high resolution output. We use the same super-resolution ratio of $3\times$ the input resolution for the texture and high resolution image domains. Respectively 108 and 52 calibrated viewpoints were originally used at resolution 768×576. We have used identical views, and also use identical 3D models except for the BIRD dataset, for which we observed large reconstruction and silhouette reprojection artifacts on the model provided. In fairness we thus only provide crops in regions where the 3D model geometry is not significantly different.

It can be generally observed that our outputs provide lower noise levels and artifacts. This is particularly visible in the BUNNY dataset, in the ear region and shadow region around the left eye. The BEETHOVEN exhibits some visibility difficulties due to the face geometry and presence of concave regions around the nose and hair, which generate artifacts for Goldluecke et al. [2013]. In contrast, our method is able to deal with these situations efficiently. A single texture domain cut is present on the nose but the discontinuity is barely visible thanks to the inter-chart terms we introduced. More accurate details and sharper pattern borders can be observed on the BIRD wing and tail, notably in the feather textures.

4.5.3 Additional Real-World Dataset: Is SR always necessary?

A question that is naturally raised, is if the Super-Resolution appearance estimation is always needed. One could expect that, when the object is captured from high definition cameras simple blending methods of high resolution images could be adequate enough to regain appearance details.

This however does not hold. In the theoretical analysis of Section 4.3, we claim that the proposed SR model always leads to estimations with finer details. It is the underlined principle of existing sub-pixel misalignments, due to which hidden visual information is revealed.

To demonstrate this we introduce TOMAS dataset, which consists of 68 images of size 2048×2048 pixels, captured in KINOVIS studio (see Figure 4.11). There are two main challenges

behind this dataset. First, the resolution of input observations is high, that in turn automatically increases the base line, which Super-Resolution appearance estimation needs to exceed. Second, there are many high frequency details, which appearance estimation algorithm should reconstruct.

Super-Resolution appearance estimation is applied to compute a texture with magnification number 2; the dimension of the computed super-resolved appearance is 4096×4096. Images are then rendered at the same view points and compared to the input and the ones computed with the initialized texture. SR successfully regains additional details compared to the input images. In Figure 4.12, close ups show that new hidden information is retrieved even in the case where the input observations are of high quality and demonstrate that Super-Resolution is always informative regardless the input observations.

4.5.3.1 Influence of Texture Prior

We here demonstrate how the prior information over the texture influences the SR estimation. Knowledge over the unknown appearance is expressed in the SR estimation through the λ parameter. High values of this parameter correspond to textures with larger smooth patches, while low values favor textures with finer details, namely small smooth patches with strong edges.

Applying SR appearance estimation with decreasing λ parameter can be seen in the context of coarse to fine approaches and a similar paradigm can be established. As a first step, in order to smooth out the noise a large parameter λ is defined and thus a smooth texture is computed as solution. Then for the next steps, the coarse computed texture is set as initialization and the algorithm is applied by imposing this time smaller contribution of the TV prior. This whole iterative process can be continued until a visually appealing result is achieved.

SR appearance estimation is applied on TOMAS dataset three times sequentially while decreasing λ parameter. Close ups of Figure 4.13 shows the details computed at the end of the last round.

4.6 Discussion

Through the previous evaluation we prove experimentally that the proposed model succeeds to combine the observations of a 3D object and to compute a high-resolution appearance. Introducing a single appearance space and expressing the input low resolution images as noisy observations we build a linear model which exploits the non-redundant appearance information. The experiments on the synthetic dataset demonstrate the robustness of the model to all the sources of noise that appear through the capturing process in the multi-view case.

Although the noise that appears due to the 3D nature of the problem (aka geometric and calibration noise) adds an additional degree of difficulty, it enhances at the same time the effect of aliasing. This translates to more propagated information, new hidden appearance shared among the views and thus super-resolution is possible.

The stochastic approach that was adopted in the current super-resolution multi-view framework allows to introduce the uncertainty of the noise as well as the probabilistic nature of the unknown appearance. The prior information of the unknown texture and the motion field regularizing the

problem and a unique solution is provided for the appearance. The coarse to fine new paradigm that is established through the regularization parameters enables to compute textures with even more details. A natural extension of the current framework would be to learn these parameters as well as to incorporate in the inference procedure the computation of the blurring kernel which in the current work is assumed known. It would be also of great practical value to theoretically predict how much can the appearance of 3D objects be super-resolved.

4.7 Conclusion

We have presented a novel framework to efficiently model the appearance of subjects observed from multiple viewpoints. It is a compact representation, which builds on the idea to express all the appearance information shared among the views in a common texture space and to super-resolve it. The unknown appearance (texture) which is formulated as an image (vector) generates the image observations at different viewpoints and compensates for all sources of noise through a motion field which is embedded in the linear operator.

The availability of multiple observations corresponds to hidden non-redundant appearance information which makes the super-resolution feasible. Due to geometric, calibration inaccuracies related to the 3D nature of the problem and due to noise originated from the 2D nature of the problem aliasing effects are prominent. Super-resolution however benefits from them and a high detailed appearance is computed.

The idea of exploiting the non-redundant appearance information to solve for a high resolution compact appearance opens a new research direction towards the exploitation of even more appearance samples available in time. Appearance information is also fused in time and can be accessed through video sequences, facts that give rise to the next research question. Taking full advantage of available appearance observation shared among views and fused in time becomes the subject of the following chapter.

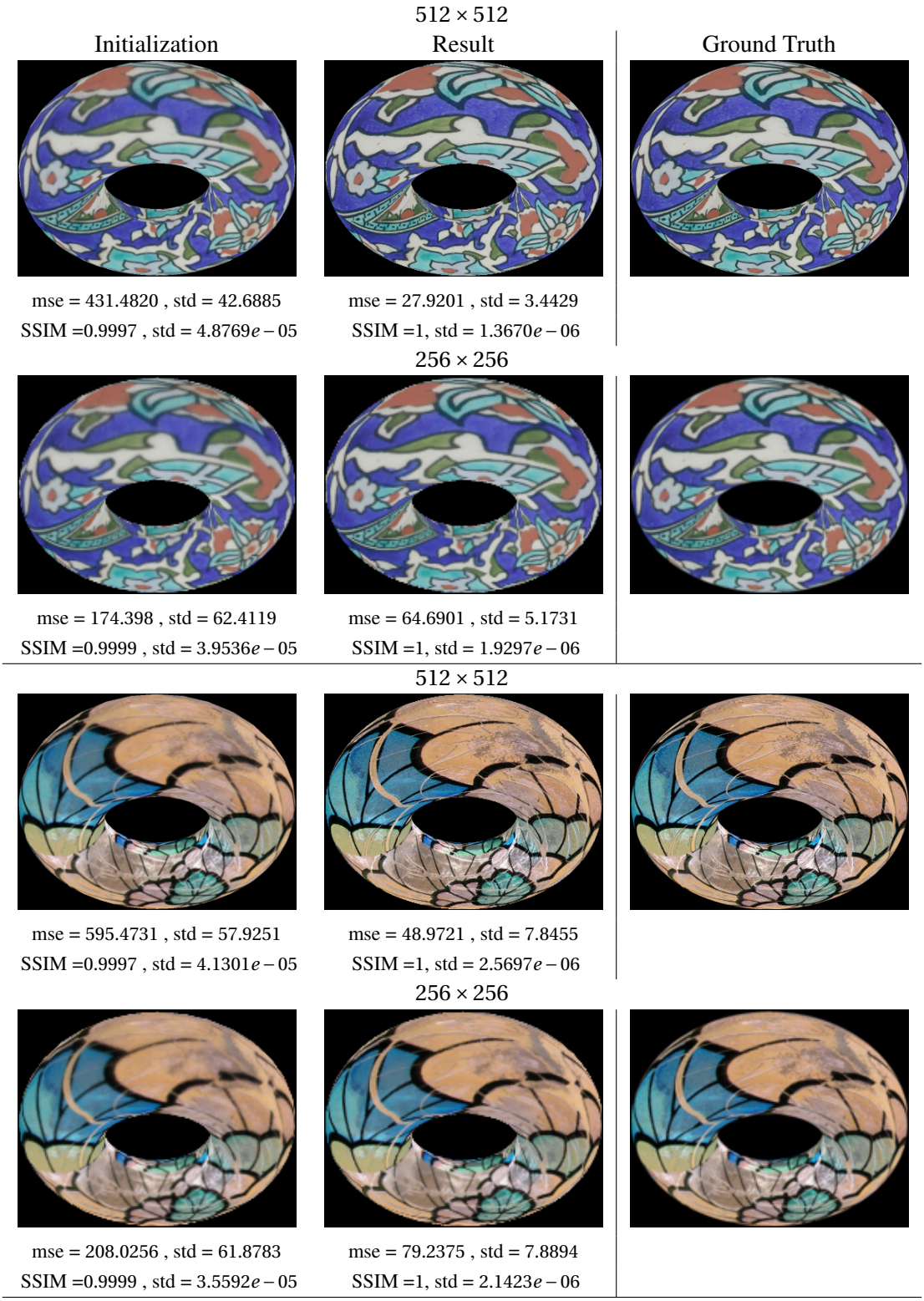

Figure 4.3 – Performance evaluation with respect to magnification number on TORUS dataset. For magnification number 2, the images reconstructed by the super-resolved texture are indistinguishable from ground truth.

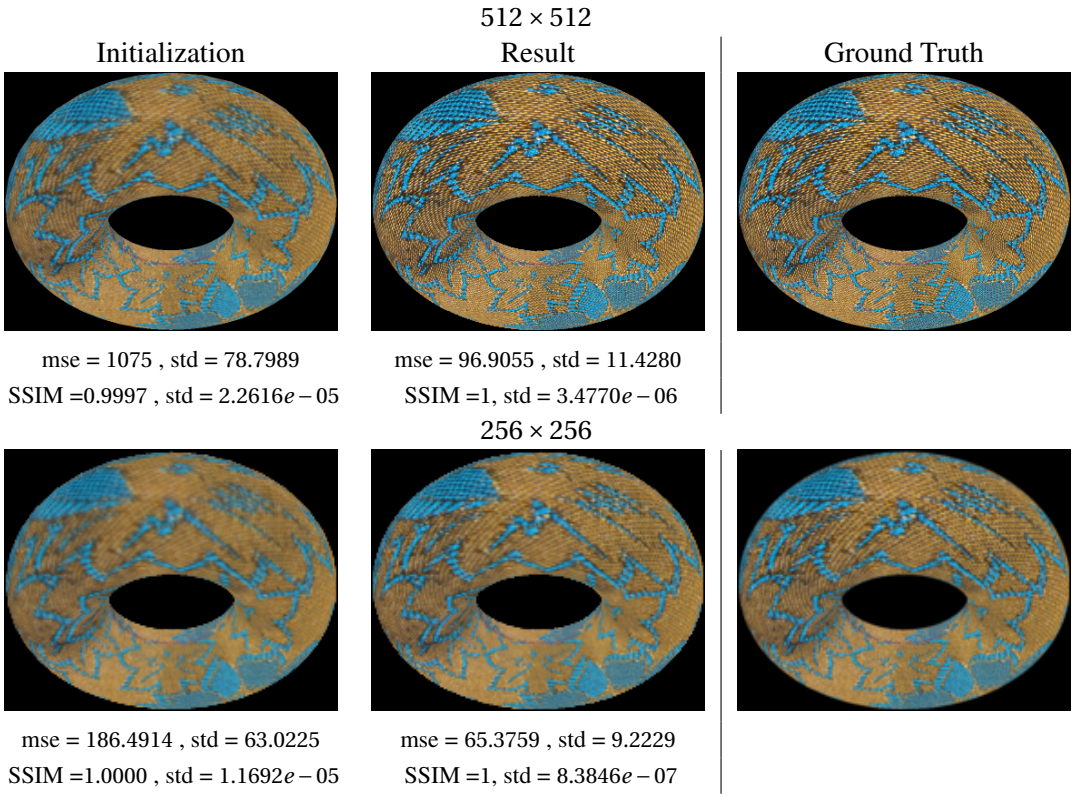

Figure 4.4 – Performance evaluation with respect to magnification number on TORUS dataset for a texture with finer details. Even with a more challenging texture, justified also by the increase of both MSE and SSIM metrics in the initialization step,the generated images with the super-resolved texture are indistinguishable from ground truth.

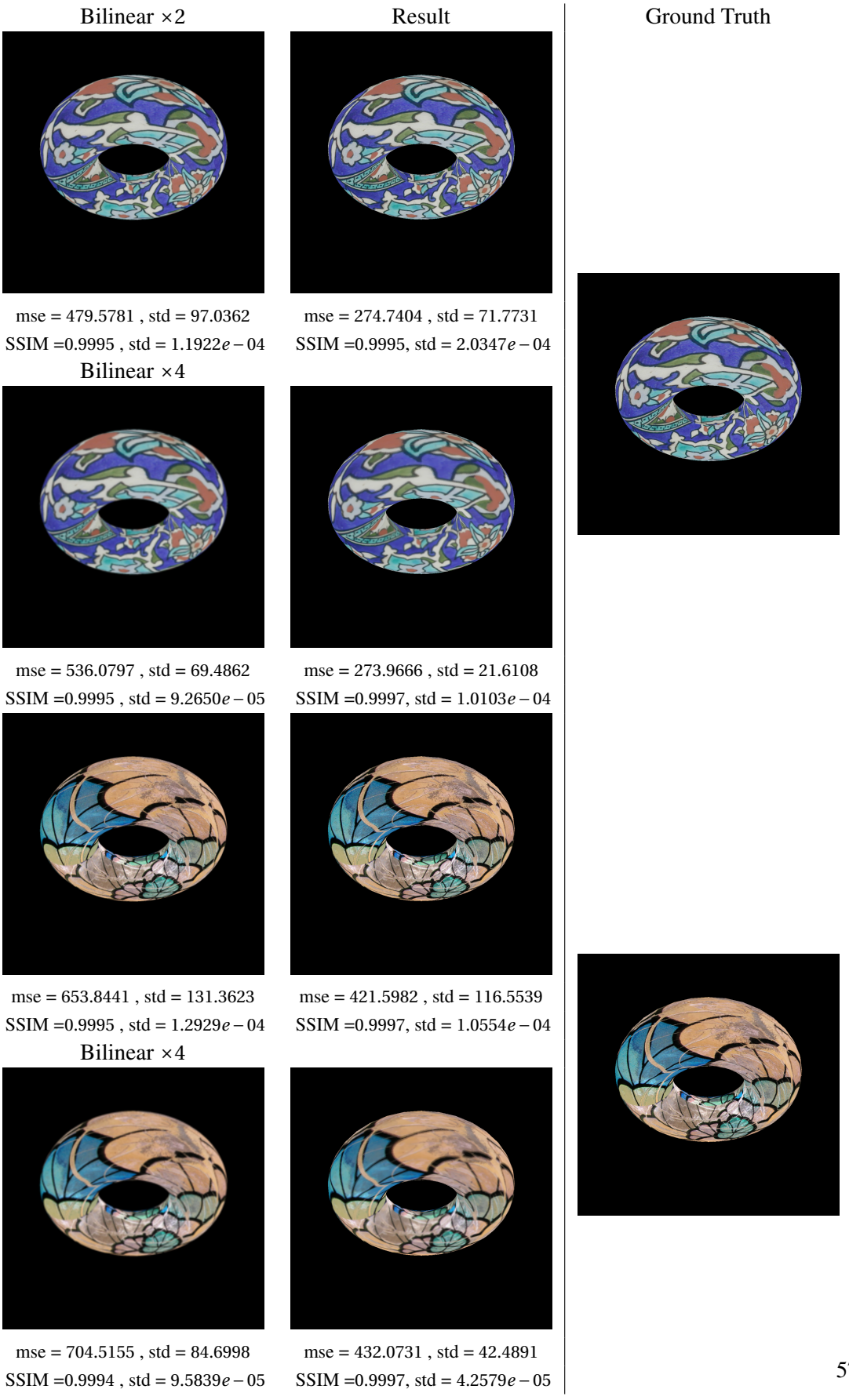

Figure 4.5 – Comparison on TORUS dataset for two textures. While the magnification number increases bilinear interpolation leads to more blurry results compared to SR.

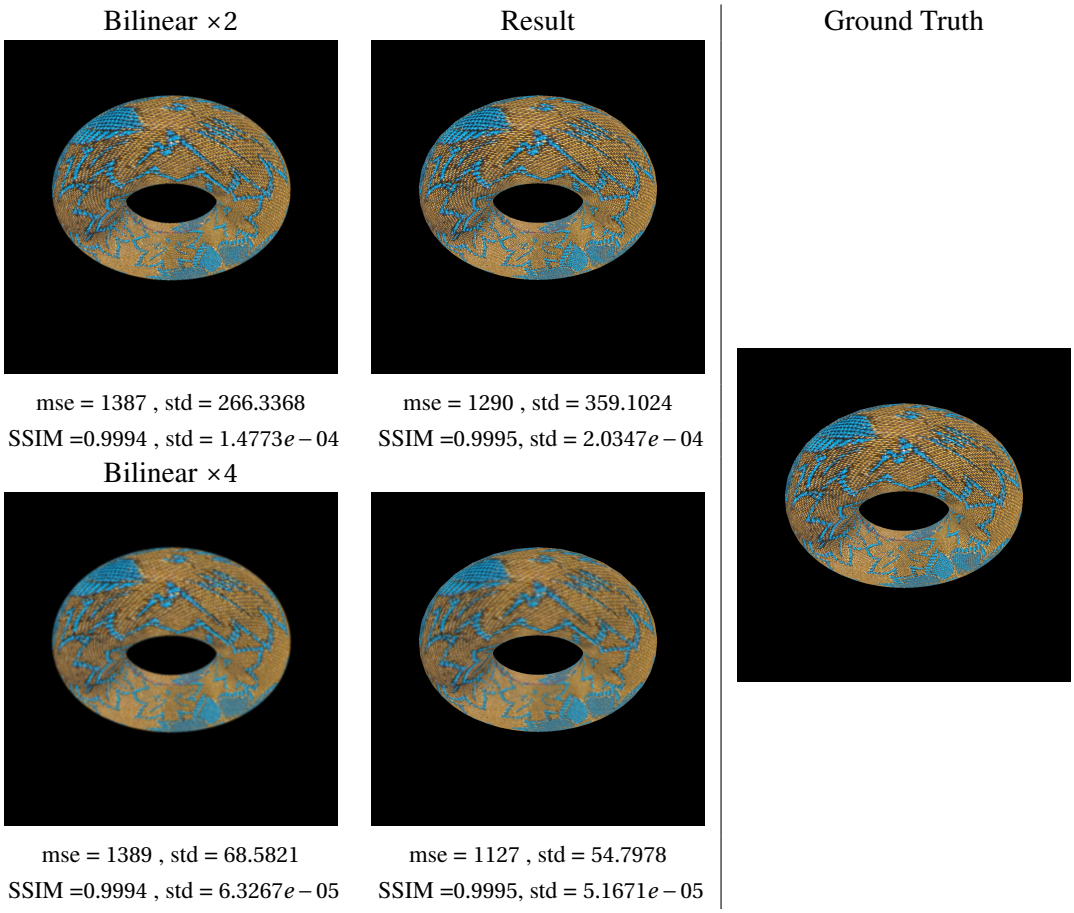

Figure 4.6 – Comparison on TORUS dataset for a texture with finer details. Even with a more challenging texture, justified also by the increase of both MSE and SSIM metrics in the initialization step,the super-resolved texture generates images with high frequency details.

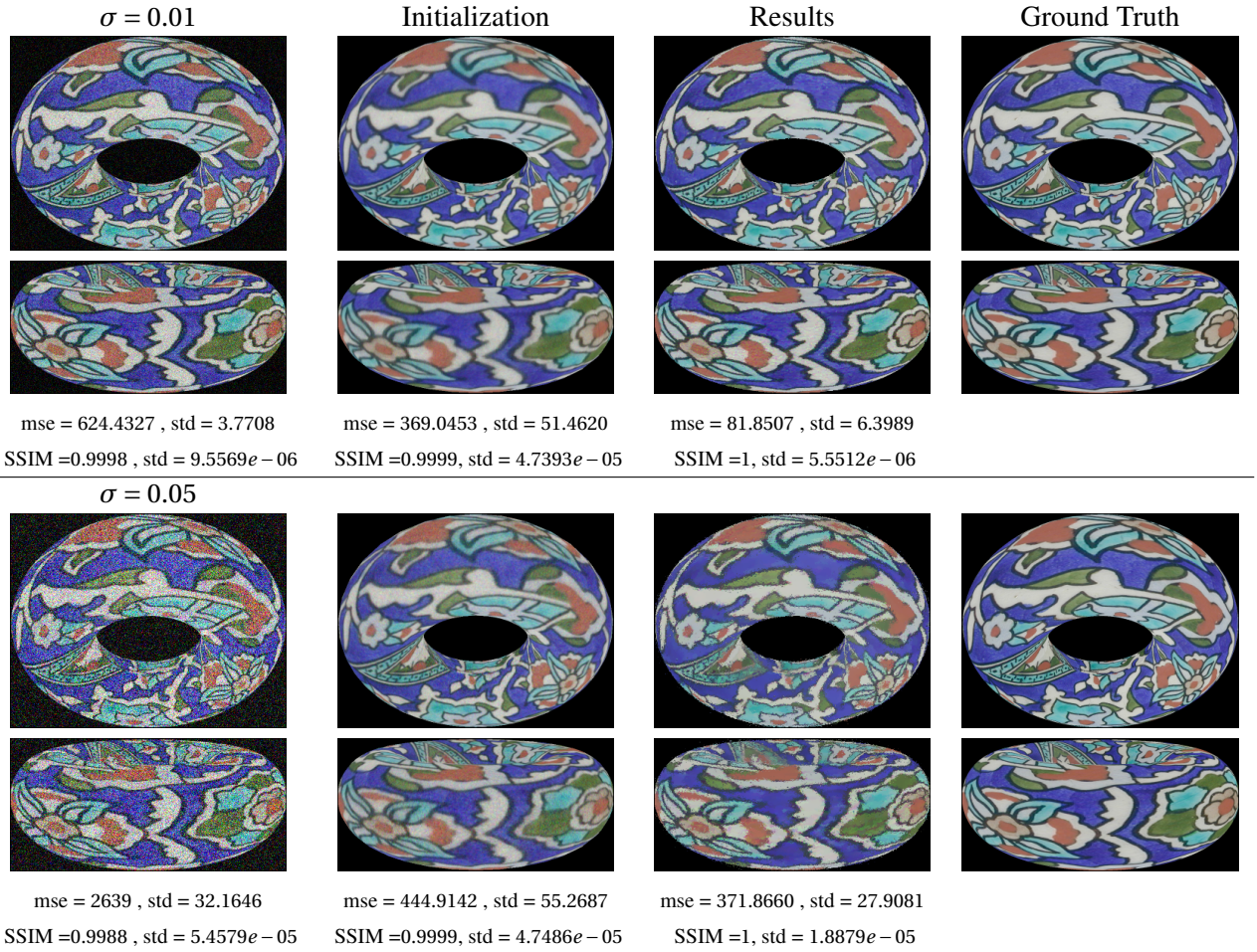

Figure 4.7 – **Our Super-Resolution system compensates for white Gaussian noise.** Multivariate Gaussian Noise with zero mean value and varying σ is added on the ground truth texture and images are rendered at different viewpoints.

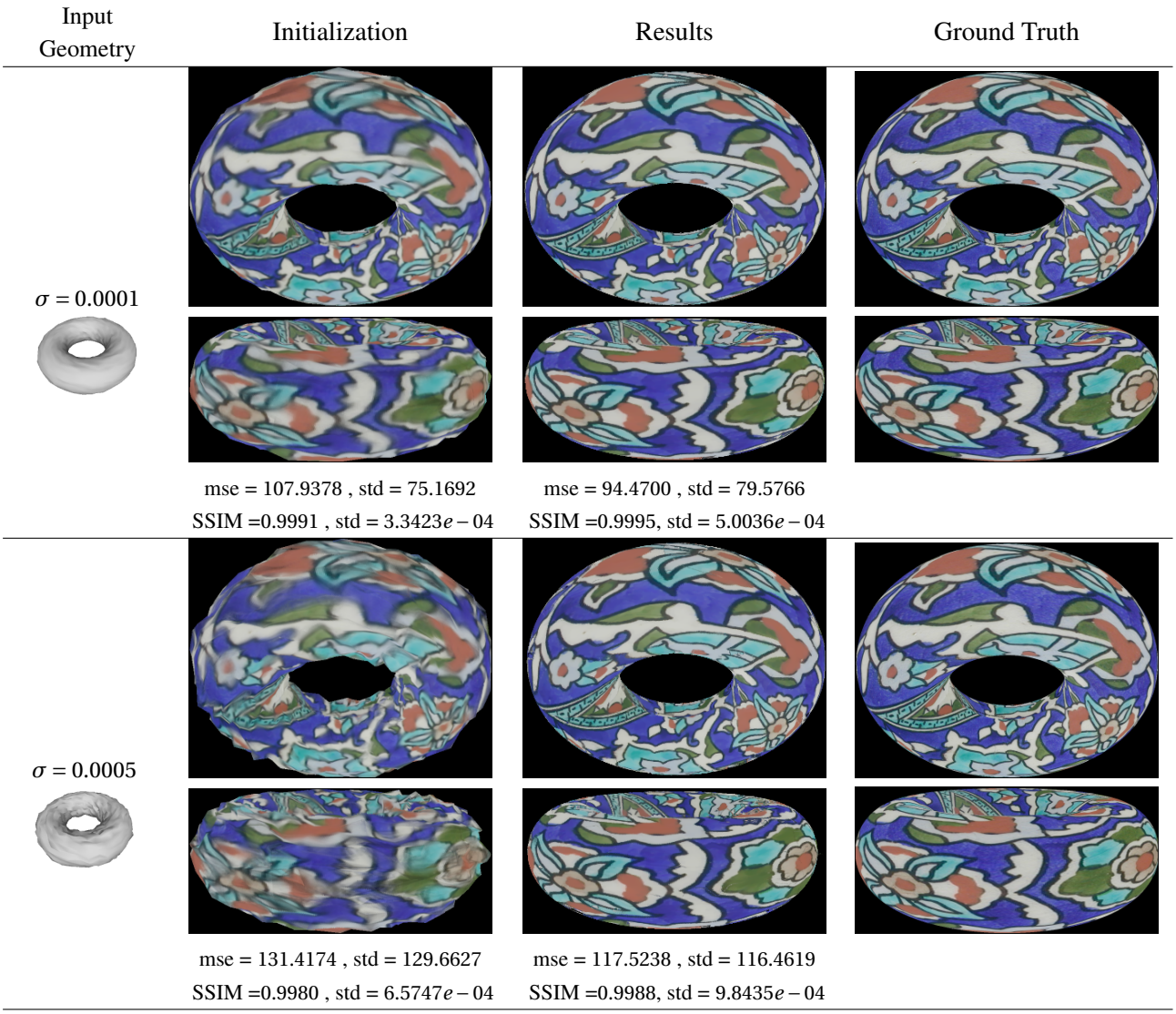

Figure 4.8 – **Our Super-Resolution system compensates for geometric noise.** Multivariate Gaussian Noise with zero mean value and varying σ is added on each vertex of the Torus to simulate the geometric noise in the reconstruction step.

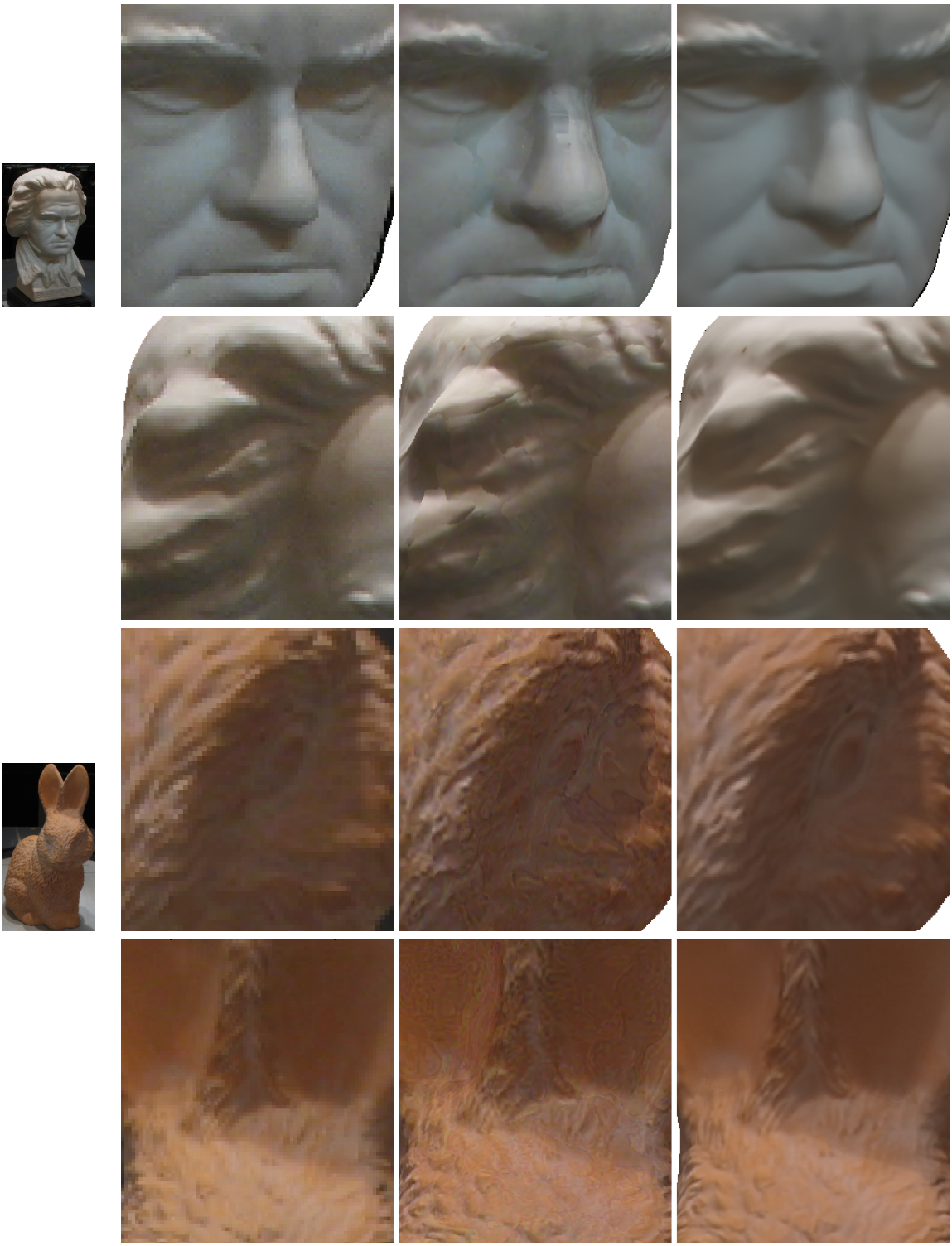

Figure 4.9 – Comparison on BEETHOVEN and BUNNY datasets. Left column: input images. Middle: output of Goldluecke et al. [2013]. Right: our algorithm. Best viewed magnified and in color. Differences on the BEETHOVENs' nose are due to the 3D model that was provided by the authors of Goldluecke et al. [2013] and kept identical for a fair comparison.

Figure 4.10 – Comparison on BIRD dataset. Left column: input images. Middle: output of Goldluecke et al. [2013]. Right: our algorithm. Best viewed magnified and in color.

(a) Capture Studio

(b) Thomas is captured from 68 cameras. Captured images have size 2048 × 2048 pixels.

Figure 4.11 – Acquisition of TOMAS dataset.

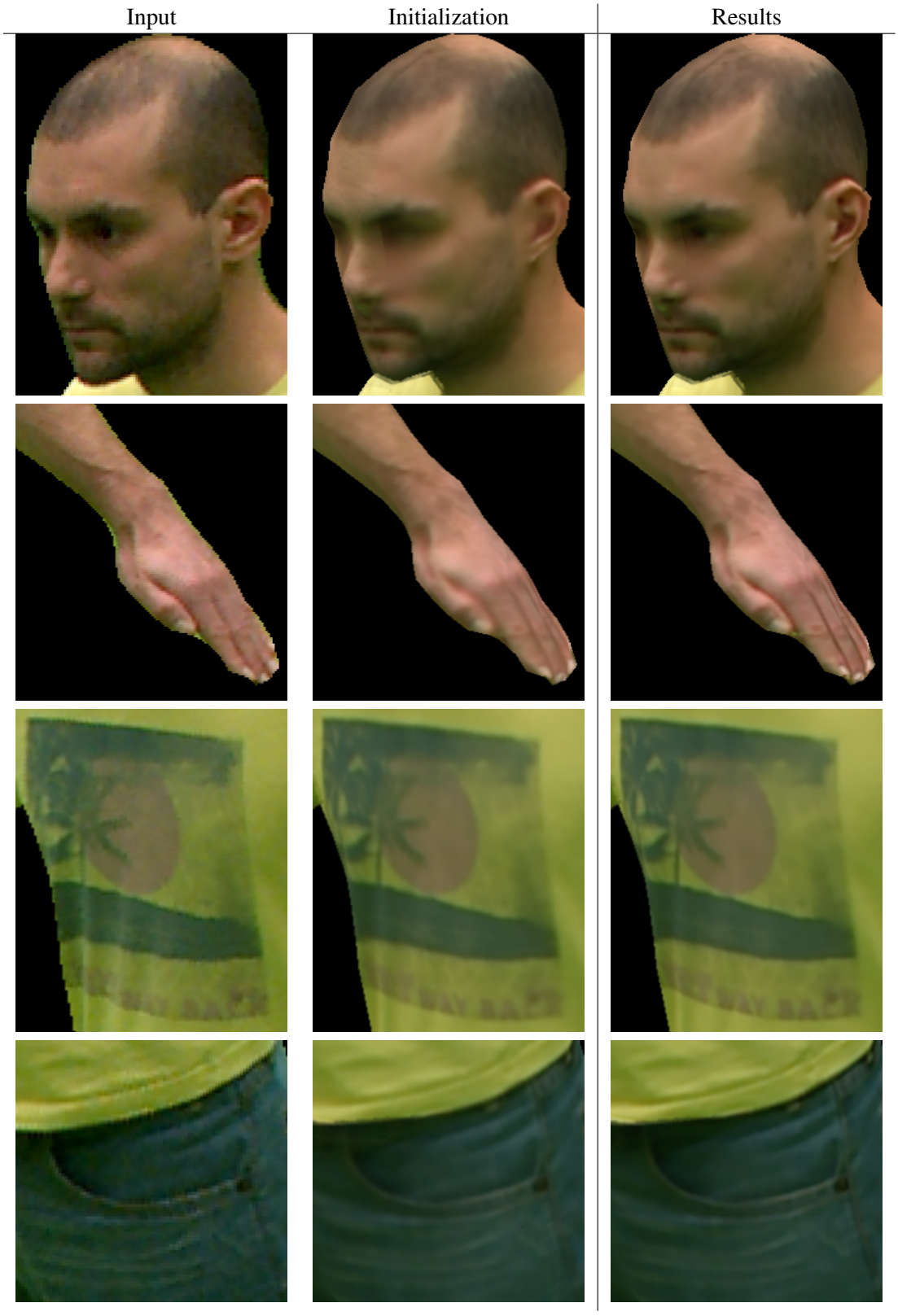

Figure 4.12 – From top to down: Detailed and focused areas around the eye and the ear. Additional veins on the arm and wrinkles on the fingers. Letters "WAY BACK" and leaves of the second tree top right. Wrinkles of T-Shirt and pocket.

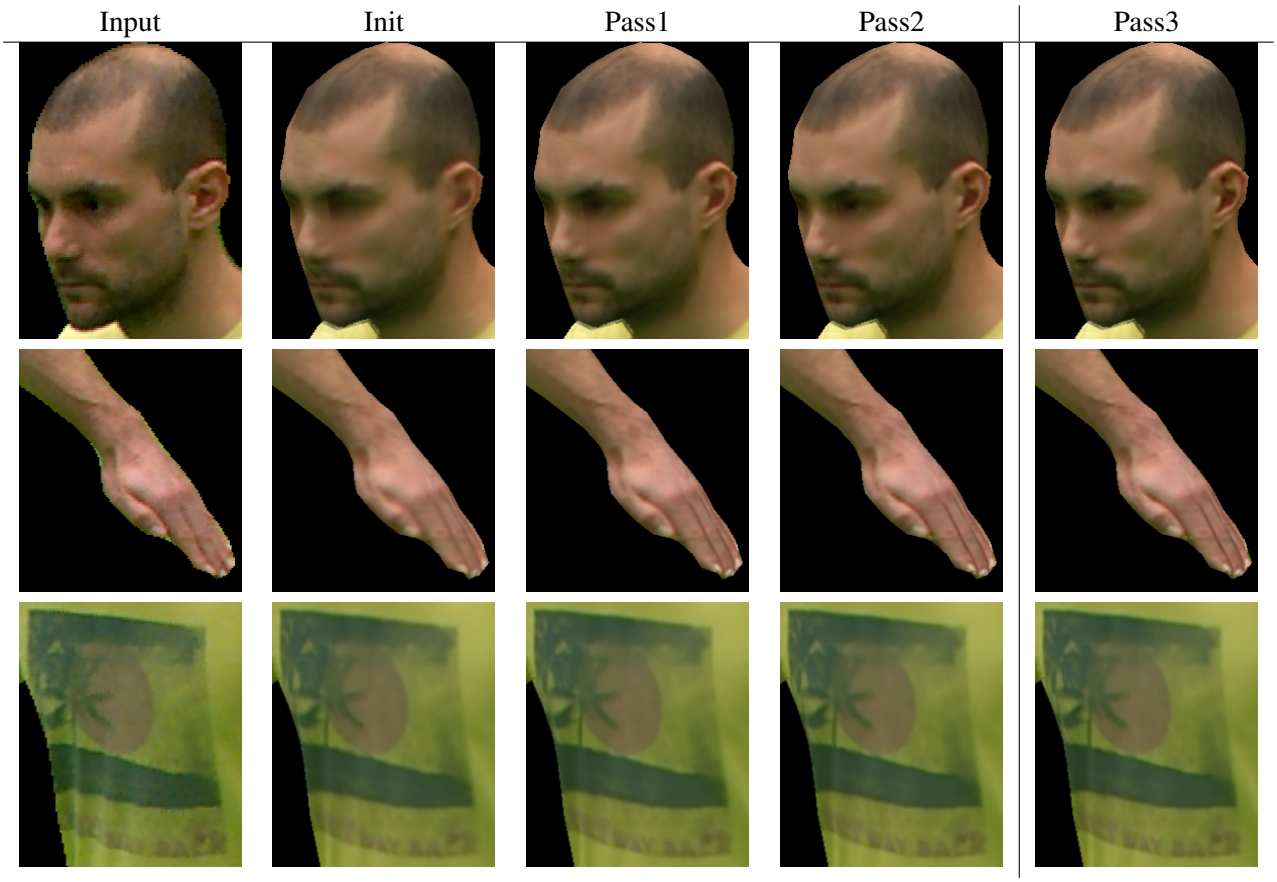

Figure 4.13 – Comparison on λ parameter. While λ decreases from left to right details are further revealed. Pass1 corresponds to $\lambda = 5e-04$, Pass2 to $\lambda = 5e-05$, and Pass3 to $\lambda = 5e-06$.

5 Temporally Coherent Super-Resolved Appearance Model

5.1 Introduction and Motivations

Through the Super-Resolution appearance model, the detailed appearance of a subject is successfully computed. This principle of SR technique to exploit the non-redundant appearance information of the input images, captured at multiple viewpoints, can be further extended and motivates the study of the current chapter.

We propose a temporally coherent super-resolved appearance model, which successfully generalizes our previous multi-view appearance super-resolution work to the temporal domain and regains higher quality. The appearance of a subject does not change dramatically through small predefined time intervals. In such cases, where lighting conditions remain the same and no significant movements takes place, the appearance is considered to retain in overall the same characteristics. Exploiting the appearance information of the subject that is not only shared among the views but also is fused in time in order to compute a detailed appearance is the fundamental of the proposed model. The motivation of inserting the time domain in the model is twofold. First, several such input video streams are available in the multi-view context. Second, new misalignments are caused by the fusion of the common appearance over time and thus new non-redundant information is available. The proposed model is view-independent, compact, linear, which are all properties inherited by the SR model presented in Chapter 4. The major contribution is the ability to exploit viewpoint as well as time appearance redundancy. As opposed to the work of Tung [2008], where video sequences are super-resolved at each viewpoint, our proposed model utilizes the video cues in a view-independent framework. Contrary to the work of Goldluecke and Cremers [2009a], which only deals with the multi-view static aspect, our model proposes in addition, temporal extension.

In Section 5.2 the conditions, which allow for the temporal Super-Resolution of the appearance, are discussed and the image formation model is provided. The probabilistic approach of solving the super-resolved appearance is then presented in Section 5.3 as a continuation of the proposed method in the static case. In Section 5.4, experimental validation demonstrates the additional quality, that is regained through the temporal case and the contributions as well as the shortcomings brought forward, are discussed in 5.5.

5.2 Super-Resolution in Multi-View Scenario Through Temporal Accumulation

The visual details regained from the Super-Resolution appearance model of Chapter 4, motivate the exploitation of more observations in order to boost the quality of the appearance. Towards that direction and since several video sequences are available we propose a model to take full advantage of the temporal accumulation of all views. Inspiration is drawn from the vastly explored area of video Super Resolution and in particular based on the work of Liu and Sun [2011] we develop a temporal coherent Super-Resolution appearance model. It successfully generalizes the previous SR model to time domain and achieves higher quality appearance.

The main assumption of the proposed framework is the idea that the appearance of a captured subject does not change dramatically during time and thus captured images at different time instances are considered observations of the same appearance. However, these conditions are not always met, especially when the subjects of interest are of arbitrarily deformable nature, such as human actors, whose movement during a time interval can be large. In these cases, big occlusions, changes in light are often apparent and the appearance undergoes large variations. In order for the assumption of common appearance to hold, a specific time interval needs to be accounted, where the subject of interest undergoes small movement variations and hence the appearance remains similar. And it comes without any loss of generality to consider small time windows and to develop the model for such scenario. But even in these cases, a consistent temporal accumulation of appearance data can only be achieved by realigning the relevant parts of the texture from one temporal frame to another, and accounting for sources of geometric variability. Fortunately, recent progress in non-rigid surface tracking methods Cagniart et al. [2010] offer a path to resolve such issues, which we open with this work.

In this context of predefined narrow time intervals, where the subject is modeled by a mesh of fixed connectivity it is feasible to super-resolve the common appearance by exploiting the available visual information shared across time and among viewpoints. The main principle of SR, discussed in 4.2.2, is the existence of information, which differs in sub-pixel precision. This non-redundant information appears even stronger in the temporal case since any method which imposes fixed connectivity can be imperfect and possibly results in misalignments. These additional sub-pixel shifts are exploited by the proposed model to serve an even more detailed appearance.

5.2.1 Image Formation Model

The proposed image formation model which integrates the observations across time is described in Figure 5.1. The same parametric representation of 4.3.2 is used to express each image captured at a specific viewpoint and time instance, as an observation of the common appearance. Following similar analysis to 4.3.2, at each time instance, the common appearance is mapped to the corresponding pose and projected into the high resolution image space through the projection operator. The major difference of that step is the information enclosed in that operator. In the multi-view static case, the geometric noise introduced misalignments in the form of sub-pixel shifts. In the temporal extension, due to inaccurate tracking, stronger misalignments appear and

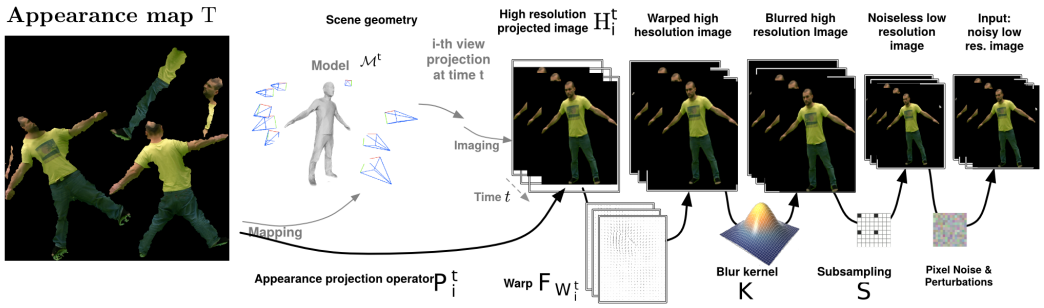

Figure 5.1 – Summary of image formation model and problem notation.

they are also reflected through the same operator. Although these shifts originate from different factors, we postulate and will verify in experiments that they all can be uniformly expressed by warp fields, defined in the high resolution image space. The warped images are then blurred and down-sampled according to the acquisition process, at the end of which the thermal sensor noise is similarly expressed through the additive term.

Let T be the unknown super-resolved appearance which is considered to be common for the time interval $t \in \{1, \cdots, n_t\}$. A temporally coherent mesh model of the object is assumed, *i.e.* whose connectivity is fixed but of varying pose $\{\mathcal{M}^t\}$, obtained by tracking the surface tracking Cagniart et al. [2010]. Let $\{\mathrm{I}_i^t\}$ the set of input images and $\{\mathrm{H}_i^t\}$ be the set of the corresponding high resolution images. Each high resolution image H_i^t is generated from the common appearance T through the projection operator, which encodes the geometric noise as well as the noise due to inaccurate tracking. Let $\hat{\mathbf{P}}_i^t$ be this projection operator. The mapping of a texel at location k of the texture space $\mathbb{T}$ to the noisy geometric point $\hat{\phi}(k)$ is time-independent due to fixed connectivity of the mesh across time and then it is inaccurately projected, due to imperfect camera calibrations, on the image H_i^t at the pixel $\hat{\pi}_i^t \circ \hat{\phi}(k)$. Under the assumption that the appearance remains similar, the value of that pixel is equal to the appearance value, to which the texel is mapped through the appearance function $T \circ \hat{\phi}(k)$. Overall, the high resolution image is generated according to:

$$\mathrm{H}_i^t = \hat{\mathbf{P}}_i^t \mathrm{T}. \tag{5.1}$$

Since this noise is visualized as sub-pixel shifts, it can be further expressed through warp fields over the noiseless high resolution image. Let $\mathbf{P}_i^t$ the noiseless projection operator and let $\mathbf{F}_{\mathrm{W}_i^t}$ the linear warping operator. The high resolution image is written as follows:

$$\mathrm{H}_i^t = \hat{\mathbf{P}}_i^t \mathrm{T} = \mathbf{F}_{\mathrm{W}_i^t} \mathbf{P}_i^t \mathrm{T}. \tag{5.2}$$

To complete the observation model the capturing process needs to be simulated. At this stage, the blur operator **K**, the sub-sampling operator **S** remain the same during the time interval and only the thermal noise with the color filtering noise can be considered time dependent. Let N_i^t be the

additive term to describe this noise in the capturing process. The low resolution image I_i^t at the output of the capturing system in view $\{i, t\}$ is simulated by:

$$\mathrm{I}_i^t = \mathbf{SK}\mathrm{H}_i^t + \mathrm{N}_i^t. \tag{5.3}$$

The full image formation model in the multi-view context, which describes how the low resolution image is generated by the high resolution appearance (texture) is written as follows:

$$\begin{aligned} \mathrm{I}_i^t &= \mathbf{SKF}_{\mathrm{W}_i^t}\mathbf{P}_i^t\mathrm{T} + \mathrm{N}_i^t \\ \mathrm{I}_i^t &= \mathbf{A}_i^t\mathrm{T} + \mathrm{N}_i^t. \end{aligned} \tag{5.4}$$

where, under the assumption that every parameter is written in lexicographic order, we introduce the sparse linear operator $\mathbf{A}_i^t = \mathbf{SKF}_{\mathrm{W}_i^t}\mathbf{P}_i^t$ for each view $\{i, t\}$ has $w_{\mathrm{I}_i^t} \times h_{\mathrm{I}_i^t}$ rows and $w_\mathrm{T} \times h_\mathrm{T}$ columns and the noise term is of size $w_{\mathrm{I}_i^t} \cdot h_{\mathrm{I}_i^t} \times 1$. According to equation 5.4 the set of low resolution images $\{\mathrm{I}_i^t\}$ results from the measurement of the appearance T of a subject with varying pose $\{\mathscr{M}^t\}$ in the time interval $t \in \{1, \cdots, n_t\}$. Solving for this common appearance is thus rephrased as solving for the appearance, which best explains the set of low resolution images.

5.3 Probabilistic Model Extended for Temporal Case

The above inverse problem is expressed in a probabilistic framework to describe the forward process of image generation as a stochastic process. Following similar analysis with the static case, any uncertainty in the image generation pipeline is modeled through the generative probabilistic model of equation 5.4.

In 5.3.1 the graphical model is presented to describe the image generation through random fields. The analytic probability distributions are similar to the ones used in the static case and are inserted in the formalization of the Maximum a Posteriori solution (MAP) of the appearance, which is computed through a two-stage iterative algorithm in 5.3.2.

5.3.1 Graphical Model

In the generative probabilistic model of equation 5.4 the main difference compared to the static case is the insertion of the time factor of the parameters W_i^t and N_i^t and I_i^t. The extension of the model in the time domain is thus translated in considering these parameters as samples of random fields. At each view $\{i, t\}$ a probability distribution is assigned to each corresponding field to describe its statistical behavior. The unknown high resolution texture T is time-independent and it is also considered sample of a random field. The projection operator $\mathbf{P}_i^t$, the sub-sampling operator $\mathbf{S}$ and the Gaussian blur kernel operator $\mathbf{K}$ (PSF) of the image formation model are known and by representing each random variable as a node and expressing their dependency by edges the corresponding graphical model is presented in Figure 5.2. The structure of the graph and the conditional probability distribution specify the joint probability distribution over all the random variables as follows:

$$p(\{\mathrm{I}_i^t\}, \{\mathrm{W}_i^t\}, \mathrm{T}) = p(\{\mathrm{I}_i^t\}|\{\mathrm{W}_i^t\}, \mathrm{T})\, p(\{\mathrm{W}_i^t\}, \mathrm{T}). \tag{5.5}$$

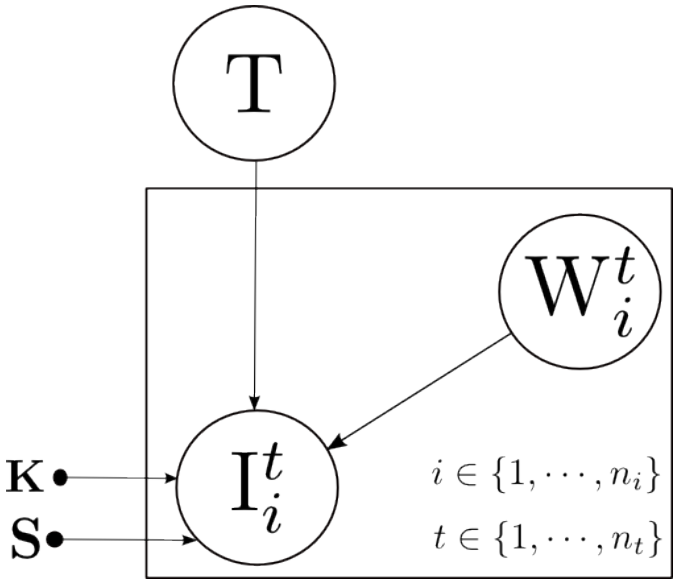

Figure 5.2 – The plate represents $n_i \times n_t$ nodes for the variable of image and $n_i \times n_t$ for the variable of warp field. The temporal extension adds more observations.

The generation of the image at each view $\{i, t\}$ is independent of the one at any other view point and time instance. Inserting the notion of independence as in the static case, extended also in the temporal domain, the graphical model is described by the joint distribution as follows:

$$p(\{\mathrm{I}_i^t\}, \{\mathrm{W}_i^t\}, \mathrm{T}) = \prod_{i,t} p(\mathrm{I}_i^t | \mathrm{W}_i^t, \mathrm{T}) \prod_{i,t} p(\mathrm{W}_i^t) p(\mathrm{T}). \tag{5.6}$$

5.3.2 Maximum A Posteriori Solution Through Bayesian Framework

To solve for the unknown high resolution appearance, Bayesian inference is applied on the graphical model of the temporal case. The corresponding study in the static case revealed the ill-posed behavior of the inverse problem and the need of Maximum A Posteriori Solution. At this point, only the maximum a posteriori solution is provided as a point estimator of the appearance. In particular, the Maximum A Posteriori estimators are written as follows:

$$\begin{aligned} \{\mathrm{T}^*_{\mathrm{MAP}}, \{\mathrm{W}^{t*}_{\mathrm{iMAP}}\}\} &= \arg\max_{\mathrm{T}, \{\mathrm{W}_i^t\}} p(\mathrm{T}, \{\mathrm{W}_i^t\} | \{\mathrm{I}_i^t\}) \\ &= \arg\max_{\mathrm{T}, \{\mathrm{W}_i^t\}} \frac{p(\{\mathrm{I}_i^t\} | \{\mathrm{W}_i^t\}, \mathrm{T}) p(\{\mathrm{W}_i^t\}, \mathrm{T})}{p(\{\mathrm{I}_i^t\})} \\ &= \arg\max_{T, \{\mathrm{W}_i^t\}} \prod_{i,t} p(\mathrm{I}_i^t | \mathrm{T}, \mathrm{W}_i^t) P(\mathrm{T}) \prod_{i,t} p(\mathrm{W}_i^t). \end{aligned} \tag{5.7}$$

The probability distributions of the random variables are analytically formulated similarly to the static case by adding the notion of time as additional component. Thus the analytic representation

of the MAP solution is written as follows:

$$\{\mathrm{T}^*_{\mathrm{MAP}}, \{\mathrm{W}^{\mathrm{t}^*}_{\mathrm{i\,MAP}}\}\} = \arg\min_{\mathrm{T},\{\mathrm{W}^t_i\}} \sum_{i,t} (\mathrm{I}^t_i - \mathbf{A}^t_i \mathrm{T})^\top \mathrm{D}_i (\mathrm{I}^t_i - \mathbf{A}^t_i \mathrm{T}) + \theta TV_I(\mathrm{T}) + \sum_{i,t} \nu \left(\|\nabla u^t_i\| + \|\nabla v^t_i\| \right). \tag{5.8}$$

According to the two-step iterative algorithm, introduced in the static case, the warp field is firstly estimated through the following:

$$\mathrm{W}^{\mathrm{t}^*}_{\mathrm{i\,MAP}} = \arg\min_{\mathrm{W}^t_i} \nu \left(\|\nabla u^t_i\| + \|\nabla v^t_i\| \right) + (\mathrm{I}^t_i - \mathbf{SKF}_{\mathrm{W}^t_i} \mathbf{P}^t_i \mathrm{T})^\top \mathrm{D}_i (\mathrm{I}^t_i - \mathbf{SKF}_{\mathrm{W}^t_i} \mathbf{P}^t_i \mathrm{T}).$$

Given the corresponding warping operator $\mathbf{F}_{\mathrm{W}^{t^*}_{\mathrm{i\,MAP}}}$ the MAP solution of the appearance in computed through the following:

$$\mathrm{T}^*_{\mathrm{MAP}} = \arg\min_{\mathrm{T}} \sum_{i,t} (\mathrm{I}^t_i - \mathbf{A}^t_i \mathrm{T})^\top \mathrm{D}_i (\mathrm{I}^t_i - \mathbf{A}^t_i \mathrm{T}) + \lambda \|\nabla \mathrm{T}\|.$$

The stopping criteria for this alternating minimization algorithm are the maximum number of iterations and the convergence of the relative norm of the objective function to a predefined threshold.

5.4 Experimental Evaluation

To evaluate our approach on the temporal aspect, we use four synchronized datasets, TOMAS and three new GOALKEEPER, BACKPACK, ACTOR. The datasets were acquired with three different setups and camera models so as to maximize testing variability. All 3D models were obtained using silhouette-based reconstruction techniques and thus yield largely imperfect models. GOALKEEPER consists of 21 calibrated viewpoints at 1024×1024, which we down-sample to 512×512 for the purpose of evaluation. BACKPACK consists of 15 viewpoints of a person, with resolution 1624×1224. ACTOR consists of 11 viewpoints in resolution 1920×1080.

5.4.1 Quantitative Evaluation

There are several difficulties in designing an experiment to quantify this improvement, such as the absence of ground truth data in texture space for real datasets. Synthetic datasets are less than ideal for image restoration and super-resolution problems: a significant conclusion can only be achieved if the different sources of variability are correctly introduced and simulated: sensor noise, calibration error, local reconstruction errors, specularities, temporal misalignments. Instead, we focus here on showing the temporal improvement by running our algorithm on a $2\times$ downsampled version of the GOALKEEPER dataset, and comparing our reprojected result with the higher resolution inputs using the mean squared error metric (MSE). Figure 5.3 shows the result of this experiment, with convergence curves from one frame (static case) to three frames, and MSE's evaluated on the 21 input views. Several observations can be made from these curves. First, they illustrate convergence of the iterations toward the high resolution ground truth. Second the temporal improvement leveraged by our algorithm is validated in two forms: acceleration of

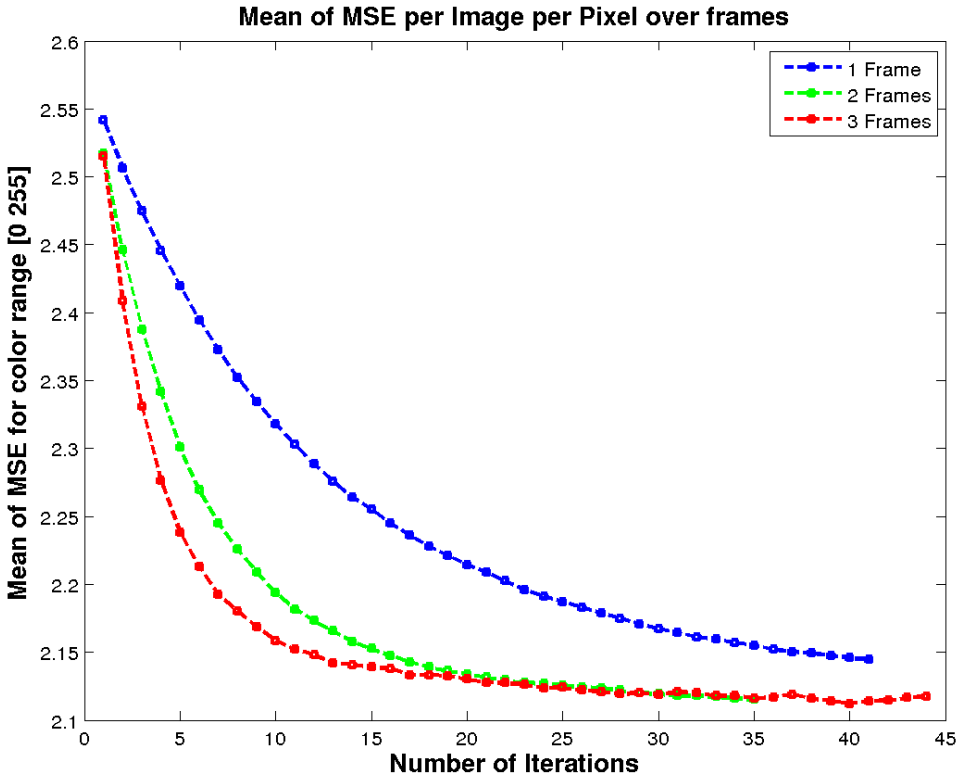

Figure 5.3 – Results on GOALKEEPER dataset. We computed the mean value over frames of the Mean Square Error between our output and high resolution ground truth image. The resolution of input images is 512×512 and of the super-resolved output images is 1024×1024. We use a time step of 2 in experiments, corresponding to an acquisition frequency of 15Hz.

the rate of convergence using more temporal frames, and improvement of the final result quality over using only one temporal frame.

5.4.2 Qualitative Evaluation

In Figure 5.4 rendered images with the reconstructed appearance are presented to demonstrate the quality improvement of temporal extension compared to static case. The ACTOR dataset is arguably the most difficult one, with lower views and higher noise levels both in the images and the reconstruction. We focus on small motions of the three subjects, and test the method for 2 to 7 frames. Significant improvements can be seen in the figure through temporal accumulation. Blurring is successfully removed and high frequency details are correctly aligned and retrieved.

5.4.3 Temporal Evolution of Appearance

The appearance is temporally coherent only in small time intervals, where changes in motion and lighting conditions are not large. The proposed temporal SR appearance estimation makes use of this assumption and additional details are recovered.

We here validate this assumption and demonstrate that the appearance evolves over time. SR appearance estimation is applied at different time instances on TOMAS dataset. Close ups of the initialization and the computed super-resolved textures demonstrate in Figure 5.5 the similar nature of the appearance variability across sequential time steps (e.g. folds of T-Shirt and pens). Observing the temporal evolution of the computed appearance, two cases can be identified. There is appearance information that is temporally consistent (e.g. the colors of the T-Shirt and the pens, the logo of the T-Shirt) while there is information that varies (e.g. folds are changing). In this spatio-temporal dataset, low frequency can thus be interpreted as the temporally consistent visual information while high frequency as the non consistent one. These observations are of great interest, since they provoke further study of novel more compressive appearance representations to encode the underlying variability.

5.5 Discussion

The experimental evaluation of the proposed temporal coherent Super-Resolution appearance model shows that the temporal accumulation can indeed increase the quality of the appearance. The observations through time provide in the appearance space additional non-redundant information, which can be exploited. The length of the time intervals, where the proposed model is applied, is required to be small in order to capture small variations of motion and appearance. This restriction however comes from the inherent limitation of the appearance modelling problem. Considering time intervals, where the changes in the motion are large, the appearance can be extremely different and the assumption of temporal coherency is not anymore valid. When images are rendered at a time instance not included in the interval, where the appearance is super-resolved, artifacts are created and the visual result is not realistic. This observation is an indication that the appearance is always subject to a given geometry, which remains constant in a predefined time interval and is considered to be accurate enough. Despite the regain of visual details due to the extension of the model in the time domain, there is still a particular nature of visual artifacts related to the geometry. In the proposed model, any inaccuracy in the geometric reconstruction is compensated at a certain degree by the warp fields per viewpoint and per time instant. This local way of handling the geometry however could limit the visual quality, since the problem of appearance modelling could benefit in a more global way by the common geometric information shared among the views and across time. Elaborating more on this aspect of geometric modelling is the research direction of the next chapter.

5.6 Conclusion

In this chapter the direction towards the time domain has been explored and the temporal aspect has been successfully integrated in the proposed framework. Under the assumption that changes in the motion of the captured subject are not large in a predefined time interval, a temporally coherent Super-Resolution appearance model is defined to exploit the non redundant information available in that time interval.

It is a novel representation, which efficiently leverages the principle of SR not only in the multi-view but also in the temporal context. Additional details are retrieved and higher visual

quality appearance is achieved. The major contribution of the proposed model is the property of time independence up to a specified time interval. The appearance representation is view-independent and time-independent under the constraint that the time instance belongs to the predefined time interval.

The limitation of the model to a predefined time interval, where motion changes are small, reveals the correlation of the geometry to the appearance. The appearance is subject to the geometry of the model and thus the visual quality can benefit from it when the geometric aspect is integrates in the model in a more coherent way. Exploring how geometry and appearance, as two separate modalities, interfere each other is the main subject of the next chapter.

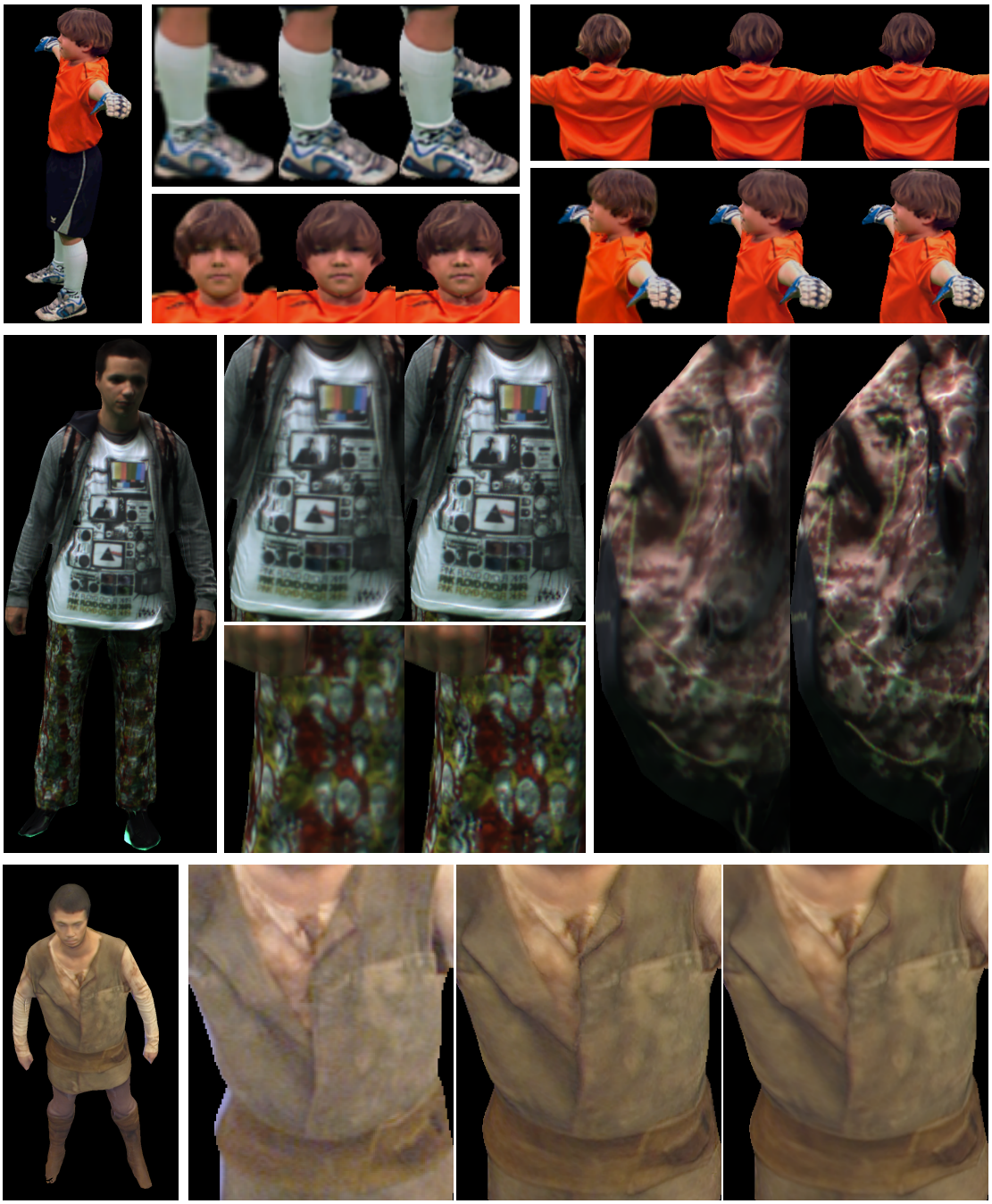

Figure 5.4 – This figure illustrates various temporal improvements and detail enhancements obtained with various acquired datasets, comparing different convergence using one or several temporal frames. Top: GOALKEEPER dataset. Left: output of Frame 3. Input is compared to Frame 1 and Frame 3 for each close-up. Middle: BACKPACK dataset. Input on left, Frame 1 and Frame 2 comparisons for close-ups. Details are revived on the backpack, T-shirt and pants. Bottom: ACTOR; left to right: full result with three frames, close-up comparison between input, against Frame 1 and Frame 3. Best viewed magnified and in color.

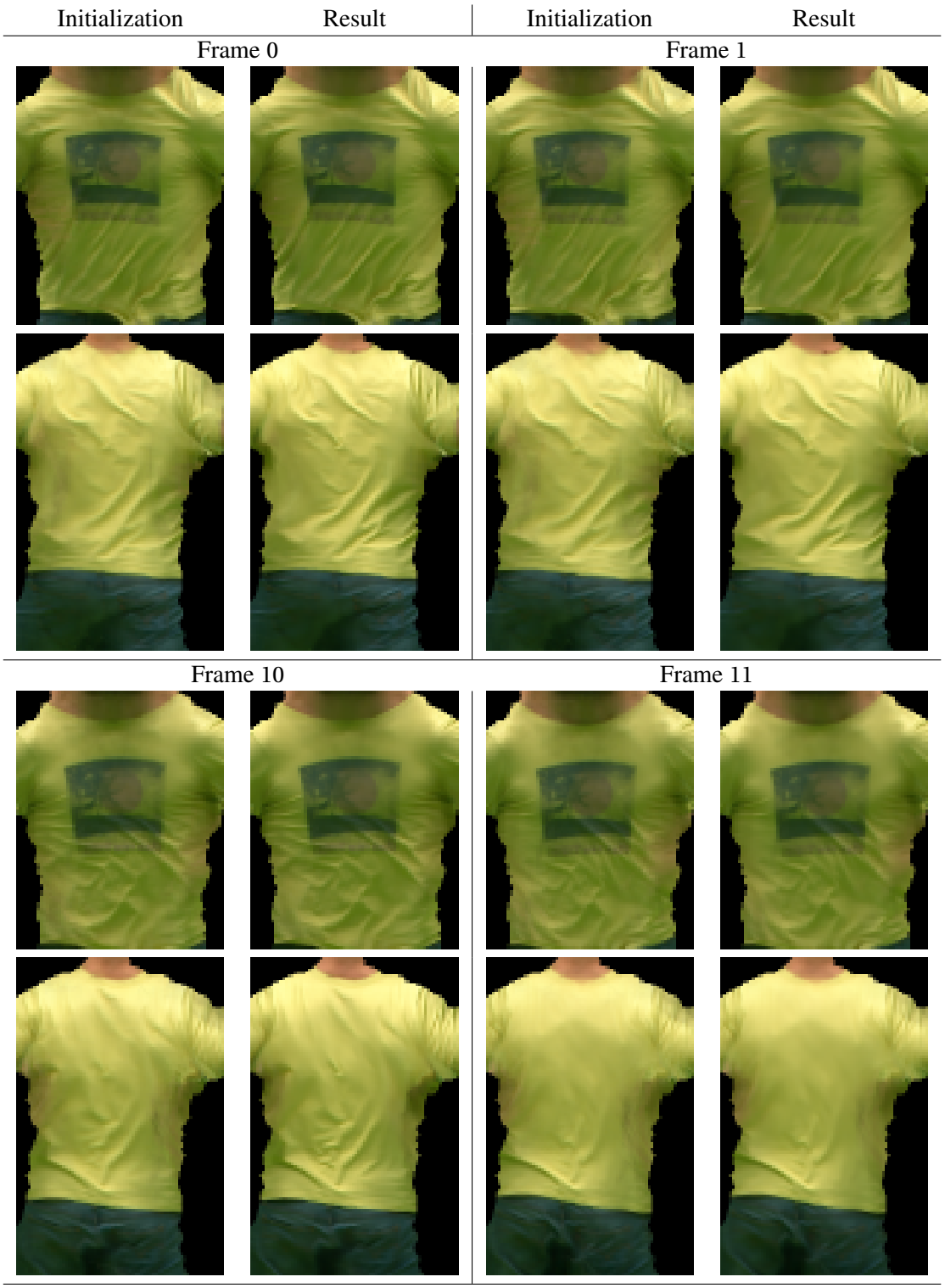

Figure 5.5 – Temporal evolution of appearance computed on TOMAS dataset. The resolution of input images is 2048×2048 and of the super-resolved output images is 4096×4096. Low appearance variability across sequential time steps.

6 Geometry and Appearance through a Super-Resolution Framework

6.1 Introduction and Motivations

The static and the temporal Super-Resolution appearance models succeed to regain visual details by exploiting the non-redundant appearance information, shared among the views and fused in time respectively. In both models however geometry is introduced as a known, constant variable and information is not explicitly exploited to enhance the appearance estimation. Expressing it as a separate modality should act as additional source of information provoking further research.

In this chapter, we propose a Super-Resolution framework, where we accommodate jointly the geometric and the appearance modelling and we demonstrate that an improvement of the quality of the appearance can be achieved. The motivation behind this direction is two-fold. First, appearance and geometry, as two separate modalities which both explain the image observations are conditionally dependent and the process of estimating them should intuitively prove mutually beneficial. Second, through the previous study and our proposed appearance model, we have shown that geometric inaccuracies are successfully compensated locally per view and frame. Accounting for them in a more global way by expressing them in the common space where geometry is defined, comes as a natural extension. Although this idea of exploiting the relationship between geometry and appearance has been introduced, few works exist, which apply Super-Resolution techniques in the multi-view context. As opposed to previous work Goldluecke and Cremers [2009a] which optimize for a displacement map defined on the normals of the surface our proposed algorithm alleviates this limitation and allows a more general geometric refinement which will best explain the image observations. Compared to the more recent work Goldluecke et al. [2013] which additionally solves for camera calibration parameters, we account for any calibration noise uniformly through a linear representation of the image generation process.

In Section 6.2 we develop the basic ideas which motivated us to define this research direction of modelling geometry and appearance through a Super-Resolution framework for the multi-view scenario and we then present the linear image formation model which materializes SR and accommodates geometry and appearance. In analogy to our previous strategy, we cast the problem into a probabilistic framework in Section 6.3. and we apply Bayesian inference. In Sections 6.4 and 6.5 we present the two step iterative algorithm to optimize for appearance and geometry accordingly. In Section 6.7 limitations, future directions are discussed and main findings are summarized.

6.2 Super-Resolution in Multi-View Scenario Accounting For Geometry and Appearance

Through our previous study we demonstrated that super-resolving the appearance of subjects is possible due to the redundancy with which it is observed across temporal frames and different viewpoints. We here extend the scenario and we introduce the geometry as a separate modality which acts as additional source of information. We propose a framework which unifies the problem of super-resolving appearance and the problem of geometric modelling and in particular we develop the image formation model in Section 6.2.1 to materialize the above concepts.

Arguably, appearance and geometry, as two separate modalities are inherently correlated and the process of estimating them should prove mutually beneficial. Intuitively we could see this interdependency of photometry and geometry through a naive example of a face, where the geometry could be either coarsely modeled as a sphere, or modeled in great detail with fine restitution of geometric features. Geometric characteristics (e.g nose, lips, eyes) are expressed in the form of visual information in the estimated appearance of both cases. The more precise the geometry is the more on focus these visual details are, which sustains the statement that geometric accuracy drives to appearance accuracy. Looking at the opposite direction, we could also claim that appearance can lead to geometric enhancement. We thus introduce the notion of bidirectional relation of geometry and appearance and we further elaborate on it. This aspect has been also studied by previous researchers, who have aspired to exploit both of the modalities in order to increase the accuracy of their reconstruction, e.g (Aubry et al. [2011], Delaunoy and Pollefeys [2014], Pujades et al. [2014]), a fact that also manifests in the emergence of works based on data fusion, e.g (Morris et al. [2001], Cabezas et al. [2014]).

As an additional support towards the exploitation of this bidirectional relation, comes the efficiency of our proposed appearance model to account for any geometric noise. The provided appearance model succeeds to compensate for the geometric noise through the two dimensional warp fields defined on image space. These motion fields were locally defined at each view and time frame and the idea to express them in a more global framework comes as a natural continuation. Revisiting the causes of the geometric noise, discussed in Chapter 4.3, reveals that these two dimensional fields comprise information about the reconstruction and calibration. We believe that it is possible to extract the information related to the geometry and to directly express it on it.

To that goal we introduce the geometry as an additional variable and we extend our previous image formation model accordingly to integrate it. We introduce a framework to unify the reconstruction of appearance and geometry. Given the updated image observation model, appearance and geometry are both optimized to explain the image observations. Interestingly, previous works (Goldluecke and Cremers [2009a], Goldluecke et al. [2013]) adopt also this research direction. As opposed to the first one, which optimizes for a displacement map defined on the normals of the surface, our proposed algorithm alleviates this limitation and allows a more general geometric refinement which best explains the image observations. Compared to their more recent work (Goldluecke et al. [2013]) which on top of the geometric and texture optimization, optimizes for camera calibration parameters, we follow a different strategy, where we uniformly account for

both calibration and reconstruction and then extract the component related to reconstruction only, in order to enhance the geometry.

6.2.1 Image Formation Model

In the same spirit of our previous development, we introduce the parametric representation of the image observation model to explain how the low resolution observations are generated by the unknown high-resolution appearance, to describe the levels of degradation in the capturing process and to accommodate the geometry.

The texture is projected into the high resolution image through the projection operator. This operator expresses the dependency on the geometry and on the camera calibration parameters but we introduce only the geometry as additional variable since we assume that camera parameters are constant and we only model geometry. However, inaccuracies of both reconstruction and calibration can appear and they jointly lead to sub-pixel shifts in the image space. We model these shifts with two dimensional warps over the projected high resolution images. Then the warped high resolution images are convolved and down-sampled to simulate the capturing process. Thermal noise and color filtering noise are expressed through an additive term.

Let G be the geometric realization of the model $\mathscr{M}$, which is defined through a spatial localization of the vertices $\mathrm{G} := \{\mathrm{V}_l, l = 1, \ldots, \mathscr{V}\} \subset \mathbb{R}^3$ where $\mathscr{V}$ the number of vertices. Let $\hat{\mathrm{G}}$ be the geometric approximation due to reconstruction noise and let $\hat{\mathbf{P}}^i_{\hat{\mathrm{G}}}$ be the projection operator at viewpoint i, which expresses the noise due to reconstruction and calibration inaccuracies. The generated high resolution image is thus expressed as:

$$\mathrm{H}_i = \hat{\mathbf{P}}^i_{\hat{\mathrm{G}}} \mathrm{T}. \tag{6.1}$$

This noise takes the form of sub-pixel shifts and it is simulated further by warp fields over the noiseless high resolution images. Let $\mathbf{P}^i_{\mathrm{G}}$ be the noiseless projection operator, expressing the ideal case, without any reconstruction and calibration noise.The high resolution image is written as follows:

$$\mathrm{H}_i = \hat{\mathbf{P}}^i_{\hat{\mathrm{G}}} \mathrm{T} = \mathbf{F}_{\mathrm{W}_i} \mathbf{P}^i_{\mathrm{G}} \mathrm{T}. \tag{6.2}$$

The low resolution image I_i at viewpoint i is generated by convolving and down-sampling the high resolution image and then adding the additive noise term to express the thermal sensor and color filtering noise. The full image formation model is written as follows:

$$\begin{aligned} \mathrm{I}_i &= \mathbf{SKF}_{\mathrm{W}_i} \mathbf{P}^i_{\mathrm{G}} \mathrm{T} + \mathrm{N}_i \\ \mathrm{I}_i &= \mathbf{A}^i_{\mathrm{G}} \mathrm{T} + \mathrm{N}_i, \end{aligned} \tag{6.3}$$

where under the same lexicographic assumption introduced already, the sparse operator $\mathbf{A}^i_{\mathrm{G}} = \mathbf{SKF}_{\mathrm{W}_i} \mathbf{P}^i_{\mathrm{G}}$ for each view $\{i\}$ has $w_{\mathrm{I}_i} \times h_{\mathrm{I}_i}$ rows and $w_{\mathrm{T}} \times h_{\mathrm{T}}$ columns and the noise term is of size $w_{\mathrm{I}_i} \cdot h_{\mathrm{I}_i} \times 1$. Solving for both appearance and geometry is rephrased as solving for the appearance and geometry which both explain the set of low resolution images.

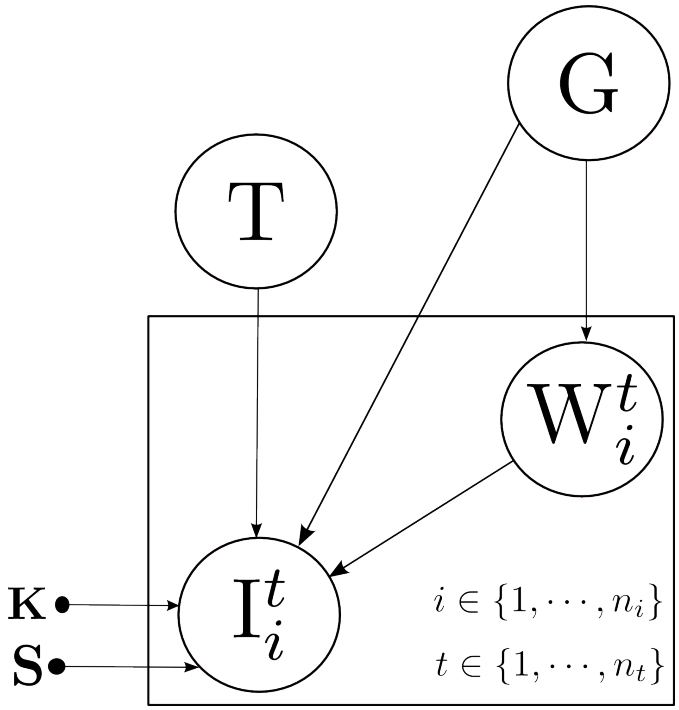

Figure 6.1 – Geometry is integrated in the problem as an additional node connected to the nodes of appearance and warp field. Appearance is subject to the geometry and probabilistically depends on it. Warp fields compensate geometric noise and thus they depend on the geometry.

6.3 Probabilistic Model Accounting For Geometry

Casting the image generation process into a probabilistic framework, all the parameters are treated as stochastic variables to express every uncertainty. The corresponding graphical model is presented in 6.3.1 and in 6.3.2 prior knowledge of the additional unknown geometry is expressed through the analytic representation of the probability distribution.

6.3.1 Graphical Model

In the image generation process, the high resolution texture T, the warp field W_i, the observed low resolution images I_i and the geometry G are considered to be samples of random fields. The dependency among them is expressed in Figure 6.1 and it is quantitatively defined by conditional probability distributions. The joint probability distribution over all the random variables is written as follows:

$$p(\{I_i\}, T, \{W_i\}, G) = p(\{I_i\}|T, \{W_i\}, G)\, p(T, \{W_i\}, G). \tag{6.4}$$

The conditional probability $p(\{I_i\}|T, \{W_i\}, G)$ expresses how probable is for the set of low resolution images $\{I_i\}$ to be generated by the known texture T, the known geometry G and the set of warping operators $\{W_i\}$. The structure of the graph determines further the factorization that can take place. The set of warping fields $\{W_i\}$ is introduced to compensate for the geometric

inaccuracies and thus given the geometry G it is independent from the appearance T. The joint probability distribution $p(\mathrm{T}, \{\mathrm{W}_i\}, \mathrm{G})$ is written as follows:

$$p(\mathrm{T}, \{\mathrm{W}_i\}, \mathrm{G}) = p(\mathrm{T})\, p(\{\mathrm{W}_i\}|\mathrm{G})\, p(\mathrm{G}). \tag{6.5}$$

Inserting the notion of view independence in the image generation process the graphical model is described by the joint probability distribution:

$$p(\{\mathrm{I}_i\}, \mathrm{T}, \{\mathrm{W}_i\}, \mathrm{G}) = \prod_i p(\mathrm{I}_i|\mathrm{T}, \mathrm{W}_i, \mathrm{G})\, p(\mathrm{T}) \prod_i p(\mathrm{W}_i|\mathrm{G})\, p(\mathrm{G}). \tag{6.6}$$

6.3.2 Geometric Prior Information

The statistical dependency of the parameters as well as their statistical behavior is quantitatively defined through the analytic representation of the corresponding distributions. The choice of these representations is made on the criterion of reflecting in the most realistic way the nature of the image generation process. Given the current graphical model the observation model expresses how the low resolution image is generated given the appearance, the warp fields and the geometry and corresponds to the distribution of the additive noise term. Under the same assumption of the zero mean independent white Gaussian noise, and its statistical independence among the views the conditional probability of the set of the low resolution images $\{\mathrm{I}_i\}$ given the high resolution texture T, the set of the warping matrices $\{\mathrm{W}_i\}$ and the geometry G is expressed as:

$$\begin{aligned} p(\{\mathrm{I}_i\}|\{\mathrm{W}_i\}, \mathrm{T}, \mathrm{G}) &= \prod_{i,t} p(\mathrm{I}_i|\mathrm{W}_i, \mathrm{T}, \mathrm{G}) \\ &= \prod_{i,t} \frac{1}{Z(\mathrm{D}_i)} e^{\sum_{i,t} -(\mathrm{I}_i - \mathbf{A}_{\mathrm{G}}^i \mathrm{T})^\top \mathrm{D}_i (\mathrm{I}_i - \mathbf{A}_{\mathrm{G}}^i \mathrm{T})}. \end{aligned} \tag{6.7}$$

The texture model expresses that the appearance is a natural image and the probability is written as:

$$p(\mathrm{T}) = \frac{1}{Z_{\mathrm{T}}(\lambda)} e^{-\lambda \|\nabla \mathrm{T}\|}. \tag{6.8}$$

Assuming that the geometry is known, dense optical flow is used as the motion field to compensate for the geometric noise. As it was in 4.4.2.3 explained, that noise exhibits a sparse behavior and the occasional transitions of the flow field are introduced through the TV norm as the potential function in the probability distribution. Given the geometry and under the assumption of statistical independence among viewpoints, the conditional probability of the set of the warp fields $\{\mathrm{W}_i\}$ is expressed as:

$$\begin{aligned} p(\{\mathrm{W}_i\}|\mathrm{G}) &= \prod_{i,t} p(\mathrm{W}_i|\mathrm{G}) \\ &= \prod_{i,t} \frac{1}{Z_{\mathrm{W}_i}(\gamma)} e^{-\nu\left(\|\nabla u_i^t\| + \|\nabla v_i^t\|\right)}. \end{aligned} \tag{6.9}$$

The geometry as an additional variable of the graphical model is also considered sample of

a random field to which probability distribution is assigned at the reconstruction stage. This probability distribution writes as $p(\mathrm{G}) = \frac{1}{Z_{\mathrm{G}}(\gamma)} e^{-\gamma\Phi(\mathrm{G})}$, where $\Phi(\mathrm{G})$ encodes information about the spatial localization of vertices's and thus it describes the general structure of the object. In the current thesis, two main classes of objects are studied. The first category includes objects with smooth surfaces and quadratic model is used to describe the probability distribution, which the geometry follows. The second category corresponds the case, where objects consist of smooth patches interrupted by strong edges, properties that are expressed through more sparse priors.

6.3.2.1 Shapes with Smooth Surfaces

The smoothness property of the geometry can be described by introducing a Gaussian distribution of the second order derivative operator applied on the geometry. This quadratic form leads to over smoothed geometry and results in an overall shrinkage. To maintain the smoothing property and to eliminate the shrinkage effect, second order of laplacian is used instead and the prior distribution of the geometry is written as follows:

$$p(\mathrm{G}) = \frac{1}{Z_{\mathrm{G}}(\gamma)} e^{-\gamma\|\mathbf{L}^2\mathrm{G}\|_2^2}, \tag{6.10}$$

where $\mathbf{L} = I_d - D^{1/2}\tilde{W}D^{-1/2}$ is the normalized laplacian matrix, $\tilde{W}$ the local normalized averaging with conformal weights, $D = \mathrm{diag}(d_i)$ the diagonal matrix with $d_i = \sum w_{ij}$ and $Z_{\mathrm{G}}(\gamma)$ the normalization factor.

6.3.2.2 Shapes with Discontinuities

For the other category of objects which consist of smooth patches, interspersed by occasional strong edges, the geometry does not follow the probability distribution described before. In this case, the spatial localization of the vertices of the mesh introduces a sparsity of large derivatives, which is described by non-quadratic terms in the potential function. In particular, the L1 norm of the gradient operator being applied on the geometry promotes the expected discontinuities and thus the probability distribution of the geometry of such objects is defined as:

$$p(\mathrm{G}) = \frac{1}{Z_{\mathrm{G}}(\gamma)} e^{-\gamma\|\mathbf{G}\mathrm{G}\|_1}, \tag{6.11}$$

where $\mathbf{G}$ is the (m, n) gradient matrix with m the number of edges, n the number of vertex and $\|x\|_1 = \sum_l |x(l)|$.

6.3.3 Bayesian Framework

The image formation model described by the graphical model depends on the appearance, geometry and two dimensional warp fields. To problem is thus reformulated as an optimization with respect to all these parameters. We seek appearance, motion field and geometry, which best explain the observed images and at the same time preserve their own characteristic nature

imposed at reconstruction stage. The Maximum a Posterior estimators are written as follows:

$$\begin{aligned}\{\mathrm{T}^*_{\mathrm{MAP}}, \{\mathrm{W}^*_{\mathrm{iMAP}}\}, \mathrm{G}^*_{\mathrm{MAP}}\} &= \arg\max_{\mathrm{T},\{\mathrm{W}_i\},\mathrm{G}} p(\mathrm{T}, \{\mathrm{W}_i\}, \mathrm{G}|\{\mathrm{I}_i\}) \\ &= \arg\max_{\mathrm{T},\{\mathrm{W}_i\},\mathrm{G}} \frac{p(\{\mathrm{I}_i\}|\mathrm{T}, \{\mathrm{W}_i\}, \mathrm{G})p(\mathrm{T}, \{\mathrm{W}_i\}, \mathrm{G})}{p(\{\mathrm{I}_i\})} \\ &= \arg\max_{T,\{\mathrm{W}_i\},\mathrm{G}} \prod_i p(\mathrm{I}_i|\mathrm{T}, \mathrm{W}_i, \mathrm{G})P(\mathrm{T}|\mathrm{G}) \prod_i p(\mathrm{W}_i|, \mathrm{G})p(\mathrm{G}).\end{aligned} \tag{6.12}$$

Inserting the analytic formulations of the distributions, the equation 6.12 is equivalent to:

$$\{\mathrm{T}^*_{\mathrm{MAP}}, \{\mathrm{W}^*_{\mathrm{iMAP}}\}, \mathrm{G}^*_{\mathrm{MAP}}\} =$$

$$\arg\min_{\mathrm{T},\{\mathrm{W}_i\},\mathrm{G}} \sum_i (\mathrm{I}_i - \mathbf{A}^i_{\mathrm{G}}\mathrm{T})^\top \mathrm{D}_i(\mathrm{I}_i - \mathbf{A}^i_{\mathrm{G}}\mathrm{T}) + \lambda\|\nabla\mathrm{T}\| + \sum_i \nu(\|\nabla u_i\| + \|\nabla v_i\|) + \gamma\Phi(\mathrm{G}). \tag{6.13}$$

Optimizing simultaneously for appearance T, motion fields $\{\mathrm{W}_i\}$ and geometry G is intrinsically hard. We thus use the motion fields $\{\mathrm{W}_i\}$ as a variable to decouple the optimization of the appearance from the optimization of the geometry and a coordinate-descent optimization procedure is followed. We propose a two-stage iterative algorithm and in the first step we solve for the appearance and the motion fields which best explain the low resolution images. We then solve for the geometry which explains the motion fields. These two steps can be iterated to improve both appearance and geometry.

6.4 Optimizing for Appearance

We consider that geometry is known and we solve for the appearance and the two dimensional motion fields which generate according to the image formation model images close to the observed ones. This optimization step resembles the one develop in the Super-Resolution framework of static case 4.4.3.1. Maximum a Posteriori estimators are provided for both unknowns through a two stage iterative algorithm, whose development due to the similarity of the one presented in 4.4.3, will be omitted in this Section.

Let the estimated geometry at iteration k be G_k. The sparse measurement operator of equation 6.3 writes $\mathbf{A}_{\mathrm{G}_k} = \mathbf{SKF}_{\mathrm{W}_{ik}}\mathbf{P}_{\mathrm{G}_k}$ and the image formation model at viewpoint i becomes

$$\begin{aligned}\mathrm{I}_i &= \mathbf{SKF}_{\mathrm{W}_{ik}}\mathbf{P}^i_{\mathrm{G}_k}\mathrm{T} + \mathrm{N}_i \\ \mathrm{I}_i &= \mathbf{A}^i_{\mathrm{G}_k}\mathrm{T} + \mathrm{N}_i.\end{aligned}$$

Let T^*_k, W^*_{ik} be the super-resolved appearance and motion field at this iteration, which are

computed through the minimization of the following objective function:

$$\{\mathrm{T}_k^*, \{\mathrm{W}_{ik}^*\}\} = \arg\min_{\mathrm{T},\{\mathrm{W}_{ik}\}} \sum_i (\mathrm{I}_i - \mathbf{A}^i_{\mathrm{G}_k}\mathrm{T})^\top \mathrm{D}_i(\mathrm{I}_i - \mathbf{A}^i_{\mathrm{G}_k}\mathrm{T}) + \lambda\|\nabla\mathrm{T}\| + \sum_i \nu(\|\nabla u_i\| + \|\nabla v_i\|). \quad (6.14)$$

The MAP estimators are written as follows:

$$\begin{aligned}\mathrm{W}_{ik}^* = {} & \arg\min_{\mathrm{W}_i} \nu(\|\nabla u_i\| + \|\nabla v_i\|) \\ & + (\mathrm{I}_i - \mathbf{SKF}_{\mathrm{W}_{ik}}\mathbf{P}^i_{\mathrm{G}_k}\mathrm{T})^\top \mathrm{D}_i(\mathrm{I}_i - \mathbf{SKF}_{\mathrm{W}_{ik}}\mathbf{P}^i_{\mathrm{G}_k}\mathrm{T}),\end{aligned} \quad (6.15)$$

$$\mathrm{T}_k^* = \arg\min_{\mathrm{T}} \sum_i (\mathrm{I}_i - \mathbf{A}^i_{\mathrm{G}_k}\mathrm{T})^\top \mathrm{D}_i(\mathrm{I}_i - \mathbf{A}^i_{\mathrm{G}_k}\mathrm{T}) + \lambda\|\nabla\mathrm{T}\|. \quad (6.16)$$

6.5 Optimizing for Geometry

Given the super-resolved appearance we then optimize for the geometry. The central idea is to first explain the photometric error by means of a 2D motion field, which uniformly accounts for the misalignments in the image space and then to account for this error by optimizing for the geometry. Given the texture and the current estimated geometry we solve for the 2D motion field which when applied on the projected high resolution image generates a warped image, which in turn when convolved and down-sampled best explains the low resolution image observation. This motion field can be interpreted as the projection of the geometric noise due to calibration and reconstruction inaccuracies. Thus, given the signature of this noise, we extract the information related only to reconstruction by regularizing the solution. These two steps of optimizing motion fields and geometry are iterated and after a certain number, a predefined accuracy is achieved. Maximizing the posterior probability and dropping all the terms independent of the geometry and motion field, the point estimators are written as follows:

$$\{\mathrm{G}_k^*, \{\mathrm{W}_{ik}^*\}\} =$$

$$\arg\min_{\mathrm{G}_k,\{\mathrm{W}_{ik}\}} \sum_i (\mathrm{I}_i - \mathbf{SKF}_{\mathrm{W}_{ik}}\mathbf{P}^i_{\mathrm{G}_k}\mathrm{T}_k^*)^\top \mathrm{D}_i(\mathrm{I}_i - \mathbf{SKF}_{\mathrm{W}_{ik}}\mathbf{P}^i_{\mathrm{G}_k}\mathrm{T}_k^*) + \sum_i \nu(\|\nabla u_i\| + \|\nabla v_i\|) + \gamma\Phi(\mathrm{G}_k). \quad (6.17)$$

6.5.1 Warp Estimates

We first show how to solve for the warp fields by minimizing the photometric error. Given the estimated super-resolved appearance T_k^* and given the estimation G_k of geometry, each optimal warp field is independently computed through:

$$\mathrm{W}_{ik}^* = \arg\min_{\mathrm{W}_{ik}} \nu(\|\nabla u_i\| + \|\nabla v_i\|) + (\mathrm{I}_i - \mathbf{SKF}_{\mathrm{W}_{ik}}\mathbf{P}^i_{\mathrm{G}_k}\mathrm{T}_k^*)^\top \mathrm{D}_i(\mathrm{I}_i - \mathbf{SKF}_{\mathrm{W}_{ik}}\mathbf{P}^i_{\mathrm{G}_k}\mathrm{T}_k^*). \quad (6.18)$$

Similar strategy to Liu and Sun [2011] is followed to solve this equation and the estimation of W_{ik} is initialized with the results W_{ik}^{*} of the previous appearance optimization step.

6.5.2 Geometry

In the second step, we show how to optimize the geometry accounting for the estimated warp fields. The intuition behind this is that we are solving for a geometry, which when mapped to the super-resolved texture, projected into the high resolution space, warped according to the estimated motion field and then convolved and down-sampled best fits the observation. This step results in a non linear in the geometry objective function and we introduce an approximation to linearize it. The idea is that the projection of the optimal geometry should be close to the projection of the old geometry displaced by the estimated motion field. The solution to the previous optimization is approximated by minimizing an equivalent re-projection error maintaining also the statistical behavior imposed at reconstruction stage.

Let $\mathbf{F}_{\mathrm{W}_{ik}}^{*}$ be the warping operator, which corresponds to the optimal estimated field. The image observation model writes as $\mathrm{I}_i = \mathbf{SKF}_{\mathrm{W}_{ik}}^{*}\mathbf{P}_{\mathrm{G}_k}^{i}\mathrm{T}_k^{*} + \mathrm{N}_i$. The optimal geometry is computed through the minimization:

$$\mathrm{G}_k^{*} = \arg\min_{\mathrm{G}_k} \sum_i (\mathrm{I}_i - \mathbf{SKF}_{\mathrm{W}_{ik}}^{*}\mathbf{P}_{\mathrm{G}_k}^{i}\mathrm{T}_k^{*})^{\top}\mathrm{D}_i(\mathrm{I}_i - \mathbf{SKF}_{\mathrm{W}_{ik}}^{*}\mathbf{P}_{\mathrm{G}_k}^{i}\mathrm{T}_k^{*}) + \gamma\Phi(\mathrm{G}_k). \tag{6.19}$$

Due to projection operator $\mathbf{P}_{\mathrm{G}_k}^{i}$ in the data term, the above objective function is not linear in the geometry and we propose the following linearization step. Let I_{uv}^{ik} be the coordinates of the projected geometry G_k at view i and $\tilde{I}_{uv}^{ik}$ the coordinates updated by the estimated warp fields, $\tilde{I}_{uv}^{ik} = I_{uv}^{ik} + \mathrm{W}_{ik}^{*}$. These can be interpreted as the new observations. The equivalent of equation 6.19 is written as:

$$\mathrm{G}_k^{*} = \arg\min_{\mathrm{G}_k} \sum_i \|\pi_i\mathrm{G}_k - \tilde{I}_{uv}^{ik}\|_2^2 + \gamma\Phi(\mathrm{G}_k). \tag{6.20}$$

Developing further the intuition behind this minimization, the optimal geometry is the one updated by an optimal 3D displacement field. We thus need to solve for a 3D displacement field, which when applied over the current geometry yields a new one whose projection into the image space is close to the coordinates updated by the 2D motion field (new observations).

Let dG_k^{*} be this 3D update in vectorized form. It consists of $3\mathcal{V} \times 1$ elements, where $\mathcal{V}$ the number of vertices and 3 the spatial(x, y, z) coordinates. The optimal geometry is thus given by $\mathrm{G}_k^{*} = \mathrm{G}_k + \mathrm{dG}_k^{*}$ and the minimization problem is rephrased as:

$$\mathrm{dG}_k^{*} = \arg\min_{\mathrm{dG}_k} \sum_i \|\pi_i(\mathrm{G}_k + \mathrm{dG}_k) - (I_{uv}^{ik} + \mathrm{W}_{ik}^{*})\|_2^2 + \gamma\Phi(\mathrm{G}_k + \mathrm{dG}_k). \tag{6.21}$$

Re-factorizing the terms of the above objective function, the minimization problem becomes:

$$\mathrm{dG}_k^{*} = \arg\min_{\mathrm{dG}_k} \|\boldsymbol{A}\cdot\mathrm{dG}_k - \boldsymbol{b}\|_2^2 + \gamma\Phi(\mathrm{G}_k + \mathrm{dG}_k). \tag{6.22}$$

The matrix $\boldsymbol{A}$ is a matrix with $\sum_{i=1}^{N_C} 2\mathrm{Vrt}_i$ rows and $3\mathcal{V}$ columns, where Vrt_i the number of visible

vertices at camera c_i and N_C the number of cameras. An analytic derivation is provided in Appendix A.

The prior statistical behavior of the geometry restricts the space of solutions. This regularization step is necessary, since the process of expressing the optimal warp field into the 3D space in terms of a 3D displacement field adds an additional degree of freedom, resulting in ill-posed system of equations. In order to have the unique 3D displacement vector of an arbitrary vertex, 4 independent equations are needed, which means that this vertex needs to be seen by two cameras that are not in parallel positioned and two two-dimensional motion vectors are available as observations. However due to the configuration of the cameras, it may happen that a vertex is visible from more than two and thus more observations are provided, which yields an overdetermined system. There are also cases, where a vertex is visible by only one camera and thus the 2 equations are not adequate to serve a unique solution. Overall the problem is ill-posed and the prior of the geometry acts as the necessary regularization step.

We now introduce for each case of geometric shape the analytic representation of the prior distribution and we present the corresponding algorithmic approaches.

- **Smooth Surfaces**

  According to 6.10 the potential function is $\Phi(\mathrm{G}_k + \mathrm{dG}_k) = \|\mathbf{L}^2(\mathrm{G}_k + \mathrm{dG}_k)\|_2^2$ and the minimization becomes a least squares problem of the form:

$$\begin{aligned} \mathrm{dG}_k^* &= \arg\min_{\mathrm{dG}_k} \|\boldsymbol{A}\cdot \mathrm{dG}_k - \boldsymbol{b}\|_2^2 + \gamma\|\mathbf{L}^2(\mathrm{G}_k + \mathrm{dG}_k)\|_2^2 \\ &= \arg\min_{\mathrm{dG}_k} \|\boldsymbol{A}^{'}\cdot \mathrm{dG}_k - \boldsymbol{b}^{'}\|_2^2. \end{aligned} \tag{6.23}$$

  The matrix $\boldsymbol{A}^{'}$ is of size $\sum_{i=1}^{N_C} 2\mathrm{Vrt}_i + \mathcal{V}$ rows and $3\mathcal{V}$ columns, the vector $\boldsymbol{b}^{'}$ is $(\sum_{i=1}^{N_C} 2\mathrm{Vrt}_i + \mathcal{V}) \times 1$ and their mathematical formulation is analytically presented in Appendix A.

- **Surfaces with Discontinuities**

  According to equation 6.11 the potential function is $\Phi(\mathrm{G}_k + \mathrm{dG}_k) = \|\mathbf{G}(\mathrm{G}_k + \mathrm{dG}_k)\|_1$ and due to this non quadratic term, the previous derivation does not hold. We introduce an auxiliary variable of 3D displacement field to decouple the regularization term from the data term and the optimal geometry is now computed in two sub-steps. First, the optimal auxiliary displacement field is estimated as the one which when applied on the given estimated geometry results in a new representation, close to the given and which best explains the 2D observed warp field. At the end of this step the auxiliary geometry is computed by updating the current estimate with the optimal auxiliary displacement field. The optimal geometry is thus defined as the one, which is close to the previous estimated auxiliary geometry and satisfies the sparsity constraint.

  Let dU_k be the auxiliary variable of the 3D displacement field and U_k the corresponding auxiliary geometry defined as $\mathrm{U}_k = \mathrm{G}_k + \mathrm{dU}_k$. The minimization problem is reformulated as follows:

$$\{\mathrm{dU}_k^*, \mathrm{G}_k^*\} = \arg\min_{\mathrm{dU}_k, \mathrm{G}_k} \|\boldsymbol{A}\cdot \mathrm{dU}_k - \boldsymbol{b}\|_2^2 + \frac{1}{2\theta}\|\mathrm{U}_k - \mathrm{G}_k\|_2^2 + \gamma\|\mathbf{G}\mathrm{G}_k\|_1. \tag{6.24}$$

Given the current estimate of the geometry G_k, the minimization of the first step corresponds to a least squares problem, expressed as:

$$\begin{aligned} dU_k^* &= \arg\min_{dU_k} \|\boldsymbol{A} \cdot dU_k - \boldsymbol{b}\|_2^2 + \frac{1}{2\theta}\|U_k - G_k\|_2^2 \\ &= \arg\min_{dU_k} \|\boldsymbol{A}' \cdot dU_k - \boldsymbol{b}'\|_2^2. \end{aligned} \tag{6.25}$$

The analytic derivation of system is presented in Appendix A. Let U_k^* be the optimal auxiliary geometry. It is computed as $U_k^* = G_k + dU_k^*$. Let G_k^* be the optimal geometry, it is given by:

$$G_k^* = \arg\min_{G_k} \frac{1}{2\theta}\|U_k^* - G_k\|_2^2 + \gamma\|\mathbf{G}G_k\|_1. \tag{6.26}$$

Optimizing a L2 data term with a L1 has been extensively studied in Chambolle [2004] and off the shelf implementations are provided.

6.6 Experimental Evaluation

To validate the improvement in the appearance estimation, we apply the geometry and appearance Super-Resolution estimation on different datasets and we compare the results with the simple appearance Super-Resolution appearance estimation framework. We regenerate images with the reconstructed appearance and geometry in one case and with the super-resolved appearance in the other case and compare them visually.

6.6.1 Comparison with Simple Super-Resolution Appearance

Experiments are performed on the middlebury datasets TEMPLERING, DINORING. TEMPLERING dataset consists of 47 images of size 640×480 pixels and DINORING of of 48 images of size 640×480 pixels. In both cases, initial 3D reconstructions are obtained using a visual hull implementation (Franco and Boyer [2009]) and a texture atlas is computed (Sheffer et al. [2005]). Given the input images and the initial geometry, the proposed framework is applied to enhance geometry and super-resolve appearance with up-sampling factor of 2. The reconstructed texture is of size 1280×560 and images are rendered using both reconstructed appearance and geometry. To demonstrate the improvement in the quality by jointly accounting for geometry and appearance we apply the simple Super-Resolution appearance estimation and we qualitatively compare the rendered images.

It is interesting to note that each dataset represents a different category of shape distribution. The temple in TEMPLERING dataset consists of small smooth patches interrupted by strong edges, a property that is mathematically expressed through the sparse prior discussed in 6.3.2. DINORING dataset on the other side is an example of objects with smooth surfaces and a Gaussian prior over the shape, introduced in 6.3.2, is applied. The corresponding algorithms are thus validated. We should also remark that the appearance information in the case of TEMPLERING dataset includes high details compared to the DINORING dataset, which can be more informative

and may lead to better estimations.

It can be generally observed that the proposed appearance geometry Super-Resolution framework provides higher visual quality compared to the simple Super-Resolution appearance estimation method in both datasets, as shown in Figures 6.2, 6.3 and 6.4. This is particularly visible in the TEMPLERING dataset, where the rich photometric information can lead to remarkable visual improvement of the geometry. In Figures 6.2 and 6.3 rendered images at two different viewpoints of the TEMPLERING demonstrate the regain in geometric details and the focus and sharpness in appearance. DINORING dataset appears to be more challenging since the underlined texture is not informative to restrict the estimation problem and local minima in geometric update could appear. In this case the prior over the surface regularizes the problem and geometric details are recovered, as shown in Figure 6.4.

6.7 Conclusion

In this chapter we have introduced a framework which jointly enhances geometry and super-resolves appearance of 3D objects. In particular we have proposed an iterative algorithm to optimize for high quality textures and for geometric displacements to best fit the observations.

The qualitative experimental evaluation of the proposed geometry and appearance Super-Resolution estimation demonstrates that it is possible to achieve higher visual quality of appearance by jointly optimizing both parameters. Sharpness of the resulted textures increase considerably and geometric details are also recovered. These reconstructed geometric features could be interpreted as additional high frequency geometric information which can be further leveraged to add effects into the rendering. It is important to note that the geometric optimization intends to increase the quality of the appearance and the realism of the rendered images and not to necessarily lead to an estimated geometry closer to the ground truth. This is mainly due to the fact that the misalignments, which geometry aims to explain, do not originate only from deviations from the true geometry.

In overall, the conditional dependence of appearance and geometry of an object given a set of captured images proves to be beneficial to the problem of appearance modelling. Although this is by itself a contribution, it opens the avenue for new directions. Investigating how this dependence in a Super-Resolution framework can be exploited to achieve high quality geometric reconstructions is an interesting future work.

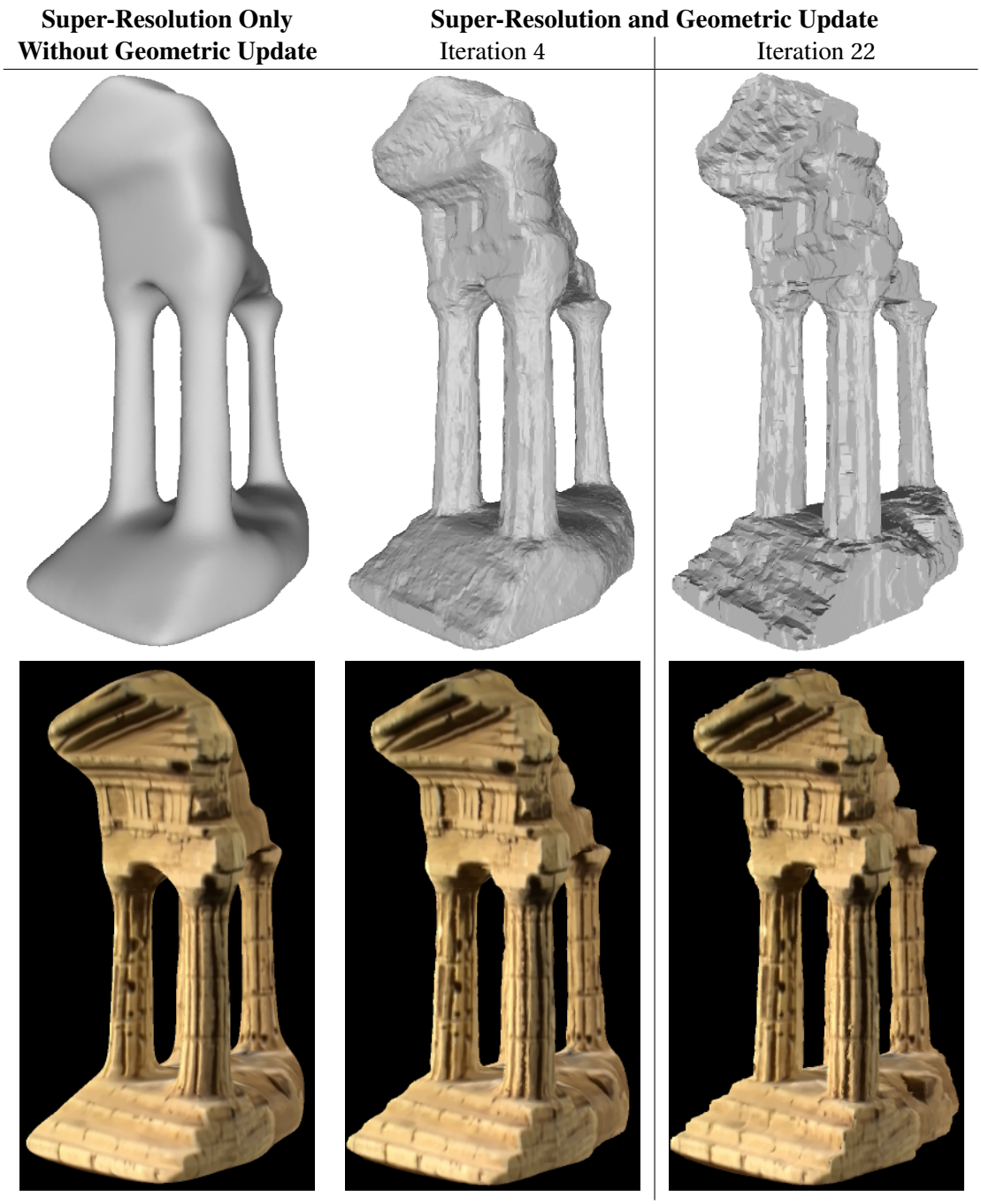

Figure 6.2 – Comparison on TEMPLERING. Crisp details are reconstructed through the geometry and appearance Super-Resolution estimation.

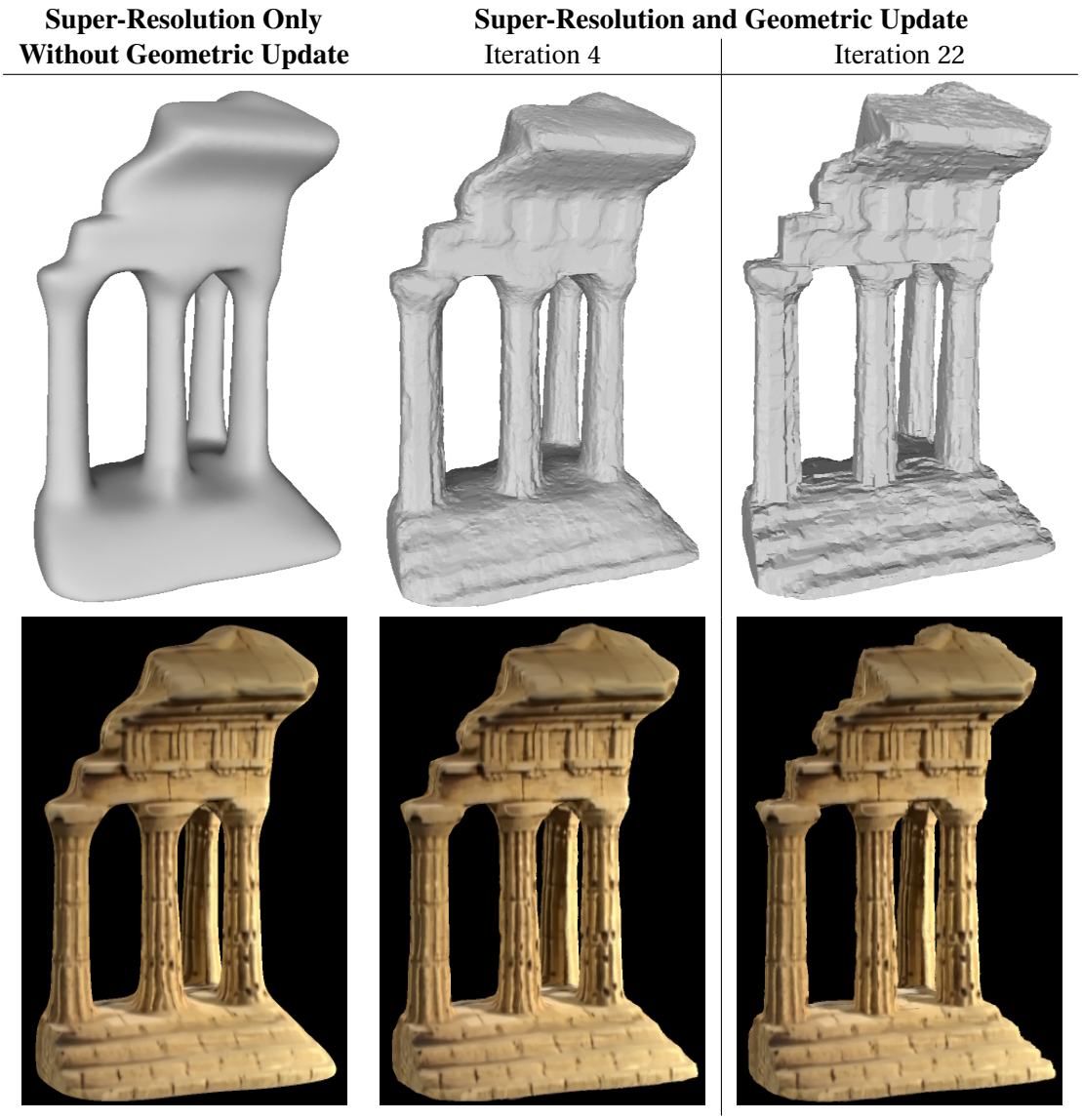

Figure 6.3 – Comparison on TEMPLERING. The sparse prior on the geometry successfully regains details. Sharp and more realistic renderings are achieved.

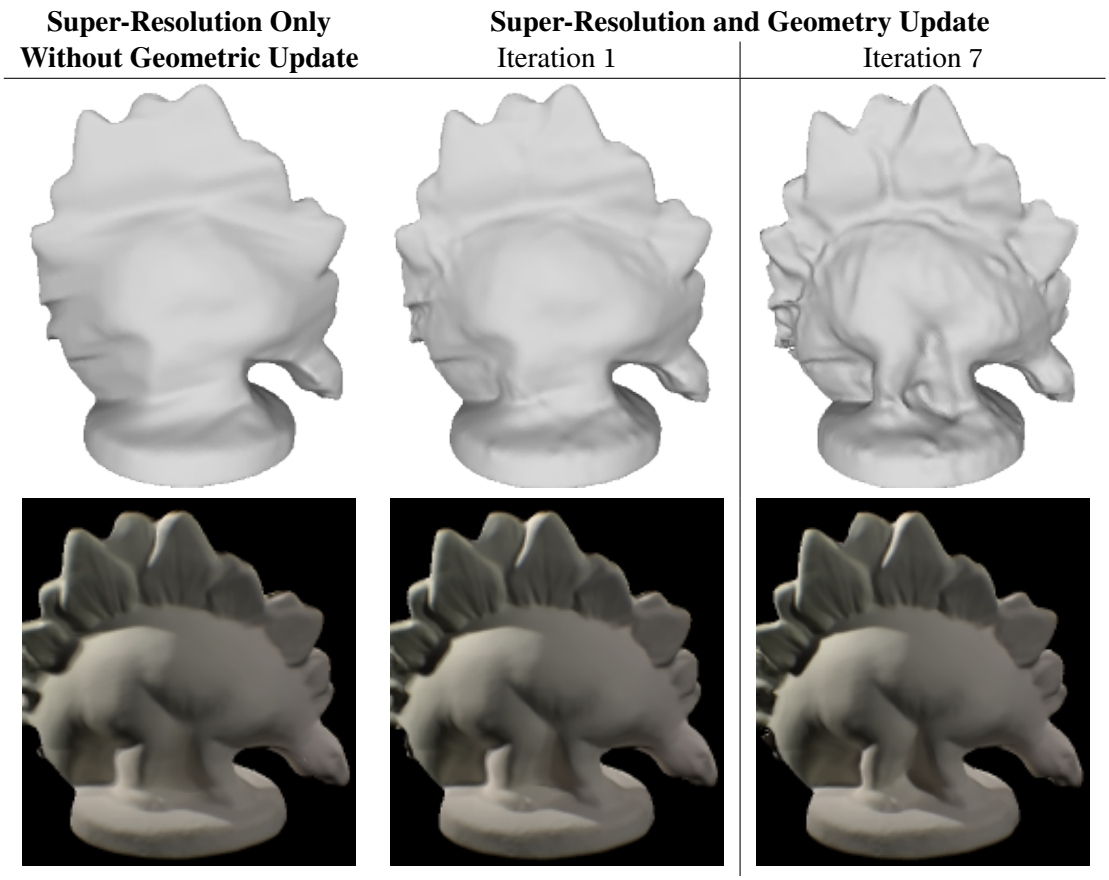

Figure 6.4 – Comparison on DINORING dataset. Focused areas around the legs and the backbone.

7 Application: Dynamic Appearance Representations

7.1 Introduction and Motivations

The previous chapters have established the feasibility to provide a high quality, compact appearance representation of an object captured at multiple viewpoints and through time by exploiting the available visual information and by leveraging in optimal way the relation between the geometric modality and the appearance. The performance of the proposed model is however bounded by the size of the temporal segments. Considering large time intervals, where the object undergoes several movements and the lighting conditions may change dramatically, the assumption of a temporally consistent appearance is not anymore valid and further study needs to be carried out.

In this chapter, we relax our previous assumption of small appearance variabilities, we consider the case of dynamic scenarios and we aspire to tackle some of its challenges. In particular, we address the problem of building efficient appearance representations of shapes observed from multiple viewpoints and in several movements. With the emergence of 3D multi-view video capturing systems, a tremendous amount of image data, capturing moving subjects, is now available. The need for more compact representations, which will replace these datasets, is thus inevitable. While geometric representations for spatio-temporal models of moving objects have been widely studied, appearance information, as provided by the observed images, is mainly considered on a per frame basis, and no method yet addresses the case where several temporal sequences of a shape are available. This is despite the obvious visual redundancy with which the appearance of subjects is observed and stands as an additional support towards a more global strategy. We propose a per subject representation that builds on PCA to identify the underlying manifold structure of the appearance information relative to a shape. The resulting Eigen representation encodes shape appearance variabilities due to viewpoint and motion, with Eigen textures, and due to local inaccuracies in the geometric model, with Eigen warps. This research direction builds on three key points. First, our previously proposed appearance representation serves as an efficient tool to study globally the evolution of the intrinsic appearance properties. Second, the finding of our previous study that optical flow can successfully compensate geometric inaccuracies, advocates its consideration in the general study of appearance variability. Third, linear subspaces have been proven to be suitable for encoding visual variability (Nishino et al. [2001], Cobzas and Jägersand [2002]). As opposed to the work of Nishino et al. [2001], which

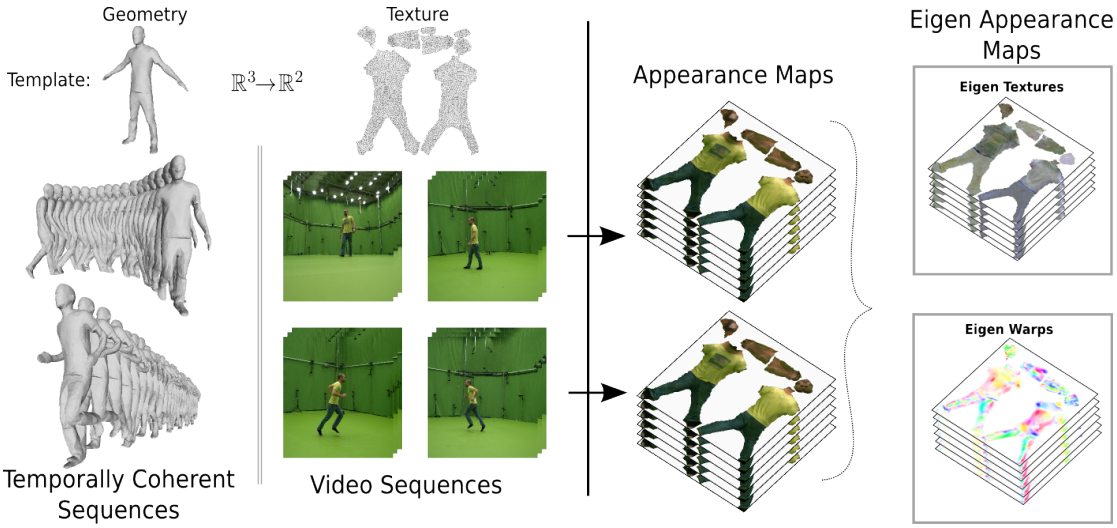

Figure 7.1 – Overview: Time consistent shape modelling provides datasets of appearance maps. Our proposed method exploits the manifold structure of these appearance information through PCA decomposition to generate the Eigen appearance maps relative to a shape.

uses linear subspaces to encode the appearance variability of static objects under light and viewpoint changes at the polygon level, we use linear subspaces for full body appearance and over multiple sequences. Compared to the work of Cobzas and Jägersand [2002], which also builds a local basis to capture geometric and photometric variations for each divided region of the image, the proposed dynamic appearance representation unifies this concept in the common texture space, which in turns results in a view independent and thus more compact representation.

In Section 7.2 we discuss about the underlying variabilities in the appearance data sets. Linearities, non-linearities, low and high frequency information are identified as intrinsic properties of these spatio-temporal samples. The insight we gain through this analysis drives us to recast in Section 7.3 the problem into a linear dimensionality reduction problem yielding the Eigen Appearance Maps. The performance of the method is evaluated in Section 7.4 and direct applications of it are proposed in Section 7.5. Limitations and future directions are discussed and summarized in 7.6 and 7.7.

7.2 Dynamic Nature of Appearance Maps

This ability to now record datasets of subject motions results in a massive amount of image information. While dynamic shape representations have been extensively studied, from temporally coherent representations over a single sequence, to shape spaces that can encode both pose and subject variabilities over multiple sequences and multiple subjects, appearance representations have received less attention in this context.

Our proposed appearance representation serves as an efficient way to study globally the evolution of the intrinsic appearance properties. Through the compact view-independent appearance representation, we express in a common appearance space the visual information per time frame. Considering large time intervals we obtain multiple appearance maps, which stand as spatio-

temporal samples. In so doing we create large datasets of texture maps, which encode appearance variabilities evolving across time. A careful observation on these spatio-temporal data reveals basic intrinsic properties, which can be further leveraged to provide efficient representations. In particular, two main categories can be identified: the first is related to the property of temporal consistency; the second is based on the property of linearity. Visual characteristics that appear to be permanent across time are defined as low frequency information while characteristics that arise in an inconsistent way correspond to high frequency. Non-linearities in appearance are observed when scene undergoes large geometric variations, while linearities are mainly due to illumination changes and can be identified in cases with small geometric variations.

These observations are of great interest in the development of our next research idea. Arguably, the nature of the appearance maps is highly dynamic. Linearities due to illumination changes, non-linearities due to geometric inaccuracies, temporal consistent information encoding low frequency appearance information are all important intrinsic properties of the spatio-temporal appearance samples, which can be exploited in order to provide efficient representations.

7.3 Eigen Appearance Maps

To exploit the intrinsic appearance properties and encode the appearance variability of a dynamic subject, observed over one or several temporal sequences, we recast the problem into a dimensionality reduction problem. As an additional support towards this direction stand the findings of previous works Nishino et al. [2001], Cobzas and Jägersand [2002], which demonstrate that linear subspaces are suitable for encoding visual variability. To adopt such direction it is important to eliminate the existing non-linearities. From our previous study it has been shown that optical flow successfully compensates geometric inaccuracies, which advocates its consideration in the general study.

Our strategy is to remove these non-linearities with state-of-the-art geometric and image-space alignment techniques, so as to reduce the problem to a single texture space, where the remaining image variabilities can be straightforwardly encoded in linear subspaces. Among the remaining variabilities in the aligned texture maps, we distinguish those caused by illumination changes and those caused by geometric inaccuracies. To separately compensate for these misalignments we introduce realignment warps in the texture domain resulting in a set of deformation fields. We call them warp fields. Both sets of aligned texture maps and warp fields can be decomposed using PCA, yielding Eigen textures and Eigen warps respectively. The full appearance information of all subject sequences can then be compactly stored as linear combinations of Eigen textures and Eigen warps. The main steps of the method below are depicted in Figure 7.2 and detailed in the following sections.

1. Texture deformation fields that map input textures to, and from, their aligned versions are estimated using optical flows. Given the deformation fields, Poisson reconstruction is used to warp textures.

2. PCA is applied to the aligned maps and to the texture warps to generate the Eigen textures and the Eigen warps that encode the appearance variations due to, respectively, viewpoint, illumination, and geometric inaccuracies in the reference model.

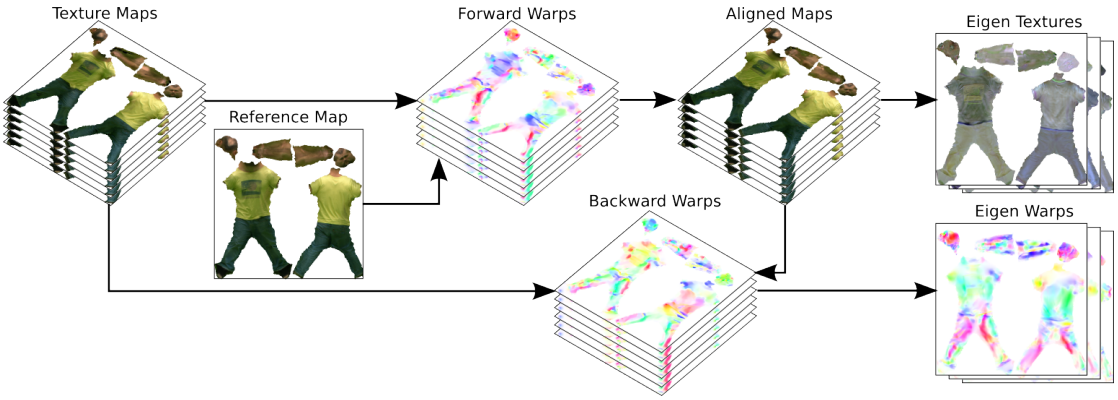

Figure 7.2 – Method pipeline from input textures (left) to eigen maps (right).

Note that due to texture space discretization, the warps between textures are not one-to-one and, in practice, two separate sets of warps are estimated. Forward warps map the original texture maps to the reference map. Backward warps map the aligned texture maps back to the corresponding input textures (see Figure 7.2).

7.3.1 Eliminating Non-Linearities

To eliminate the geometric non-linearity, we first apply state of the art tracking techniques (Allain et al. [2015]) to remove the non-linear pose component by aligning a single subject-specific template to all the subject's sequence. By aligning sequence geometries to a single template shape, we can then extract the texture maps of a subject over different motion sequences in a common texture space using our previous model. Textures are then aligned in order to reduce geometric errors resulting from calibration, reconstruction and tracking imprecisions. Such alignment is performed using optical flow, as described below, and with respect to a reference map taken from the input textures.

An exhaustive search of the best reference map with the least total alignment error over all input textures is prohibitive since it requires N^2 alignments given N input textures. We follow instead a medoid shift strategy over the alignment errors. The alignment algorithm first initializes the reference map as one texture from the input set. All texture maps are then aligned to this reference map, and the alignment error is computed as the cumulative sum of squared pixel differences between the reference and the aligned texture maps. The medoid over the aligned texture maps, with respect to alignment error, then identifies the new reference map. These two steps, alignment and medoid shift, are iterated until the total alignment error stops decreasing.

7.3.1.1 Dense texture correspondence with optical flow

The warps $\{w_k\}$ in the alignment algorithm, both forward and backward in practice, are estimated as dense pixel correspondences with an optical flow method (Sánchez Pérez et al. [2013]). We mention here that the optical flow assumptions: brightness consistency, spatial coherency and temporal persistence, are not necessarily verified by the input textures. In particular, the brightness

consistency does not hold if we assume appearance variations with respect to viewpoint and illumination changes. To cope with this in the flow estimation, we use histogram equalization as a preprocessing step, which presents the benefit of enhancing contrast and edges within images. Additionally, local changes in intensities are reduced using bilateral filtering, which smooths low spatial-frequency details while preserving edges.

7.3.1.2 Texture warping

Optical flows give dense correspondences $\{w\}$ between the reference map and the input textures. To estimate the aligned textures $\{\mathrm{T}\}$, we cast the problem as an optimization that seeks the texture map which, once moved according to w, best aligns with the considered input texture both in the color and gradient domains. Our experiments show that solving over both color and gradient domains significantly improves results as it tends to better preserve edges than with colors only. This is also demonstrated in works that use the Poisson editing for image composition, (e.g. Pérez et al. [2003], Chen et al. [2013]) or interpolation, (e.g. Linz et al. [2010], Mahajan et al. [2009]). We follow here a similar strategy. Other aligning methods can be considered.

7.3.2 Eigen Textures and Eigen Waprs

At the end of the previous step, non-linearities are successfully eliminated and statistical analysis on the aligned texture maps can be carried out. The remaining appearance variabilities in the aligned texture maps are linear and they are mainly caused due to illumination changes and small geometric variations. The last ones correspond to small misalignments, originating from local reconstruction inaccuracies and registration errors. To cope with these fine-scale misalignments, we estimate realignment warps in the texture domain, which due to their low-magnitude are also linear. We thus perform Principal Component Analysis on the textures and on the warp data separately to find the orthonormal bases that encode the main modes of variation in the texture space and in the warp space independently. We refer to vectors spanning the texture space as Eigen textures, and to vectors spanning the warp space as Eigen warps.

Both sets of aligned texture maps and warps are expressed in a vectorized form, including only pixels that fall inside the active regions within texture maps. Let $\{\mathrm{T}_i\}_{i\in[1..F]}$ be the set of aligned texture maps and $\{\mathrm{W}_i\}_{1\leq i\leq F-1}$ be the set of computed warps, where F the total number of frames available for the subject under consideration. We note the Eigen textures with T_{E_i} and the Eigen warps with W_{E_i}.

We consider first the texture maps and we present the Principal Component Analysis applied on them. We start by computing the mean image $\bar{\mathrm{T}}$ and the centered data matrix M from aligned texture maps $\{\mathrm{T}_i\}_{i\in[1..F]}$:

$$\bar{\mathrm{T}} = \frac{1}{F}\sum_k \mathrm{T}_k \ , \ M = \begin{bmatrix} | & & | \\ \mathrm{T}_1 - \bar{\mathrm{T}} & \dots & \mathrm{T}_F - \bar{\mathrm{T}} \\ | & & | \end{bmatrix}. \tag{7.1}$$

Traditionally, the PCA basis for this data is formed by the Eigen vectors of the covariance matrix

MM^T, of size $N \times N$,where N is the dimension of the vectorized representation of active texture elements. Due to the high dimensionality of texture maps, it becomes computationally prohibited to perform such analysis. To give orders of magnitude for our datasets, $N = 22438995$ and $F = 207$ for the TOMAS dataset, and $N = 25966476$ and $F = 290$ for the CATY dataset that will be presented in the next section. To surpass this limitation we make use of the fact that the non zero eigen values of MM^T are equal to the non zero Eigen values of $M^T M$, of size $(F \times F)$ this time, and that they are at most: $\min(F, N) - 1$. Based on this observation, and since $F << N$ in our experiments, we solve the characteristic equation $\det(MM^T - \alpha \cdot I_N) = 0$ by performing Singular Value Decomposition on the matrix $M^T M$, as explained in Turk and Pentland [1991a]:

$$M^T M = D \Sigma D^T \ , \ D = \begin{bmatrix} | & & | \\ V_1 & \dots & V_F \\ | & & | \end{bmatrix}, \tag{7.2}$$

where D contains the $(F-1)$ orthonormal Eigen vectors $\{V_i\}$ of $M^T M$, and $\Sigma = \text{diag}(\alpha_i)_{1 \leq i \leq F}$ contains the eigen values $\{\alpha_i\}_{1 \leq i \leq F-1}$. We can then write:

$$M^T M V_i = \alpha_i V_i \ , \ i \in [1..F-1], \tag{7.3}$$

and hence:

$$MM^T \underbrace{MV_i}_{\mathrm{T}_{E_i}} = \alpha_i \underbrace{MV_i}_{\mathrm{T}_{E_i}} \ , \ i \in [1..F-1], \tag{7.4}$$

where T_{E_i} are the Eigen vectors of MM^T.

The set $\{\mathrm{T}_{E_i}\}$ forms the orthonormal basis of the aligned texture maps after normalization, namely the Eigen textures and they encode the main modes of variation in the texture space. A similar analysis is carried out to obtain the orthonormal basis of the warp space $\{\mathrm{W}_{E_i}\}_{1 \leq i \leq F-1}$, the Eigen warps.

7.3.3 Texture generation

Given the Eigen textures and the Eigen warps, and as shown in Figure 7.3, a texture can be generated by first creating an aligned texture by linearly combining Eigen textures and second de-aligning this new texture using another linear combination of the Eigen warps.

7.4 Performance Evaluation

To validate the estimation quality of our method, we apply our estimation pipeline to several datasets, project and warp input data using the built eigenspaces, then evaluate the reconstruction error. To distinguish the different error sources, we evaluate this error both in texture space before projection, and in image domain by projecting into the input views, as compared to the original views of the object and the texture before any reconstruction in texture space, estimated in our pipeline using Tsiminaki et al. [2014]. For the image error measurement, we use the 3D model that was fitted to the sequence, as tracked to fit the test frames selected Allain et al. [2015], and

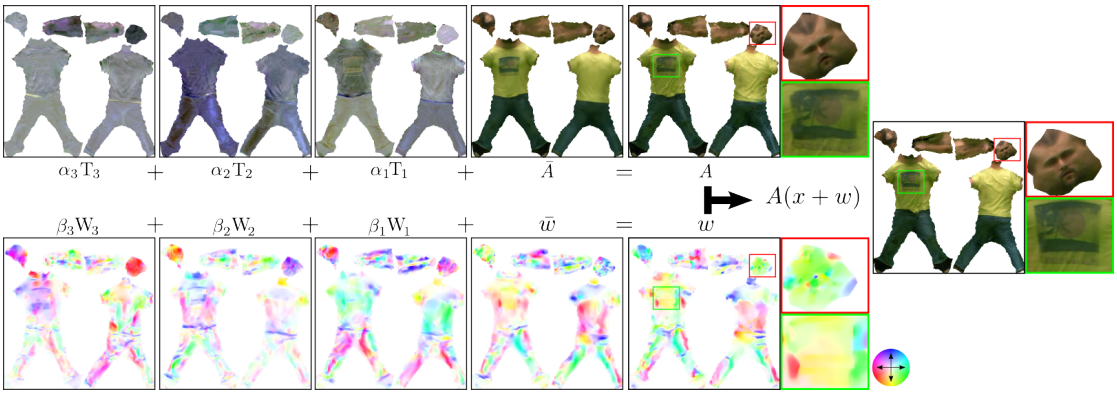

Figure 7.3 – Texture map generation by linear combination.

render the model as textured with our reconstructed appearance map, using a standard graphics pipeline. In both cases, we use the structural similarity index (SSIM) Wang et al. [2004] as metric to compare to the original. All of our SSIM estimates are computed in the active regions of the texture and image domains, that is on the set of texels actually mapped to the 3D model in the texture domain, and only among actual silhouette pixels in the image domain.

We study in particular the compactness and generalization abilities of our method, by examining the error response as a function of the number of eigen components kept after constructing the linear subspaces, and the number of training images selected. For all these evaluations, we also provide the results for a naive PCA strategy, where only a set of eigen appearance maps are built in texture space and use to project and reconstruct textures, to show the performance contribution of including the Eigen warps.

For validation, we used two multi-sequence datasets: (1) the TOMAS dataset which consists of 4 different sequences left, right, run and walk with 207 total number of frames and 68 input views each captured at resolution 2048x2048 pixels per frame; and (2) the CATY dataset: low, close, high and far jumping sequences with 290 total number of frames and 68 input views each captured at resolution 2048x2048 pixels per frame.

7.4.1 Estimation Quality and Compactness

We study the quality and compactness of the estimated representation by plotting the SSIM errors of reconstructed texture and image estimates of our method against naive PCA, for the two multi-sequence datasets (Figure 7.4). Note that all texture domain variability could be trivially represented by retaining as many Eigen textures as there are input images, thus we particularly examine how the quality degrades with the fraction of Eigen components kept. In the case of image domain evaluations, we plot the average SSIM among all viewpoints. Our method outperforms naive PCA in image and texture domains on both datasets, achieving higher quality with a lower number of Eigen components, and only marginally lower quality as the number of components grows, where the method would be anyway less useful. Higher number of Eigen components marginally favors naive PCA, because naive PCA converges to input textures when increasing the Eigen textures retained by construction, whereas our method hits a quality plateau

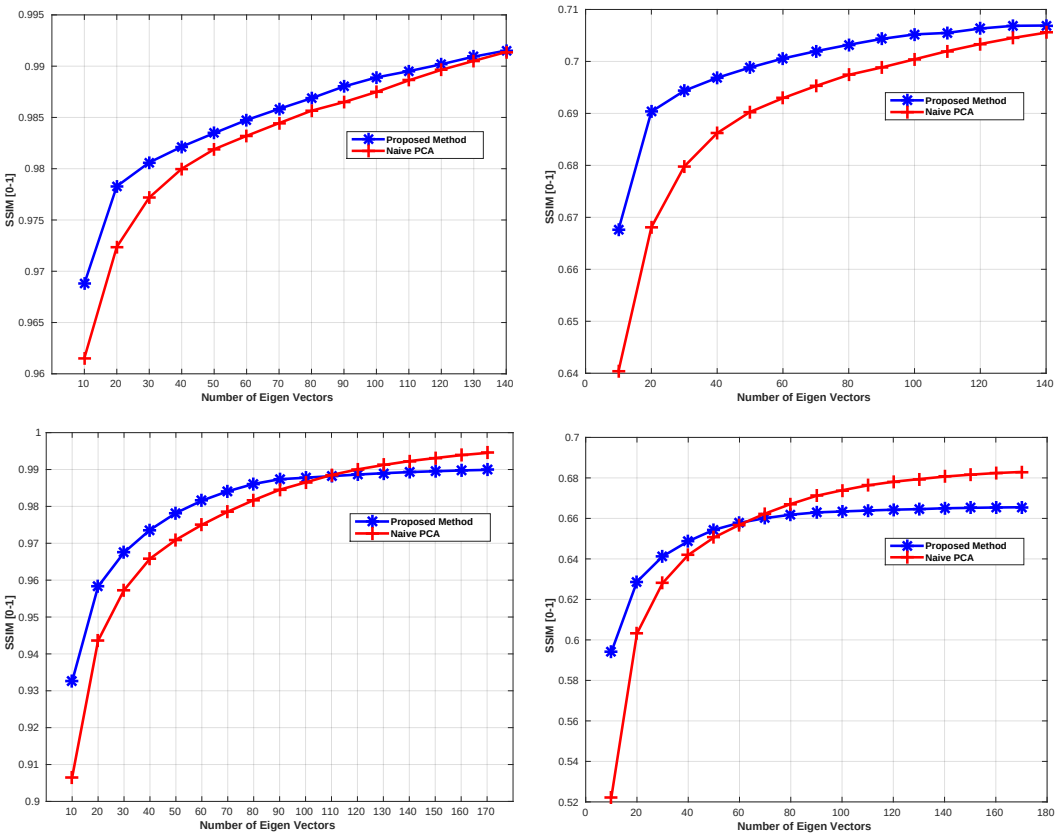

Figure 7.4 – Reconstruction error on TOMAS and CATY dataset from top to down in Texture and Image Domain from left to right.

due to small errors introduced by texture warp estimation and decomposition. For both datasets, virtually no error (0.98 SSIM) is introduced by our method in the texture domain with as low as 50 components, a substantially low fraction compared to the number of input frames (207 and 290). This illustrates the validity of the linear variability hypothesis in texture domain. The error is quite higher in the image domain (bounded by 0.7) for both our method and naive PCA, because measurements are then subject to fixed upstream errors due to geometric alignments, projections and image discretizations. Nevertheless, visually indistinguishable results are achieved with 50 Eigen components (images and warps), with a significant compactness gain.

7.4.2 Generalization ability

In the previous paragraph, we examined the performance of the method by constructing an Eigen space with all input frames. We here evaluate the ability of the model to generalize, *i.e.* how well the method reconstructs textures from input frames under a reduced number of examples that don't span the whole input set. For this purpose, we perform an experiment using a varying size training set, and a test set from frames not in the training set. We use a training set comprised of randomly selected frames spanning 0% to 60% of the total number of frames, among all sequences and frames of all datasets, and plot the error of projecting the complement frames

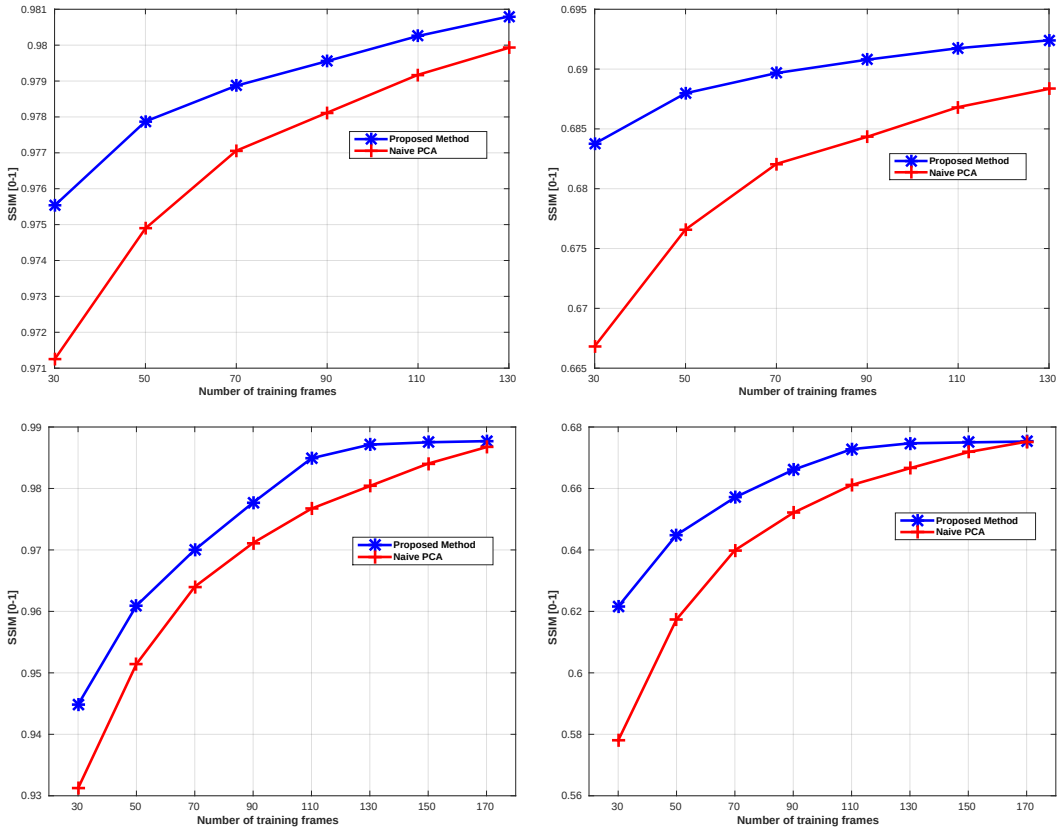

Figure 7.5 – Generalization error on THOMAS and CATY dataset from top to down in Texture and Image Domain from left to right.

on the corresponding Eigen space (Figure 7.5). The experiment shows that our representation produces a better generalization than naive PCA, *i.e.* less training frames need to be used to reconstruct a texture and reprojections of equivalent quality. For the TOMAS dataset, one can observe than less than half training images are needed to achieve similar performance in texture space, and a quarter less with the CATY dataset.

7.5 Applications

We investigate below two applications of the appearance representation we propose. First, the interpolation between frames at different time instants and second, the completion of appearance maps at frames where some appearance information is lacking due to occlusions or missing observations during the acquisition.

7.5.1 Interpolation

In our framework, appearance interpolation benefits from the pre-computed warps and the low dimensionality of our representation to efficiently synthesize compelling new appearances with reduced ghosting-artefacts. It also easily enables extension of appearance interpolation from pairwise to multiple frames. Assume that shapes between two given frames are interpolated

using a standard non-linear shape interpolation, for instance Xu et al. [2005]. Consider then the associated aligned textures and associated warps at the given frames. We perform a linear interpolation in the Eigen texture and Eigen warp spaces respectively by blending the projection coefficients of the input appearance maps. Poisson warping, as introduced in section 7.3.1.2 is used to build de-aligned interpolated texture with the interpolated backward warp. Figure 7.6 compares interpolation using our pipeline to a standard linear interpolation for 4 examples with the CATY and TOMAS datasets. Note that our method is also linear but benefits from the alignment performed in the texture space to reduce interpolation artefacts, as well as from the simplified computational aspects since interpolation applies to projection coefficients only.

7.5.2 Completion

As mentioned earlier, appearance maps can be incomplete due to acquisition issues. For instance, as shown in Figure 7.7, during the running sequence the actor TOMAS bends his knees in such a way that the upper parts of his left and right shins become momentarily hidden to the acquisition system. This results in missing information for those body parts in the texture maps and over a few frames. Such issue can be solved with our texture representation by omitting the incomplete frames when building our appearance representations, and then projecting these incomplete appearance maps in the Eigen spaces and reconstructing them using the projection coefficients and Poisson texture warping. Figure 7.7 shows two examples of this principle with occluded regions. Note however, that while effectively filling gaps in the appearance map, this completion might yet loose appearance details in regions of the incomplete map where information is not duplicated in the training set.

7.6 Discussion

The performance evaluation of our proposed methods in terms of quality, compactness and generalization ability and the comparison with a naive PCA strategy demonstrate the high fidelity of our method. The qualitative and quantitative results verify the soundness of our initial insights regarding the intrinsic properties of the appearance maps. In particular, the hypothesis on the existence of the linear variability in the texture domain as well as the hypothesis on the compensation of the non linearities due to geometric inaccuracies through warp fields are both validated.

Direct applications of the proposed appearance representation demonstrate the significance of the method. New appearances with reduced ghosting-artefacts can be efficiently synthesized by performing interpolation in the Eigen texture and Eigen warp spaces. Appearance maps with incomplete parts due to acquisition issues can be successfully completed by projecting them in our proposed Eigen texture and Eigen warp spaces. An application which was not extensively studied and is of great interest is compression. Although in terms of compactness property the method requires few components to provide good quality representations, there is not a thorough theoretical analysis of this aspect. Among the limitations, the representation performances are dependent on the underlying geometries. Future strategies that combine both shape and appearance information would thus be of particular interest. The proposed model

could also be extended to global representations over populations of subjects.

7.7 Conclusion

In this Chapter, appearance variabilities of a subject observed from multiple viewpoints and in different motions are studied and encoded through a novel efficient representation.

Intrinsic appearance properties per time frame are firstly expressed in the reconstructed appearance maps of a subject in different motion. These appearance maps yield large spatio-temporal datasets, where intrinsic properties can be easily extracted. On one side, no-linearities can be mainly observed in cases where the scene undergoes large geometric variations. On the other side, linearities can be identified in cases with small geometric variations and significant illumination changes.

We firstly eliminate the non-linearities and we propose a straightforward representation which builds on PCA and decomposes into Eigen textures and Eigen warps. Appearance variations due to viewpoint and illumination changes are encoded through the Eigen textures while the visual variability due to geometric modelling imprecisions is encoded through the Eigen warps. The framework was evaluated on 2 datasets and with respect to: (i) its ability to accurately reproduce appearances with compact representations; (ii) its ability to resolve appearance interpolation and completion tasks.

By casting the problem of efficient appearance representation as a linear dimensionality reduction, we arguably propose a baseline in the research community.

Figure 7.6 – Interpolation examples using linear interpolation (left) and our pipeline (right). From left to right: Input frames, Interpolated models, and a close-up on the texture maps (top) and the rendered images (bottom).

Figure 7.7 – Completion examples. From left to right: Input and completed models, close-up on input and completed texture maps (top) and rendered images (bottom).

8 Conclusions

8.1 Summary and main findings

Recent years have seen the emergence of image-based 4D modelling with many applications in various domains of virtual-augmented reality, sports science, social studies, entertainment and advertising. Given multiple video sequences, captured from different viewpoints, image-based 4D modelling methodologies represent dynamic scenes with spatio-temporal models. These models comprise geometric and appearance information allowing for scenes to be recorded and reused. In order to achieve high realism, both forms of information need to be replicated with great details. Geometry and appearance can be interpreted as the counterparts to the desired realism. This thesis has studied the appearance aspect and has aspired to bridge some of the gaps of the state-of-the-art.

Due to the massive amount of visual data and due to the increasing demands of real-world applications, a new reality of the appearance modelling problem has been established. On one hand, the easily accessible image data offer great potentials for high fidelity appearance reconstructions. On the other hand, they lead to more complex scenarios and as such require more elaborated ways of processing. Furthermore, real-world applications request fast renderings and high compression rates. State-of-the-art methodologies appear to be insufficient to cope with the new requirements. We have investigated these limitations and concretely, we have proposed a compact, view-independent and high resolution appearance representation.

As a first step, we introduced a common space to express the available visual information and thus a compact representation of the appearance. In this manner, multi-camera observations are interpreted as different realizations of one common appearance and we defined a linear model to exhibit directly this relation. Through the common texture space, appearance redundancy as well as view-independent appearance properties have been exploited. Leveraging this linear appearance model, we studied the general conditions of the appearance estimation in the multi-view case and identified the main principles that govern Super-Resolution techniques.

Stepping on these findings we enhanced the model with the Super-Resolution (SR) fundamentals integrated in the multi-view context. The proposed model simulates the image capturing process with a concatenation of linear operations, through which the unknown appearance (texture) generates the image observations at several viewpoints. The Super-Resolution model benefits from the non redundant appearance information which is encoded in the multiple obser-

vations and is expressed in the form of aliasing due to geometric, calibration inaccuracies related to the 3D nature of the problem and due to noise originated from the 2D nature of the problem. A unique, view-independent, high resolution appearance representation has been thus provided to allow for the object to be rendered not only at given viewpoints but also at new ones with fine details.

An extension to the temporal domain has been introduced as a natural evolution of the model. We considered small time intervals, where the appearance of the object does not change dramatically and we introduced a temporally coherent Super-Resolution appearance representation. Additional visual information fused in time has been exploited and has led to appearance reconstructions with greater details. The appearance representation is not only view-independent but also time-independent under the constraint that the time instance belongs to the predefined time interval.

To exploit the interdependency of geometry and photometry in a more coherent way, we integrated geometry as an additional variable. Driven by the fact that jointly estimating appearance and geometry should prove mutually beneficial and driven also by the success of the already presented model to compensate locally geometric inaccuracies, we advanced the framework to jointly enhance geometry and super-resolve appearance of 3D objects. Qualitative experimental evaluation demonstrated that it is possible to achieve higher visual quality of appearance by jointly optimizing both parameters.

To account for more complex dynamic scenarios we assumed large time intervals, where the object undergoes several movements, and we introduced a view-independent representation which builds on PCA and decomposes the underlying appearance variability into Eigen textures and Eigen warps. Appearance linearities due to viewpoint and illumination changes have been encoded through the Eigen textures while non-linearities due to geometric modelling imprecisions have been encoded through the Eigen warps. It has been shown that the proposed framework is able to accurately reproduce appearances with compact representations and to resolve appearance interpolation and completion tasks.

8.2 Discussion of methods

Concluding our study on the problem of appearance modelling in the multi-view case scenario, it is natural to ask oneself whether and to what extent the main questions and limitations of the state-of-the-art have been answered.

Regarding the demand on compactness of appearance modelling, our proposed model has represented the appearance by a view-independent, two dimensional texture. Instead of storing the captured images for new renderings, we have demonstrated that it is feasible to use only one image (texture map) that encloses the whole visual information and can be reused. This arguably leads to a significant decrease of storage as well to an increase of transmission speed. On top of that, due to its property of view-independency, the object can be rendered at new viewpoints allowing for real time renderings.

In terms of high quality the proposed texture map is shown to be a super-resolved appearance representation. We do not simply exploit the appearance redundancy through naive weighting

schemes but we go beyond this step and we explore and unfold the hidden non-redundant visual information. The novelty behind our proposed model is the idea to turn the existent and inevitable geometric and calibration inaccuracies into a source of information. We have further extended this idea to the temporal domain to exploit the additional misalignments due to tracking imperfections in small temporal segments. Introducing small time intervals the corresponding temporally coherent, super-resolved appearance representations allows for the object to be rendered with high accuracy not only at new viewpoints but also at new time instances in these predefined intervals. This translates into a larger decrease of storage and it facilitates even more the real time rendering process.

The price however, that one needs to pay for such a high quality appearance, is the computational cost. Super-Resolution algorithms are computationally complex and thus slow. The execution time of our proposed algorithm is in the range of 15 minutes to 30 per iteration depending on the dataset, number of views and number of frames. Due to these limitations, Super-Resolution technique is still not often present in real-world applications. In order to fill this gab and transfer the research results to the market, the algorithms need to be faster. Such an increase of speed from the research-level code can indeed be achieved using code optimization and faster hardware.

The geometric modelling that has been introduced in the framework as an additional step has been shown to be mutually beneficial for appearance and geometry. The achieved geometric enhancement allows for the possibility to capture visual effects such as high-lights, reflections and leads to further improvement of appearance. It is important however to note that the main goal of the geometric step is to improve the appearance by compensating in a more global way for the existing misalignments. In particular, it does not necessarily lead to a geometric reconstruction closer to the ground truth shape of the object, since the misalignments observed in the image space are caused not only by deviations from the true geometry. The increase of complexity is also another factor that should be considered since this step in the loop corresponds to an additional non convex on the geometry optimization step.

To account for the complexity of dynamic scenarios, where the subject undergoes several movements, we have introduced a compressive, view-independent representation which builds on PCA and decomposes the underlying appearance variability into Eigen textures and Eigen warps. This representation has been shown to resolve appearance interpolation in cases, where the source and target differ greatly. In addition to that, through this representation occluded regions are successfully filled in. It is interesting to note that representing appearance variability in large temporal segments is fundamentally one of non-linear dimensionality reduction problems. Yet, by firstly eliminating non-linearities and by proposing a linear PCA representation, appearance variations are successfully encoded and we thus propose a baseline for future.

8.3 Future Work

Future work naturally entails evaluating the limits of our proposed Super-Resolution appearance model to provide a deeper understanding of its performance. Quantitative and qualitative evaluations have demonstrated that higher magnification numbers lead to reconstructions of

higher quality. On the opposite side, the higher the magnification number the larger the memory requirements. To achieve a trade-off between quality and complexity, a magnification number of 2 has been experimentally demonstrated to be sufficient enough. A more sophisticated evaluation is however still missing and a theoretical analysis should be beneficial. Several previous works have addressed this issue in the monocular case. Lin and Shum Lin and Shum [2004] suggest that this factor is limited by the range of 1.6×5.7. A theoretical study of Robinson and Milanfar [2004] demonstrates that there is no single answer to the question of how many frames are needed, as it mostly depends on the amount of useful information encoded in the images. The level of complication increases even more in the case of 4D appearance modelling. The limits of the magnification number with respect to the number of frames really depend on the quality of the captured images, on the geometry of the scene, on the number of frames in which the object is at least partially visible and on the quality of tracking algorithms in the temporal domain. Despite the inherent difficulties it would be really informative to have an approximation of these bounds since this would help to decrease the complexity of the whole framework.

From the technical perspective, there is still space for further optimization, including making the flow, image and geometric update estimations parallel, more compact data-structures, C++ inner loop. All these enhancements will facilitate the experimental study of the trade-off between quality and complexity as well as the experimental evaluation on several datasets and under multiple conditions. An insight into the fundamental principles and bottlenecks of the geometric relaxation step with respect to the appearance modelling could be also achieved.

Another interesting avenue for further investigation regards the study of the geometric modelling in the appearance estimation framework. In the current thesis, the geometric enhancement aims to compensate for the misalignments apparent on the captured images and as such it does not necessarily lead to reconstruction closer to ground truth shape. It is thus of great interest to focus more on the geometric aspect and to investigate how the super-resolved appearance can drive the geometric reconstruction. In previous works Seitz et al. [2001], visual information together with regularization steps are used to drive the reconstruction problem. However, the problem is highly non convex and approximated visual information results in local minima. Since through our proposed model, an appearance representation with high details is now available we can expect more detailed geometric reconstructions. Extending even more the idea of Super-Resolution to shape modelling, we could even impose an up-sampling factor for the shape and solve for the high resolution geometry, which when is textured by the super-resolved appearance and observed by the multiple cameras generates the captured images.

Furthermore, several interesting directions can be found in dynamic scenarios. A study on the compression rates should be helpful to quantify the gain achieved through the proposed PCA representation. The proposed model could also be extended to global representations over populations of subjects. Based on the knowledge we obtained about the linearities and the non-linearities of appearance variations, other non-linear dimensionality reduction methods could be also proposed. Finally, since the representation performances are dependent on the underlying geometries, strategies that combine both shape and appearance information should be attempted.

Last but not least, it would be a great challenge to apply the proposed appearance modelling technique in scenarios with simple monocular capture of uncontrolled scenes. The emergence of

low-cost sensors has yield a new reality, where 3D capture from monocular video is now possible. Creating 3D content from footage taken from video camera or even mobile phone is indeed not an easy task since due to the low-cost sensors and uncontrolled scenes, these environments are more noisy compared to studios equipped with highly technical multi-camera rigs and green screens. This however should not appear as an obstacle of applying our Super-Resolution appearance modelling method. It is important to note the several key points that characterize these scenarios and serve as additional support towards that direction. First, the sources of noise are of similar nature with the ones studied in the multi-view context which, to name but a few, include misalignments, motion blur, shot noise during capturing process. Second, the demand on compactness is even higher due to memory limitations. Third, the main principle of Super-Resolution governs also these environments. The captured images are multiple observations which differ in sub-pixel precision and as such they encode new hidden, visual information. For all these reasons, an extension of our proposed Super-Resolution appearance modelling methods into these scenarios seems to be feasible and it would be interesting to evaluate its performance.

8.4 Conclusions

The present thesis has presented the problem of appearance modelling for 4D multi-view representations. It has reviewed the state-of-the-art and has reported that the leap to a compact, view-independent, high quality appearance representation able to represent large dynamic variability is still an open problem. To target these limitations it has introduced a novel appearance representation, a Super-Resolution estimation technique to provide high quality visual details and a linear dimensionality reduction method to encode the dynamic appearance variability of objects under several movements.

Beyond demonstrating the quality and compactness of the Super-Resolution appearance model, the current research study exploits the interdependency of geometry and appearance setting a new baseline for the problem of geometric reconstruction and provides a compressive representation to reduce the dimensionality of the temporal appearance variability. The proposed study marks a shift in 4D modelling problem towards high quality compact appearance representations and opens new research directions that can further advance the field.

A Optimizing Geometry

Developing further the intuition behind this minimization problem, the optimal geometry is given by an optimal update of the current estimation. In other words we are looking for the motion field, which being applied over the geometry generates the one, whose projection into the image space is close to the coordinates, updated with the estimated 2D warp field and maintains the statistical behavior imposed by the prior distribution. Let dG_k^* be this 3D update, which is vectorized and consists of $3\mathcal{V} \times 1$ elements, where $\mathcal{V}$ the number of vertices and 3 the spatial(x, y, z) coordinates. The optimal geometry is thus given by $\mathrm{G}_k^* = \mathrm{G}_k + \mathrm{dG}_k^*$ and the minimization problem is rephrased as:

$$\mathrm{dG}_k^* = \arg\min_{\mathrm{dG}_k} \sum_i \|\pi_i(\mathrm{G}_k + \mathrm{dG}_k) - (I_{uv}^{il} + \mathrm{W}_{i\,k}^{\,*})\|_2^2 + \gamma\Phi(\mathrm{G}_k + \mathrm{dG}_k) \tag{A.1}$$

Re-factorizing the terms of the above objective function, the minimization problem becomes:

$$\mathrm{dG}_k^* = \arg\min_{\mathrm{dG}_k} \|\boldsymbol{A} \cdot \mathrm{dG}_k - \boldsymbol{b}\|_2^2 + \gamma\Phi(\mathrm{G}_k + \mathrm{dG}_k) \tag{A.2}$$

The matrix $\boldsymbol{A}$ is a matrix with $\sum_{i=1}^{N_C} 2\mathrm{Vrt}_i$ rows and $3\mathcal{V}$ columns, where Vrt_i the number of visible vertices at camera c_i and N_C the number of cameras. An analytic derivation of that matrix is provided in Appendix A.

For a vertex

$$\mathrm{V}_l = \begin{bmatrix} x_l \\ y_l \\ z_l \end{bmatrix}$$

The unknown 3D motion field dG_k is vectorized, and it consists of $3\mathcal{V} \times 1$ elements, where $\mathcal{V}$ the number of vertices and 3 for the spatial(x, y, z) coordinates. The matrix $\boldsymbol{A}$ is a matrix with $\sum_{i=1}^{N_C} 2\mathrm{Vrt}_i$ rows and $3\mathcal{V}$ columns, where Vrt_i the number of visible vertices at camera c_i and N_C

the number of cameras. It has the form

$$\begin{bmatrix} \begin{bmatrix} \boldsymbol{A_1} \end{bmatrix} \\ \vdots \\ \begin{bmatrix} \boldsymbol{A_i} \end{bmatrix} \\ \vdots \\ \begin{bmatrix} \boldsymbol{A_{N_C}} \end{bmatrix} \end{bmatrix} \begin{bmatrix} \mathrm{dV}_1 \\ \vdots \\ \mathrm{dV}_l \\ \vdots \\ \mathrm{dV}_{\mathcal{V}} \end{bmatrix} = \begin{bmatrix} \begin{bmatrix} \boldsymbol{b_1} \end{bmatrix} \\ \vdots \\ \begin{bmatrix} \boldsymbol{b_i} \end{bmatrix} \\ \vdots \\ \begin{bmatrix} \boldsymbol{b_{N_C}} \end{bmatrix} \end{bmatrix} \tag{A.3}$$

The corresponding equation for each camera c_i writes $\boldsymbol{A_i}\,[\mathrm{dV}_1 \cdots \mathrm{dV}_l \cdots \mathrm{dV}_{\mathcal{V}}]^\top = \boldsymbol{b_i}$, where $\boldsymbol{A_i}$ is a diagonal matrix with $2\mathrm{Vrt}_i$ rows and $3\mathcal{V} \times 1$ columns and the $\boldsymbol{b_i}$ the vector of size $2\mathrm{Vrt}_i \times 1$. For each camera c_i the corresponding entry $\boldsymbol{A_i}$ is a diagonal matrix with $2\mathrm{Vrt}_i$ rows and $3\mathcal{V} \times 1$ columns of the form

$$\begin{bmatrix} A_i^1 & \emptyset & \cdots & \cdots & \emptyset \\ \emptyset & \ddots & \emptyset & \cdots & \emptyset \\ \vdots & \cdots & A_i^l & \cdots & \vdots \\ \emptyset & \cdots & \emptyset & \ddots & \emptyset \\ \emptyset & \emptyset & \cdots & \emptyset & A_i^{\mathcal{V}} \end{bmatrix} \begin{bmatrix} \mathrm{dV}_1 \\ \vdots \\ \mathrm{dV}_l \\ \vdots \\ \mathrm{dV}_{\mathcal{V}} \end{bmatrix} = \begin{bmatrix} b_i^1 \\ \vdots \\ b_i^l \\ \vdots \\ b_i^{\mathcal{V}} \end{bmatrix} \tag{A.4}$$

Let $I_{uv}^{il} = \begin{bmatrix} I_u^{il} \\ I_v^{il} \end{bmatrix}$ be the coordinates of the vertex V_l projected on view i and $\mathrm{W}_{il} = \begin{bmatrix} u_{il} \\ v_{il} \end{bmatrix}$ the corresponding estimated warp. Each entry A_i^l and b_i^l corresponds to the vertex V_l, which is visible from camera c_i and are computed as follows:

$$A_i^l = \begin{bmatrix} \pi_{i31}(I_u^{il} + u_{il}) - \pi_{i11} & \pi_{i32}(I_u^{il} + u_{il}) - \pi_{i12} & \pi_{i33}(I_u^{il} + u_{il}) - \pi_{i13} \\ \pi_{i31}(I_v^{il} + v_{il}) - \pi_{i21} & \pi_{i32}(I_v^{il} + v_{il}) - \pi_{i22} & \pi_{i33}(I_v^{il} + v_{il}) - \pi_{i23} \end{bmatrix} \tag{A.5}$$

$$b_i^l = \begin{bmatrix} \pi_{i11}x_l + \pi_{i12}y_l + \pi_{i13}z_l + \pi_{i14} - \alpha(I_u^{il} + u_{il}) \\ \pi_{i21}x_l + \pi_{i22}y_l + \pi_{i23}z_l + \pi_{i24} - \alpha(I_v^{il} + v_{il}) \end{bmatrix} \tag{A.6}$$

where $\alpha = \pi_{i31}x_l + \pi_{i32}y_l + \pi_{i33}z_l + \pi_{i34}$ and the 3×4 projection matrix at camera c_i is

$$\pi_i = \begin{bmatrix} \pi_{i11} & \pi_{i12} & \pi_{i13} & \pi_{i14} \\ \pi_{i21} & \pi_{i22} & \pi_{i23} & \pi_{i24} \\ \pi_{i31} & \pi_{i23} & \pi_{i33} & \pi_{i34} \end{bmatrix} \tag{A.7}$$

A.1 Smooth Surfaces

According to 6.10 the potential function is $\Phi(\mathrm{G}_k + \mathrm{dG}_k) = \|\mathbf{L}^2(\mathrm{G}_k + \mathrm{dG}_k)\|_2^2$ and the minimization becomes a least squares problem of the form:

$$\begin{aligned} \mathrm{dG}_k^* &= \arg\min_{\mathrm{dG}_k} \|\boldsymbol{A}\cdot \mathrm{dG}_k - \boldsymbol{b}\|_2^2 + \gamma \|\mathbf{L}^2(\mathrm{G}_k + \mathrm{dG}_k)\|_2^2 \\ &= \arg\min_{\mathrm{dG}_k} \|\boldsymbol{A}^{'}\cdot \mathrm{dG}_k - \boldsymbol{b}^{'}\|_2^2 \end{aligned} \tag{A.8}$$

The matrix $\boldsymbol{A}^{'}$ is of size $\sum_{i=1}^{N_C} 2\mathrm{Vrt}_i + 3\mathcal{V}$ rows and $3\mathcal{V}$ columns, the vector $\boldsymbol{b}^{'}$ is $(\sum_{i=1}^{N_C} 2\mathrm{Vrt}_i + 3\mathcal{V}) \times 1$

and they write as: $\boldsymbol{A}^{'} = \begin{bmatrix} \boldsymbol{A} \\ \boldsymbol{A}_{\mathrm{reg}} \end{bmatrix}$ $\boldsymbol{b}^{'} = \begin{bmatrix} \boldsymbol{b} \\ \boldsymbol{b}_{\mathrm{reg}} \end{bmatrix}$ where $\boldsymbol{A}_{\mathrm{reg}}$ and $\boldsymbol{b}_{\mathrm{reg}}$

correspond to the bilaplacian regularization. The dimension of $\boldsymbol{A}_{\mathrm{reg}}$ is $3\mathcal{V} \times 3\mathcal{V}$ and has the form:

$$\boldsymbol{A}_{\mathrm{reg}} = \gamma \left[\begin{array}{ccc|ccc|c|ccc} \mathrm{L}_{1,1}^2 & \emptyset & \emptyset & \mathrm{L}_{1,2}^2 & \emptyset & \emptyset & \cdots & \mathrm{L}_{1,\mathcal{V}}^2 & \emptyset & \emptyset \\ \emptyset & \mathrm{L}_{1,1}^2 & \emptyset & \emptyset & \mathrm{L}_{1,2}^2 & \emptyset & \cdots & \emptyset & \mathrm{L}_{1,\mathcal{V}}^2 & \emptyset \\ \emptyset & \emptyset & \mathrm{L}_{1,1}^2 & \emptyset & \emptyset & \mathrm{L}_{1,2}^2 & \cdots & \emptyset & \emptyset & \mathrm{L}_{1,\mathcal{V}}^2 \\ & \ddots & & & \ddots & & \ddots & & \ddots & \\ \mathrm{L}_{\mathcal{V},1}^2 & \emptyset & \emptyset & \mathrm{L}_{\mathcal{V},2}^2 & \emptyset & \emptyset & \cdots & \mathrm{L}_{\mathcal{V},\mathcal{V}}^2 & \emptyset & \emptyset \\ \emptyset & \mathrm{L}_{\mathcal{V},1}^2 & \emptyset & \emptyset & \mathrm{L}_{\mathcal{V},2}^2 & \emptyset & \cdots & \emptyset & \mathrm{L}_{\mathcal{V},\mathcal{V}}^2 & \emptyset \\ \emptyset & \emptyset & \mathrm{L}_{\mathcal{V},1}^2 & \emptyset & \emptyset & \mathrm{L}_{\mathcal{V},2}^2 & \cdots & \emptyset & \emptyset & \mathrm{L}_{\mathcal{V},\mathcal{V}}^2 \end{array}\right] \tag{A.9}$$

$\boldsymbol{b}_{\mathrm{reg}}$ is the vector of dimension $3\mathcal{V} \times 1$ and writes as:

$$\boldsymbol{b}_{\mathrm{reg}} = -\gamma \begin{bmatrix} \mathrm{L}_{1,1}^2 x_1 + \mathrm{L}_{1,2}^2 x_2 \cdots + \mathrm{L}_{1,\mathcal{V}}^2 x_{\mathcal{V}} \\ \mathrm{L}_{1,1}^2 y_1 + \mathrm{L}_{1,2}^2 y_2 \cdots + \mathrm{L}_{1,\mathcal{V}}^2 y_{\mathcal{V}} \\ \mathrm{L}_{1,1}^2 z_1 + \mathrm{L}_{1,2}^2 z_2 \cdots + \mathrm{L}_{1,\mathcal{V}}^2 z_{\mathcal{V}} \\ \vdots \\ \mathrm{L}_{\mathcal{V},1}^2 x_1 + \mathrm{L}_{\mathcal{V},2}^2 x_2 \cdots + \mathrm{L}_{\mathcal{V},\mathcal{V}}^2 x_{\mathcal{V}} \\ \mathrm{L}_{\mathcal{V},1}^2 y_1 + \mathrm{L}_{\mathcal{V},2}^2 y_2 \cdots + \mathrm{L}_{\mathcal{V},\mathcal{V}}^2 y_{\mathcal{V}} \\ \mathrm{L}_{\mathcal{V},1}^2 z_1 + \mathrm{L}_{\mathcal{V},2}^2 z_2 \cdots + \mathrm{L}_{\mathcal{V},\mathcal{V}}^2 z_{\mathcal{V}} \end{bmatrix} \tag{A.10}$$

where $\boldsymbol{b}_{\mathrm{reg}}$

A.2 Surfaces with Discontinuities

Let dU_k be the auxiliary variable of the 3D motion field and U_k the corresponding auxiliary geometry defined as $\mathrm{U}_k = \mathrm{G}_k + \mathrm{dU}_k$. The minimization problem is reformulated as follows:

$$\{\mathrm{dU}_k^*, \mathrm{G}_k^*\} = \arg\min_{\mathrm{dU}_k, \mathrm{G}_k} \|\boldsymbol{A}\cdot \mathrm{dU}_k - \boldsymbol{b}\|_2^2 + \frac{1}{2\theta}\|\mathrm{U}_k - \mathrm{G}_k\|_2^2 + \gamma\|\mathbf{G}\mathrm{G}_k\|_1 \tag{A.11}$$

An alternating coordinate scheme is followed and the first step, given the current estimate of the geometry G_k and through a re-factorization of the terms, boils down to a least squares problem,expressed as:

$$\begin{aligned} \mathrm{dU}_k^* &= \mathrm{argmin}_{\mathrm{dU}_k} \|\boldsymbol{A} \cdot \mathrm{dU}_k - \boldsymbol{b}\|_2^2 + \frac{1}{2\theta} \|\mathrm{U}_k - \mathrm{G}_k\|_2^2 \\ &= \mathrm{argmin}_{\mathrm{dU}_k} \|\boldsymbol{A}^{'} \cdot \mathrm{dU}_k - \boldsymbol{b}^{'}\|_2^2 \end{aligned} \tag{A.12}$$

The matrix $\boldsymbol{A}^{'}$ is of size $\sum_{i=1}^{N_C} 2\mathrm{Vrt}_i + 3\mathcal{V}$ rows and $3\mathcal{V}$ columns, the vector $\boldsymbol{b}^{'}$ is $(\sum_{i=1}^{N_C} 2\mathrm{Vrt}_i + 3\mathcal{V}) \times 1$

and they write as: $\boldsymbol{A}^{'} = \begin{bmatrix} \boldsymbol{A} \\ \frac{1}{2\theta}\mathrm{Id} \end{bmatrix}$ $\boldsymbol{b}^{'} = \begin{bmatrix} \boldsymbol{b} \\ \emptyset \end{bmatrix}$

The term that corresponds to the auxiliary variable $\frac{1}{2\theta}\|\mathrm{U}_k - \mathrm{G}_k\|_2^2$ boils down to $\frac{1}{2\theta}\|\mathrm{dU}_k\|_2^2$. It regularizes the auxiliary variable of the motion field in the cases of ill-posed behavior.

Bibliography

Edward H. Adelson and James R. Bergen. The plenoptic function and the elements of early vision. In *Computational Models of Visual Processing*, pages 3–20. MIT Press, 1991.

Benjamin Allain, Jean-Sébastien Franco, and Edmond Boyer. An Efficient Volumetric Framework for Shape Tracking. In *CVPR 2015 - IEEE International Conference on Computer Vision and Pattern Recognition*, Boston, United States, June 2015. URL https://hal.inria.fr/hal-01141207.

Cedric Allene, Jean-Philippe Pons, and Renaud Keriven. Seamless image-based texture atlases using multi-band blending. *ICPR*, pages 1–4, December 2008. ISSN 1051-4651. doi: 10.1109/ICPR.2008.4761913. URL http://ieeexplore.ieee.org/lpdocs/epic03/wrapper.htm?arnumber=4761913.

Harry C. Andrews and B. R. Hunt. *Digital Image Restoration*. Prentice Hall Professional Technical Reference, 1977. ISBN 0132142139.

Mathieu Aubry, Kalin Kolev, Bastian Goldluecke, and Daniel Cremers. Decoupling photometry and geometry in dense variational camera calibration. In *2011 International Conference on Computer Vision*, pages 1411–1418. IEEE, 2011.

S. Baker and T. Kanade. Limits on super-resolution and how to break them. *IEEE PAMI*, 24(9): 1167–1183, September 2002. ISSN 0162-8828. doi: 10.1109/TPAMI.2002.1033210. URL http://ieeexplore.ieee.org/lpdocs/epic03/wrapper.htm?arnumber=1033210.

Stefano Berretti, Alberto Del Bimbo, and Pietro Pala. Superfaces: A super-resolution model for 3d faces. In *Computer Vision–ECCV 2012. Workshops and Demonstrations*, pages 73–82. Springer, 2012.

Marc Berthod, Hassan Shekarforoush, Michael Werman, and Josiane Zerubia. Reconstruction of high resolution 3d visual information. In *Computer Vision and Pattern Recognition, 1994. Proceedings CVPR'94., 1994 IEEE Computer Society Conference on*, pages 654–657. IEEE, 1994.

Arnav V Bhavsar and AN Rajagopalan. Resolution enhancement in multi-image stereo. *IEEE transactions on pattern analysis and machine intelligence*, 32(9):1721–1728, 2010.

Volker Blanz and Thomas Vetter. A morphable model for the synthesis of 3d faces. In *Proceedings of the 26th Annual Conference on Computer Graphics and Interactive Techniques*, SIGGRAPH

'99, pages 187–194, New York, NY, USA, 1999. ACM Press/Addison-Wesley Publishing Co. ISBN 0-201-48560-5. doi: 10.1145/311535.311556. URL http://dx.doi.org/10.1145/311535.311556.

S. Borman and R.L. Stevenson. Super-resolution from image sequences-a review. In *Circuits and Systems, 1998. Proceedings. 1998 Midwest Symposium on*, pages 374–378, Aug 1998. doi: 10.1109/MWSCAS.1998.759509.

Chris Buehler, Michael Bosse, Leonard McMillan, Steven J. Gortler, and Michael F. Cohen. Unstructured lumigraph rendering. In *SIGGRAPH*, pages 425–432, 2001. URL http://dblp.uni-trier.de/db/conf/siggraph/siggraph2001.html#BuehlerBMGC01.

Peter J Burt and Edward H Adelson. A multiresolution spline with application to image mosaics. *ACM Transactions on Graphics (TOG)*, 2(4):217–236, 1983.

Randi Cabezas, Oren Freifeld, Guy Rosman, and John W Fisher. Aerial reconstructions via probabilistic data fusion. In *2014 IEEE Conference on Computer Vision and Pattern Recognition*, pages 4010–4017. IEEE, 2014.

Cedric Cagniart, Edmond Boyer, and Slobodan Ilic. Probabilistic deformable surface tracking from multiple videos. In Kostas Daniilidis, Petros Maragos, and Nikos Paragios, editors, *ECCV*, volume 6314 of *Lecture Notes in Computer Science*, pages 326–339. Springer, 2010. ISBN 978-3-642-15560-4. URL http://dblp.uni-trier.de/db/conf/eccv/eccv2010-4.html#CagniartBI10.

David Capel and Andrew Zisserman. Super-resolution from multiple views using learnt image models. In *Computer Vision and Pattern Recognition, 2001. CVPR 2001. Proceedings of the 2001 IEEE Computer Society Conference on*, volume 2, pages II–627. IEEE, 2001.

Dan Casas, Marco Volino, John Collomosse, and Adrian Hilton. 4d video textures for interactive character appearance. In *Computer Graphics Forum*, volume 33, pages 371–380. Wiley Online Library, 2014.

Dan Casas, Christian Richardt, John Collomosse, Christian Theobalt, and Adrian Hilton. 4d model flow: Precomputed appearance alignment for real-time 4d video interpolation. In *Computer Graphics Forum*, volume 34, pages 173–182. Wiley Online Library, 2015.

Antonin Chambolle. An algorithm for total variation minimization and applications. *J. Math. Imaging Vis.*, 20(1-2):89–97, January 2004. ISSN 0924-9907. doi: 10.1023/B:JMIV.0000011325.36760.1e. URL http://dx.doi.org/10.1023/B:JMIV.0000011325.36760.1e.

Subhasis Chaudhuri. *Super-resolution imaging*. Springer Science & Business Media, 2001.

Peter Cheeseman, Bob Kanefsky, Richard Kraft, John Stutz, and Robin Hanson. Super-resolved surface reconstruction from multiple images. In *Maximum Entropy and Bayesian Methods*, pages 293–308. Springer, 1996.

Shenchang Eric Chen and Lance Williams. View interpolation for image synthesis. In *Proceedings of the 20th annual conference on Computer graphics and interactive techniques*, pages 279–288. ACM, 1993.

Tao Chen, Jun-Yan Zhu, Ariel Shamir, and Shi-Min Hu. Motion-aware gradient domain video composition. *IEEE Transactions on Image Processing*, 22(7):2532–2544, 2013. URL http://dblp.uni-trier.de/db/journals/tip/tip22.html#ChenZSH13.

Dana Cobza, Keith Yerex, and Martin Jägersand. Dynamic Textures for Image-based Rendering of Fine-Scale 3D Structure and Animation of Non-rigid Motion. 21(3), 2002.

Dana Cobzas and Martin Jägersand. Tracking and rendering using dynamic textures on geometric structure from motion. In Anders Heyden, Gunnar Sparr, Mads Nielsen, and Peter Johansen, editors, *ECCV*, volume 2351 of *Lecture Notes in Computer Science*, pages 415–432. Springer, 2002. ISBN 3-540-43744-4. URL http://dblp.uni-trier.de/db/conf/eccv/eccv2002-2.html#CobzasJ02.

Dana Cobzaş, Keith Yerex, and Martin Jägersand. Dynamic textures for image-based rendering of fine-scale 3d structure and animation of non-rigid motion. In *Computer Graphics Forum*, volume 21, pages 493–502. Wiley Online Library, 2002.

Patrick L. Combettes and Jean-Christophe Pesquet. Proximal Splitting Methods in Signal Processing. In *Fixed-Point Algorithms for Inverse Problems in Science and Engineering*, pages 185–212. Springer, 2011. ISBN 978-1-4419-9568-1. doi: 10.1007/978-1-4419-9569-8. URL http://hal.inria.fr/hal-00643807.

Timothy F. Cootes, Gareth J. Edwards, and Christopher J. Taylor. Active appearance models. *IEEE Trans. Pattern Anal. Mach. Intell.*, 23(6):681–685, June 2001. ISSN 0162-8828. doi: 10.1109/34.927467. URL http://dx.doi.org/10.1109/34.927467.

R.E. Crochiere and L. Rabiner. Interpolation and decimation of digital signals ;a tutorial review. *Proceedings of the IEEE*, 69(3):300–331, March 1981. ISSN 0018-9219. doi: 10.1109/PROC.1981.11969.

Paul E. Debevec, Camillo J. Taylor, and Jitendra Malik. Modeling and rendering architecture from photographs: a hybrid geometry- and image-based approach. In *SIGGRAPH*, SIGGRAPH '96, pages 11–20, New York, NY, USA, 1996. ACM. ISBN 0-89791-746-4. doi: 10.1145/237170.237191. URL http://doi.acm.org/10.1145/237170.237191.

Amaël Delaunoy and Marc Pollefeys. Photometric bundle adjustment for dense multi-view 3d modeling. In *2014 IEEE Conference on Computer Vision and Pattern Recognition*, pages 1486–1493. IEEE, 2014.

M. Eisemann, B. De Decker, M. Magnor, P. Bekaert, E. de Aguiar, N. Ahmed, C. Theobalt, and A. Sellent. Floating Textures. *Comp. Graph. Forum*, 27(2):409–418, April 2008. ISSN 0167-7055. doi: 10.1111/j.1467-8659.2008.01138.x. URL http://doi.wiley.com/10.1111/j.1467-8659.2008.01138.x.

S. Farsiu, M. D. Robinson, M. Elad, and P. Milanfar. Fast and robust multiframe super resolution. *IEEE Transactions on Image Processing*, 13(10):1327–1344, Oct 2004. ISSN 1057-7149. doi: 10.1109/TIP.2004.834669.

Sina Farsiu, Dirk Robinson, Michael Elad, and Peyman Milanfar. Robust shift and add approach to superresolution. In *Optical Science and Technology, SPIE's 48th Annual Meeting*, pages 121–130. International Society for Optics and Photonics, 2003.

David J Field. Relations between the statistics of natural images and the response properties of cortical cells. *JOSA A*, 4(12):2379–2394, 1987.

Jean-Sébastien Franco and Edmond Boyer. Efficient polyhedral modeling from silhouettes. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 31(3):414–427, 2009.

Rik Fransens, Christoph Strecha, and Luc Van Gool. Optical flow based super-resolution: A probabilistic approach. *CVIU*, 106(1):106–115, 2007. URL http://dblp.uni-trier.de/db/journals/cviu/cviu106.html#FransensSG07.

William T. Freeman, Egon C. Pasztor, and Owen T. Carmichael. Learning low-level vision. *International Journal of Computer Vision*, 40(1):25–47. ISSN 1573-1405. doi: 10.1023/A:1026501619075. URL http://dx.doi.org/10.1023/A:1026501619075.

Yasutaka Furukawa and Jean Ponce. Accurate, dense, and robust multiview stereopsis. *IEEE transactions on pattern analysis and machine intelligence*, 32(8):1362–1376, 2010.

Daniel Glasner, Shai Bagon, and Michal Irani. Super-resolution from a single image. In *Computer Vision, 2009 IEEE 12th International Conference on*, pages 349–356. IEEE, 2009.

B. Goldluecke and D. Cremers. A superresolution framework for high-accuracy multiview reconstruction. In *Pattern Recognition (Proc. DAGM)*, Jena, Germany, 2009a.

B. Goldluecke, M. Aubry, K. Kolev, and D. Cremers. A super-resolution framework for high-accuracy multiview reconstruction. *IJCV*, 2013. To appear.

Bastian Goldluecke and Daniel Cremers. Superresolution texture maps for multiview reconstruction. In *ICCV*, volume 9, pages 1677–1684. Citeseer, 2009b.

T. S. Huang and R. Y. Tsay. Multiple frame image restoration and registration. In *Advances in Computer Vision and Image Processing*, volume 1, pages 317–339, Greenwich, 1984. JAI.

Martin Jagersand, Neil Birkbeck, and Dana Cobzas. View dependent texturing using a linear basis. In *Image and Geometry Processing for 3-D Cinematography*, pages 285–305. Springer, 2010.

Zsolt Janko and Jean-Philippe Pons. Spatio-temporal image-based texture atlases for dynamic 3-D models. In *IEEE 3DIM*, pages 1646–1653, Kyoto, Japon, October 2009. IEEE.

Tian Jing and Ma Kai-Kuang. A survey on super-resolution imaging. *Signal, Image and Video Processing*, 5(3):329–342, 2011. ISSN 1863-1711. doi: 10.1007/s11760-010-0204-6. URL http://dx.doi.org/10.1007/s11760-010-0204-6.

Aggelos K. Katsaggelos, Rafael Molina, and Javier Mateos. Super resolution of images and video.

Aggelos Konstantinos Katsaggelos. *Digital Image Restoration*. Springer-Verlag New York, Inc., Secaucus, NJ, USA, 1991. ISBN 0387532927.

Kwang In Kim and Younghee Kwon. Single-image super-resolution using sparse regression and natural image prior. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 32(6): 1127–1133, 2010.

Kazuto Kimura, Takayuki Nagai, Hiroto Nagayoshi, and Hiroshi Sako. Simultaneous estimation of super-resolved image and 3d information using multiple stereo-pair images. In *2007 IEEE International Conference on Image Processing*, volume 5, pages V–417. IEEE, 2007.

Stephane Laveau and Olivier D Faugeras. 3-d scene representation as a collection of images. In *Pattern Recognition, 1994. Vol. 1-Conference A: Computer Vision & Image Processing., Proceedings of the 12th IAPR International Conference on*, volume 1, pages 689–691. IEEE, 1994.

Victor S. Lempitsky and Denis V. Ivanov. Seamless mosaicing of image-based texture maps. In *CVPR*. IEEE Computer Society, 2007. URL http://dblp.uni-trier.de/db/conf/cvpr/cvpr2007.html#LempitskyI07.

Jed Lengyel. The convergence of graphics and vision. *Computer*, 31(7):46–53, 1998.

Hendrik P. A. Lensch, Wolfgang Heidrich, and Hans-Peter Seidel. A silhouette-based algorithm for texture registration and stitching. *Graphical Models*, 63(4):245–262, 2001. URL http://dblp.uni-trier.de/db/journals/cvgip/cvgip63.html#LenschHS01.

Z. Lin, J. He, X. Tang, and C. K. Tang. Limits of learning-based superresolution algorithms. In *Computer Vision, 2007. ICCV 2007. IEEE 11th International Conference on*, pages 1–8, Oct 2007. doi: 10.1109/ICCV.2007.4409063.

Zhouchen Lin and Heung-Yeung Shum. Fundamental limits of reconstruction-based superresolution algorithms under local translation. *IEEE PAMI*, 26(1):83–97, January 2004. ISSN 0162-8828. doi: 10.1109/TPAMI.2004.1261081. URL http://dx.doi.org/10.1109/TPAMI.2004.1261081.

Christian Linz, Christian Lipski, and Marcus Magnor. Multi-image interpolation based on graph-cuts and symmetric optical flow, August 2010. SIGGRAPH '10: ACM SIGGRAPH 2010 Posters.

Ce Liu. *Beyond Pixels: Exploring New Representations and Applications for Motion Analysis*. PhD thesis, Massachusetts Institute of Technology, May 2009.

Ce Liu and Deqing Sun. A Bayesian approach to adaptive video super resolution. *CVPR*, pages 209–216, June 2011. doi: 10.1109/CVPR.2011.5995614. URL http://ieeexplore.ieee.org/lpdocs/epic03/wrapper.htm?arnumber=5995614.

Dhruv Mahajan, Fu-Chung Huang, Wojciech Matusik, Ravi Ramamoorthi, and Peter N. Belhumeur. Moving gradients: a path-based method for plausible image interpolation. *ACM Trans. Graph.*, 28(3), 2009. URL http://dblp.uni-trier.de/db/journals/tog/tog28.html#MahajanHMRB09.

William R Mark, Leonard McMillan, and Gary Bishop. Post-rendering 3d warping. In *Proceedings of the 1997 symposium on Interactive 3D graphics*, pages 7–ff. ACM, 1997.

Peyman Milanfar. *Super-resolution imaging*. CRC Press, 2010.

Robin D Morris, Peter Cheeseman, Vadim N Smelyanskiy, and David A Maluf. A bayesian approach to high resolution 3d surface reconstruction from multiple images. In *Higher-Order Statistics, 1999. Proceedings of the IEEE Signal Processing Workshop on*, pages 140–143. IEEE, 1999.

Robin D Morris, Udo Von Toussaint, and Peter C Cheeseman. High resolution surface geometry and albedo by combining laser altimetry and visible images. *INTERNATIONAL ARCHIVES OF PHOTOGRAMMETRY REMOTE SENSING AND SPATIAL INFORMATION SCIENCES*, 34(3/W4):105–112, 2001.

Ko Nishino, Yoichi Sato, and Katsushi Ikeuchi. Eigen-texture method: Appearance compression and synthesis based on a 3d model. *IEEE Trans. Pattern Anal. Mach. Intell.*, 23(11):1257–1265, 2001. URL http://dblp.uni-trier.de/db/journals/pami/pami23.html#NishinoSI01.

Haesol Park, Kyoung Mu Lee, and Sang Uk Lee. Combining multi-view stereo and super resolution in a unified framework. In *Signal & Information Processing Association Annual Summit and Conference (APSIPA ASC), 2012 Asia-Pacific*, pages 1–4. IEEE, 2012.

Sung Cheol Park, Min Kyu Park, and Moon Gi Kang. Super-resolution image reconstruction: a technical overview. *Signal Processing Magazine, IEEE*, 20(3):21–36, 2003.

Patrick Pérez, Michel Gangnet, and Andrew Blake. Poisson image editing. *ACM Trans. Graph.*, 22(3):313–318, July 2003. ISSN 0730-0301. doi: 10.1145/882262.882269. URL http://doi.acm.org/10.1145/882262.882269.

L. C. Pickup, D. P. Capel, S. J. Roberts, and A. Zisserman. Bayesian methods for image super-resolution. *The Computer Journal*, 2007.

Lyndsey C Pickup, David P Capel, Stephen J Roberts, and Andrew Zisserman. Bayesian Image Super-resolution , Continued. (Ml).

Sergi Pujades, Frédéric Devernay, and Bastian Goldluecke. Bayesian view synthesis and image-based rendering principles. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3906–3913, 2014.

Deepu Rajan, Subhasis Chaudhuri, and Manjunath V Joshi. Multi-objective super resolution: concepts and examples. *IEEE Signal Processing Magazine*, 20(3):49–61, 2003.

Dirk Robinson and Peyman Milanfar. Statistical performance analysis of superresolution image reconstruction. In *Signals, Systems and Computers, 2004. Conference Record of the Thirty-Eighth Asilomar Conference on*, volume 1, pages 144–149. IEEE, 2004.

Stefan Roth and Michael J Black. Fields of experts. *International Journal of Computer Vision*, 82(2):205–229, 2009.

I.J Schoenberg. Cardinal interpolation and spline functions: Ii interpolation of data of power growth. *Journal of Approximation Theory*, 6(4):404 – 420, 1972. ISSN 0021-9045. doi: http://dx.doi.org/10.1016/0021-9045(72)90048-2. URL http://www.sciencedirect.com/science/article/pii/0021904572900482.

Steven M. Seitz and Charles R. Dyer. View morphing. In *Proceedings of the 23rd Annual Conference on Computer Graphics and Interactive Techniques*, SIGGRAPH '96, pages 21–30, New York, NY, USA, 1996. ACM. ISBN 0-89791-746-4. doi: 10.1145/237170.237196. URL http://doi.acm.org/10.1145/237170.237196.

Steven M Seitz, James Diebel, Daniel Scharstein, and Richard Szeliski. A Comparison and Evaluation of Multi-View Stereo Reconstruction Algorithms. 2001.

Steven M Seitz, Brian Curless, James Diebel, Daniel Scharstein, and Richard Szeliski. A comparison and evaluation of multi-view stereo reconstruction algorithms. In *2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'06)*, volume 1, pages 519–528. IEEE, 2006.

Hee Seok Lee and Kuoung Mu Lee. Simultaneous super-resolution of depth and images using a single camera. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 281–288, 2013.

Jonathan Shade, Steven Gortler, Li-wei He, and Richard Szeliski. Layered depth images. In *Proceedings of the 25th annual conference on Computer graphics and interactive techniques*, pages 231–242. ACM, 1998.

Alla Sheffer, Bruno Lévy, Maxim Mogilnitsky, and Alexander Bogomyakov. Abf++: fast and robust angle based flattening. *ACM Transactions on Graphics (TOG)*, 24(2):311–330, 2005.

Hassan Shekarforoush, Marc Berthod, and Josiane Zerubia. 3d super-resolution using generalized sampling expansion. In *Image Processing, 1995. Proceedings., International Conference on*, volume 2, pages 300–303. IEEE, 1995.

Heung-yeung Shum and Sing Bing Kang. A survey of image-based rendering techniques. In *In Videometrics, SPIE*. Citeseer, 1999.

Javier Sánchez Pérez, Enric Meinhardt-Llopis, and Gabriele Facciolo. TV-L1 Optical Flow Estimation. *Image Processing On Line*, 3:137–150, 2013. doi: 10.5201/ipol.2013.26.

Takeshi Takai, Adrian Hilton, and Takashi Mastuyama. Harmonised texture mapping. In *3DPVT*, 2010.

Hiroyuki Takeda, Sina Farsiu, and Peyman Milanfar. Kernel regression for image processing and reconstruction. *Image Processing, IEEE Transactions on*, 16(2):349–366, 2007.

Christian Theobalt, Naveed Ahmed, Hendrik P. A. Lensch, Marcus A. Magnor, and Hans-Peter Seidel. Seeing people in different light-joint shape, motion, and reflectance capture. *IEEE Trans. Vis. Comput. Graph.*, 13(4):663–674, 2007. URL http://dblp.uni-trier.de/db/journals/tvcg/tvcg13.html#TheobaltALMS07.

Jing Tian and Kai-Kuang Ma. Stochastic super-resolution image reconstruction. *Journal of Visual Communication and Image Representation*, 21(3):232–244, 2010. ISSN 10473203. doi: 10.1016/j.jvcir.2010.01.001. URL http://linkinghub.elsevier.com/retrieve/pii/S1047320310000027.

Brian C. Tom, Nikolas P. Galatsanos, and Aggelos K. Katsaggelos. Reconstruction of a high resolution image from multiple low resolution images. In Subhasis Chaudhuri, editor, *Super-Resolution Imaging*, volume 632 of *The International Series in Engineering and Computer Science*, pages 73–105. Springer US, 2002. ISBN 978-0-7923-7471-8. doi: 10.1007/0-306-47004-7_4. URL http://dx.doi.org/10.1007/0-306-47004-7_4.

RY Tsai and Thomas S Huang. Multiframe image restoration and registration. *Advances in computer vision and Image Processing*, 1(2):317–339, 1984.

Vagia Tsiminaki, Jean-Sébastien Franco, and Edmond Boyer. High Resolution 3D Shape Texture from Multiple Videos. In *CVPR 2014 - IEEE International Conference on Computer Vision and Pattern Recognition*, Columbus, OH, United States, June 2014. URL https://hal.inria.fr/hal-00977755.

Tony Tung. Simultaneous super-resolution and 3D video using graph-cuts. *CVPR*, pages 1–8, June 2008. doi: 10.1109/CVPR.2008.4587703. URL http://ieeexplore.ieee.org/lpdocs/epic03/wrapper.htm?arnumber=4587703.

Matthew Turk and Alex Pentland. Eigenfaces for recognition. *J. Cognitive Neuroscience*, 3 (1):71–86, January 1991a. ISSN 0898-929X. doi: 10.1162/jocn.1991.3.1.71. URL http://dx.doi.org/10.1162/jocn.1991.3.1.71.

Matthew Turk and Alex Pentland. Eigenfaces for recognition. *J. Cognitive Neuroscience*, 3 (1):71–86, January 1991b. ISSN 0898-929X. doi: 10.1162/jocn.1991.3.1.71. URL http://dx.doi.org/10.1162/jocn.1991.3.1.71.

M. Unser, A. Aldroubi, and M. Eden. Enlargement or reduction of digital images with minimum loss of information. *Image Processing, IEEE Transactions on*, 4(3):247–258, Mar 1995. ISSN 1057-7149. doi: 10.1109/83.366474.

Marco Volino, Dan Casas, John Collomosse, and Adrian Hilton. Optimal Representation of Multiple View Video. In *Proceedings of the British Machine Vision Conference*. BMVA Press, 2014.

Zhou Wang, A. C. Bovik, H. R. Sheikh, and E. P. Simoncelli. Image quality assessment: From error visibility to structural similarity. *Trans. Img. Proc.*, 13(4):600–612, April 2004. ISSN 1057-7149. doi: 10.1109/TIP.2003.819861. URL http://dx.doi.org/10.1109/TIP.2003.819861.

Yair Weiss and William T Freeman. What makes a good model of natural images? In *Computer Vision and Pattern Recognition, 2007. CVPR'07. IEEE Conference on*, pages 1–8. IEEE, 2007.

Dong Xu, Hongxin Zhang, Qing Wang, and Hujun Bao. Poisson shape interpolation. In *Proc. of the 2005 ACM Symposium on Solid and Physical Modeling*, 2005.

Takuma Yamaguchi, Hiroshi Kawasaki, Ryo Furukawa, and Toshihiro Nakayama. Super-resolution of multiple moving 3d objects with pixel-based registration. In Hongbin Zha, Rin-ichiro Taniguchi, and Stephen Maybank, editors, *Computer Vision – ACCV 2009*, volume 5996 of *Lecture Notes in Computer Science*, pages 516–526. Springer Berlin / Heidelberg, 2010. ISBN 978-3-642-12296-5.

Jianchao Yang and Thomas Huang. Image super-resolution: Historical overview and future challenges. *Super-resolution imaging*, pages 1–34, 2010.

Haichao Zhang, Jianchao Yang, Yanning Zhang, and Thomas S Huang. Non-local kernel regression for image and video restoration. In *Computer Vision–ECCV 2010*, pages 566–579. Springer, 2010.

Maria Zontak and Michal Irani. Internal statistics of a single natural image. In *Computer Vision and Pattern Recognition (CVPR), 2011 IEEE Conference on*, pages 977–984. IEEE, 2011.