Joint Albedo Estimation and Pose Tracking from Video

Sima Taheri, *Student Member, IEEE,* Aswin Sankaranarayanan, *Member, IEEE,* and Rama Chellappa, *Fellow, IEEE*

Abstract—The albedo of a Lambertian object is a surface property that contributes to an object's appearance under changing illumination. As a signature independent of illumination, the albedo is useful for object recognition. Single image-based albedo estimation algorithms suffer due to shadows and non-Lambertian effects of the image. In this paper, we propose a sequential algorithm to estimate the albedo from a sequence of images of a known 3D object in varying poses and illumination conditions. We first show that by knowing/estimating the pose of the object at each frame of a sequence, the object's albedo can be efficiently estimated using a Kalman filter. We then extend this for the case of unknown pose by simultaneously tracking the pose as well as updating the albedo through a Rao-Blackwellized particle filter. More specifically, the albedo is marginalized from the posterior distribution and estimated analytically using the Kalman filter, while the pose parameters are estimated using importance sampling and by minimizing the projection error of the face onto its spherical harmonic subspace, which results in an illumination-insensitive pose tracking algorithm. Illustrations and experiments are provided to validate the effectiveness of the approach using various synthetic and real sequences followed by applications to unconstrained, video-based face recognition.

Index Terms—Albedo, Pose Tracking, Spherical Harmonics, Sequential Algorithm, Kalman Filter, Rao-Blackwellized Particle Filter, intrinsic image statistics.

1 Introduction

Variations in the visual appearance of an object mostly arise due to changes in illumination and pose [2]. Therefore, understanding the interaction of the objects' surface with irradiated light (illumination) and subsequent imaging with a camera (pose) is important for a wide range of computer vision applications. When a surface exhibits Lambertian reflectance, an illumination-insensitive property of the surface is the *albedo* which is a surface reflectance property that contributes to the object's appearance under changing illumination.

Estimating the reflectance properties of human faces has been of interest for decades [1], [3]–[8]. But the proposed algorithms and the assumptions underlying the development of algorithms are application dependent. Although faces are neither exactly Lambertian nor entirely convex, many algorithms make convex Lambertian assumption for the face. Such assumptions are reasonable for applications where the goal is to find a signature that is independent of illumination for representing and recognizing faces across illumination variations [1], [9]–[11] and *not* to analyze the reflectance field of the face or to render a face [12], [13]. In this paper, we also make the convex Lambertian assumption for faces and explore a robust and computationally efficient method for recovering the albedo of a known 3D facial surface[1] from multiple images or a video. While the discussion of the paper is mainly on estimating the albedo of human face, the proposed algorithm can be applied to generic objects under Lambertian reflectance assumption. However, since we do not explicitly model nonrigid deformations, the objects either should be rigid or only have small nonrigid deformations.

Much of the progress in facial albedo estimation has been achieved using a single image of the object under unknown lighting conditions. However, most of the existing approaches are based on restrictive assumptions on objects and illumination conditions, [1], [5], or are computationally intense due to iterative optimization procedures used for obtaining the solution, [14]. On the other hand, the ability to handle multiple images goes a long way in overcoming the shortcomings of single image-based algorithms due to its inherent advantage of having more information (see Fig. 1 for a motivating example).

Multiple image-based algorithms mostly process the data in the batch mode [7], [15]–[19]. However, recursive processing of a set of images is important especially when a video is being processed. Hence our main focus is on recursive estimation of the albedo from multiple frames in a video. Since the presence of multiple images comes with the possibility of additional variations (e.g. in the pose), efficient fusion of available information over the images is important as it leads to a more accurate and robust estimate of the albedo. The goal

• *Sima Taheri and Rama Chellappa are with the Electrical and Computer Engineering (ECE) department at the University of Maryland, ({taheri,rama}@umiacs.umd.edu).*
Aswin Sankaranarayanan is with the ECE department at Rice University (saswin@rice.edu)

1. An average 3D face model is used.

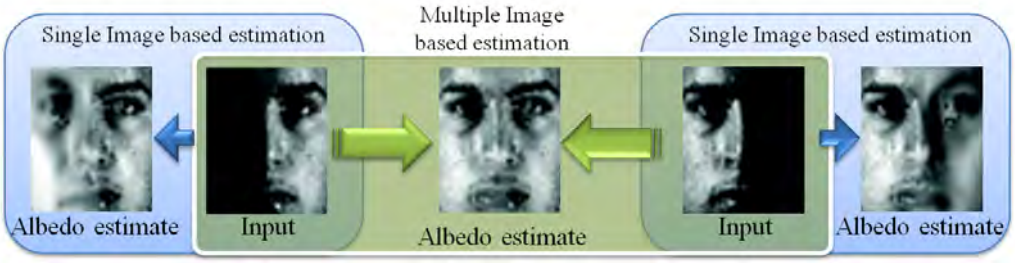


Fig. 1. Benefit of albedo estimation using multiple images. Two views of a face are shown as input images in the left and right columns and their corresponding albedos are estimated using [1]. The albedo maps shown in the right and left columns are noisy and partly based on the mean albedo. The middle column shows the improved estimated albedo map using both views of the face.

of this paper is recursive/sequential albedo estimation for improved pose tracking and recognition of faces modeled as Lambertian objects.

For a Lambertian object in a known pose and illumination, the observed image is linear in its albedo. This allows us to formulate the problem of multi-view albedo estimation as one of Kalman filtering. In particular, the unknown albedo is defined as the static state vector of the Kalman filter. However, since the pose of the face is usually unknown, the albedo estimation step is coupled with that of pose tracking to realize a joint albedo and pose estimation algorithm. We set this problem in a Bayesian inference framework and efficiently solve it using a Rao-Blackwellized particle filter. This allows us to perform efficient analytical inference using a Kalman filter over the albedo state-space and computationally intensive inference using particle filters over a smaller state-space encompassing just the pose parameters.

The joint tracking and albedo estimation approach allows us to build illumination-insensitivity into a tracking algorithm. This is achieved by defining the particle weights using the projection error onto the spherical harmonic subspace of the current observation. We demonstrate the computational and numerical advantages as well as the limitations of our algorithm using several experiments.

It is worth pointing out that the problem studied in this paper has close connections to the photometric stereo problem. Specifically, photometric stereo [20] refers to surface reconstruction of a static scene from multiple images taken under varying illumination. Under appropriate reflectance model (Lambertian [3], specular [21]), the image intensities at each pixel can be expressed in terms of unknown surface parameters (typically, the surface normal) and illumination. This static scene assumption obviates the need for accurate registration, and in many cases, surface estimates can be individually obtained at each pixel. As a result, photometric stereo and its variants (which includes structured lighting) are among the most precise methods for accurate shape recovery. In this paper, we consider the scenario of a *non-static* scene under *changing* illumination. While this puts us beyond the traditional setup of photometric stereo, some of the core concepts in photometric stereo are highly relevant to our problem formulation and solution.

The rest of the paper is organized as follows. We discuss related work in Section 2. The problem of albedo estimation is addressed in Section 3. Subsequently, a Rao-Blackwellized particle filter is proposed in Section 4 for joint pose tracking and albedo estimation. Experimental results are presented in Section 5.

2 Related Work

In this section, we discuss some previous efforts undertaken for recovering the albedo of an object from an intensity image or a sequence of images/video of an object in different poses and illuminations. Since pose estimation is required for video-based albedo recovery, we also discuss some related work on illumination-insensitive pose tracking.

2.1 Albedo estimation

Estimating the facial albedo and the surface shape, as intrinsic factors pertinent to establishing facial identity, has been the focus of computer vision researchers for a long time. While significant efforts have been made to reduce the impact of extrinsic factors such as illumination and pose, the underlying problems still persist and are difficult to solve. We categorize the proposed approaches into single image-based and multiple image-based approaches.

Single image-based approaches: Estimating the albedo, illumination direction and surface normals given a single intensity image is inherently ill-posed. Two approaches, namely shape-from-shading (SFS) approaches and model-based approaches, have been employed to make the problem more tractable.

Shape-from-shading approaches for object shape and albedo estimation make simplifying assumptions such as constant or piecewise constant albedo and known illumination direction [3], [4]. These assumptions are not valid for many real objects and limit the practical applicability of these algorithms. Other SFS algorithms reduce the intractability of general albedo maps and surface normal estimation by using appropriate domain specific constraints, such as symmetry [22], or employing a statistical model for the shape [23]–[25]. In most of these approaches the main goal is shape estimation and albedo is incorporated to completely specify the image formation process.

The Retinex algorithm [26], [27] is one of the first approaches to estimating the lightness of surfaces. This algorithm uses a simplified model of intrinsic image statistics and makes the assumption that image derivatives with a large magnitude are caused by changes in the albedo of the surface, while derivatives with a small magnitude are caused by changes in illumination. Under this model, a shading image can be constructed by calculating the derivatives of the observed image, eliminating derivatives with a large magnitude, then reconstructing the image. However, this simple characterization of shading does not hold for many surfaces and this is the main disadvantage of this algorithm as noted in [28].

Model-based approaches, on the other hand, use the statistical knowledge of the 3D object model which regularizes this problem significantly [5], [29]. Blanz *et al.* [29] recovered the shape and albedo parameters of a 3D morphable model (3DMM) in an analysis-by-synthesis fashion. In order to handle more general lighting conditions, Zhang *et al.* [5] integrated the spherical harmonic illumination representation [2], [9] into the 3DMM approach, by modeling the texture component of the face using spherical harmonic bases. They proposed a feature point-based shape recovery algorithm followed by iterative estimation of albedo and illumination coefficients. However, their method can not handle the harsh lighting conditions due to the limited information that can be extracted from a single image.

To address this problem, Wang *et al.* [14], [30] proposed an optimization algorithm for albedo estimation which is robust to harsh illumination conditions and partial occlusion. By decoupling texture, geometry and illumination and modeling them separately they handle challenging conditions such as cast shadows and saturated regions. Their algorithm works by optimizing the energy function of a Markov Random field over albedo, shape and light resulting in a computationally expensive algorithm that may converge to a local optimum solution.

Biswas *et al.* [1], [31] proposed a stochastic filtering framework for albedo estimation for a frontal face as well as a face with unknown pose. They explicitly accounted for the error in the estimate of surface normals, illumination coefficients and pose to improve the albedo estimate. In their framework, the albedo estimate for pixels corrupted with a large noise is mainly based on the prior albedo and as a consequence it leads to unreliable estimation in noisy situations. Figure 1 shows the albedo estimated using their algorithm on two poorly illuminated faces of a subject. Each face by itself gives a poor estimate of the albedo, which is partly based on the mean albedo. However, estimating the albedo by fusing the information from both images leads to a much more accurate albedo map. This motivates the problem of multi-image or video-based albedo estimation.

Multiple image-based approaches: When an object rotates in front of a camera under distant and varying illumination, the appearance of the object changes both geometrically and photometrically. These changes provide clues to both shape and albedo of the object. Most of the approaches in this category combine multi-view stereo with photometric stereo to find the correspondences across views and subsequently estimate the shape and albedo of the object [7], [15]–[19]. But these algorithms are all performing batch processing and they are computationally demanding [7].

Zhou *et al.* [6] proposed a factorization-based approach to fully recover the albedo and surface normal by imposing a rank, integrability and face symmetry constraints. But the important limitation of this algorithm as well as other multi-image based approaches is that they process all the images in a batch mode. However, it is necessary to develop algorithms that can work in the sequential mode and fuse the estimated parameters in previous frames with the newly available data while accounting for the various sources of error.

Non-Lambertian face modeling: As we mentioned earlier, the Lambertian assumption for faces is valid depending on the application. Recognizing faces across illumination variations using the albedo of the face as a signature independent of illumination has shown promising results [1], [31]. On the other hand, there are other approaches that propose non-Lambertian models for analyzing the apparent bidirectional reflectance distribution function (BRDF) of the face [12], [32]. While these non-Lambertian models are appropriate for computing photo-realistic facial animations as well as face relighting, they only bring slight improvements for face recognition as compared to the results obtained by making the Lambertian assumption [12].

2.2 Illumination-insensitive Pose Tracking

An important challenge in video-based albedo estimation is to find the pose of the face at each frame. Since albedo estimation is often coupled with a 3D shape model, a pose tracking algorithm is required to obtain the 3D configuration of the face at each frame. Several methods have been proposed for 3D face tracking, [8], [33]–[36], however, they are often sensitive to illumination variations.

Cascia *et al.* [35] formulated the tracking problem as an image registration problem in the cylinder's texture map image. To account for lighting variations, they modeled the residual error of registration as a linear combination of texture warping templates and orthogonal illumination templates. Marks *et al.* [36] proposed a generative model and stochastic filtering algorithm for 3D nonrigid object tracking with the aim of addressing the inefficiencies of template matching and optical flow-based algorithms. They combined the advantages of template matching and flow-based algorithms and performed the joint inference of 3D position, orientation and nonrigid deformations.

Xu and Roy-Chowdhury [8] proposed a bilinear space of motion and illumination in which they estimated the

pose and illumination parameters by iterative optimization. They assumed that the illumination in the first frame is uniform and hence the intensity can be used as an estimate of the albedo. This assumption narrows down the domain of videos on which the algorithm can successfully perform tracking. Although explicit illumination modeling makes the tracking algorithm reasonably robust to illumination variations, their algorithm does not adequately model the albedo while tracking the face. The readers are referred to [37] for more detailed discussions on pose tracking.

Contributions: In this paper we address joint video-based albedo estimation and illumination-insensitive pose tracking. Our contributions are as follows:

- When the pose of the object is known, we propose an efficient video-based sequential albedo estimation using the Kalman filter.
- When the pose of the object is unknown, we show that pose and albedo estimation of an object from a video sequence can be performed using a computationally efficient Rao-Blackwellized particle filter.
- And finally, we propose an approach that eliminates the need for recalculating the spherical harmonic bases at each pose by exploiting the physical properties inherent to Lambertian objects.

3 Video-based Albedo Estimation

The key idea behind the sequential albedo estimation framework proposed in this paper revolves around the linear relationship between the image observation and albedo. Under the Lambertian assumption for the face, the intensity reflected by a point p_i on the face, with the surface normal $\mathbf{n}_i$ and the albedo ρ_i, due to the lighting function l coming from direction $\mathbf{u}_l$ is modeled as:

$$I(p_i) = I_i = \rho_i \int l(\mathbf{u}_l) \max(\mathbf{n}_i.\mathbf{u}_l, 0) d\mathbf{u}_l \tag{1}$$

Lambert's cosine law is non-linear due to $\max(\mathbf{n}.\mathbf{u}_l, 0)$ which accounts for the formation of attached shadows. However, in a seminal work, Basri and Jacobs [9] showed that images of a face (specifically, any convex Lambertian object) under varying illumination are closely approximated by a 9-dimensional (linear) subspace using a spherical harmonic decomposition. Let y_{nm} be the spherical harmonic basis of *order* n and *degree* m. Note that spherical harmonic bases are functionals on a sphere, i.e, $\mathsf{y}_{nm} : \mathbb{S}^2 \mapsto \mathbb{R}$. For the rest of the paper, we parametrize $\mathbb{S}^2$ using unit norm vectors in $\mathbb{R}^3$. Any arbitrary lighting function l over a scene can be described as a function in $\mathbb{S}^2$ if the light sources are at infinity (or in practice, sufficiently far away). In such a case, the lighting function l can be described using the spherical harmonic bases as,

$$l(\mathbf{u}_l) = \sum_{n=0}^{\infty} \sum_{m=-n}^{n} l_{nm} \mathsf{y}_{nm}(\mathbf{u}_l) \tag{2}$$

Basri and Jacobs also showed that the Lambertian kernel, $\max(\mathbf{n}.\mathbf{u}_l, 0)$, acts as a smoothing filter on the light source, and the image of the object produced in that lighting condition depends heavily on the lower order spherical harmonic basis elements. Therefore the generated image can be well approximated using just the first 9 basis elements as,

$$\begin{aligned} I_i &\approx \rho_i \sum_{n=0}^{2} \sum_{m=-n}^{n} l_{nm} \alpha_n \mathsf{y}_{nm}(\mathbf{n}_i) \\ &= \rho_i \sum_{n=0}^{2} \sum_{m=-n}^{n} l_{nm} Y_{nm}(\mathbf{n}_i) \end{aligned} \tag{3}$$

where $\{\alpha_n\}$ are the coefficients of the spherical harmonic expansion of the Lambertian kernel. For a given object of d pixels described using a set of albedos $\{\rho_i\}$ and normal vectors $\{\mathbf{n}_i\}$, we can now construct the so called *spherical harmonic basis images* (SHBI) $\{\mathbf{Y}_{nm} \in \mathbb{R}^d | n = 0, 1, 2; m = -n, \ldots, n\}$ such that $\mathbf{Y}_{nm} = \{Y_{nm}(\mathbf{n}_i)\} = (\alpha_n \mathsf{y}_{nm}(\mathbf{n}_1), \alpha_n \mathsf{y}_{nm}(\mathbf{n}_2), \ldots, \alpha_n \mathsf{y}_{nm}(\mathbf{n}_d))^T$.

Using (3), the intensity observed at the ith pixel of the face with a known pose and illumination is written as

$$I_i = \rho_i Y^T(\mathbf{n}_i) L + \nu_i, \tag{4}$$

where $Y(\mathbf{n}_i) = (Y_{00}(\mathbf{n}_i), \ldots, Y_{22}(\mathbf{n}_i))^T \in \mathbb{R}^9$ encodes the pose using spherical harmonic basis values at that pixel and $L = (l_{00}, \ldots, l_{22})^T \in \mathbb{R}^9$ encodes the lighting positions. Moreover, the observation noise, $\nu = \mathcal{N}(0, \Sigma_v)$, is defined to have a multivariate Gaussian distribution and it models the surface deviation from the Lambertian assumption (specularity, cast shadow, and saturated pixels). Then, the intensity vector for the whole face can be expressed as,

$$I = diag(\rho)\mathbf{Y}L + \nu, \tag{5}$$

where $\rho = (\rho_1, \ldots, \rho_d)^T \in \mathbb{R}^d$ and the $d \times 9$ matrix $\mathbf{Y} = [\mathbf{Y_{00}}, \ldots, \mathbf{Y_{22}}]$ encodes the spherical harmonic basis images. Here, the subject intrinsic matrix $B = [diag(\rho)\mathbf{Y}]$ defines an illumination-insensitive subspace of all images of the face under arbitrary illuminations. Given an observation I, we can use the projection error onto this subspace to define an observation model that is illumination-insensitive as well.

For a video of the face where the pose is known, the albedo can be optimally updated over time using the Kalman filter. However before describing the Kalman filter framework for albedo estimation, we discuss how we can avoid recalculating the spherical harmonic basis images at each frame which makes the algorithm computationally efficient.

3.1 Head Orientation vs. Illumination Direction

Representation of the face in (3) relates the 3D structure of the face, the illumination direction and the face albedo to the generated image. This representation is suited for situations where the pose is fixed and only the illumination varies. However, when the pose changes in

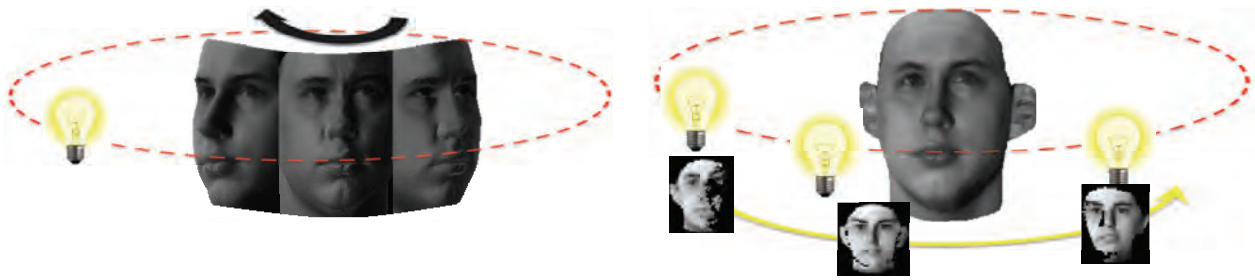


Fig. 2. **(left)** Head rotation in front of a fixed illumination source can be replaced by **(right)** the illumination source rotation in opposite direction around the fixed face along with visibility modeling. As a consequence of the physical properties inherent to Lambertian objects, these two configurations result in the same observed intensities on the face.

a video, the basis images $\{\mathbf{Y_{nm}}\}$ differ from frame to frame as the normal vectors associated with a point of the face change due to rotation. Therefore, to estimate albedo, we need to recalculate the basis images at each frame which is computationally inefficient. Xu and Roy-Chowdhury [8] addressed this problem by proposing a bilinear subspace formulation for joint illumination and motion. They combined the effects of motion, illumination and 3D structure in generating a sequence of images. But their approach still requires computing the bilinear bases at each frame which is time consuming. Here, we resolve this problem by exploiting a property of Lambert's law.

For Lambertian objects, the apparent intensity of a surface patch to an observer depends on the angle between its surface normal, $\mathbf{n}_i$, and the incident illumination direction, $\mathbf{u}_l$, and is independent of each of these directions separately. Rotation of the face changes the direction of $\mathbf{n}_i$, which leads to change of angle between $\mathbf{n}_i$ and $\mathbf{u}_l$ and change of intensity at p_i as a result. But the same change in $\langle \mathbf{n}_i.\mathbf{u}_l \rangle$ is obtained if we keep $\mathbf{n}_i$ fixed and rotate the illumination source by the same magnitude and about the same axis, but in the opposite direction.

When the face is rotated by R, while the illumination is fixed, the intensity of the point p_i changes as

$$I_R(p_i) = \rho_i \sum_n \sum_m l_{nm} Y_{nm}(R(\mathbf{n}_i)) \qquad (6)$$

where $\{Y_{nm}(R(\mathbf{n}_i))\}$ are the basis values in the new pose. Recalculating the basis images for each new pose is computationally expensive. On the other hand, rotation of spherical harmonics using the transformation matrix $D(R)$ has been studied in [2], [38], [39]. The supplementary document provides more details on the transformation matrix and a method to calculate it. Using this idea and (6), the intensity of the point p_i can be written as

$$\begin{aligned} I_R(p_i) &= \rho_i \sum_n \sum_m l_{nm} \sum_{m'} D^n_{mm'}(R) Y_{nm'}(\mathbf{n}_i) \qquad (7) \\ &= \rho_i L^T (D(R) Y(\mathbf{n}_i)) = \rho_i Y(\mathbf{n}_i)^T (D^T(R) L) \end{aligned}$$

where D is the 9×9 spherical harmonics transformation matrix [38]. This expresses the new intensity at p_i in terms of original harmonic basis values, $Y(\mathbf{n}_i)$ and the transformed illumination coefficients. As this equation suggests, the illumination coefficients should be transformed by D^T, which is the inverse of matrix D, to compensate for the head rotation. It should be noted that transforming the 9-dimensional illumination coefficients is more robust than rotating the whole $9 \times d$ harmonic basis matrix. Moreover, this is justified by the Lambertian property, as discussed before.

We can now use the same basis images, $\{\mathbf{Y_{nm}}\}$, for representing the face throughout the image sequence and only compute the new illumination coefficients for each new pose/frame. In this way we avoid recalculating the SHBI at each frame which makes the computation much faster and efficient. However, it should be noted that since rotation of the face makes some pixels disappear, a visibility test needs to be applied in order to remove the non-visible pixels from the current view. Since we have a 3D model for the face, visibility issues are solved easily. Figure (2) illustrates the idea proposed in this section.

3.2 Shape estimation

We calculate the 3D morphable shape model [40] by convex combinations of the shapes of m training examples in the Vetter dataset [41] followed by principal component analysis (PCA) as $s = \bar{s} + Sa$. The columns of S are the most significant eigenvectors s_i rescaled by their standard deviation and the coefficient a constitutes a pose-insensitive low-dimensional coding of a face. We can either use the mean shape, $\bar{s}$, throughout the process or compute a more accurate estimate of the 3D shape using the approach presented in [5].

Registering the 3D shape model to the face in the first frame can be performed using the method proposed by Zhang *et. al* [5]. For a set of pre-selected feature points on the morphable model, we find the corresponding landmarks, s_{img}, on the first frame of the test video[2]. We set the initial coefficient a_0 to zero and register the average shape, $\bar{s}$, to the first frame using the algorithm proposed by [42]. This gives us the initial rotation, translation and scale parameters. We define the shape error at feature points as the difference between s_{img} and the new shape information of feature points in the model that was rendered by the recovered projection

2. We picked fifteen landmarks manually on the face in the first frame, but it can also be performed automatically using face and facial component detection method

parameters. Then the vector of shape parameters, a, can be updated using the method proposed in [5]. We iterate through the shape parameter updating procedure until the amount of update falls below a threshold. Then the final shape parameters are used to get the face 3D shape model. This estimated shape model is used throughout the frames. Figure 3 shows a sample of the face with landmarks on it along with the registered shape with the face texture warped to it.

3.3 Albedo Estimation using the Kalman Filter

As mentioned earlier, there are many single image-based algorithms for robust albedo estimation. However, to obtain an accurate estimate of the albedo from a sequence of images (sequential mode), estimates from individual images must be fused. We use the Kalman filter to fuse the information over time. Though the Kalman filter was originally designed to estimate the state of a time-varying system, it can be used on static processes as well. There are many instances where the Kalman filter has been used in this fashion [43], [44]. In fact, the classic textbook by Maybeck [45] introduces the Kalman filter using a static example. It should be noted that when we apply the Kalman filter to a static process like the albedo map, the state transition model is given as $\rho_t = \rho_{t-1}$ and it is noiseless. We use the Kalman filter to sequentially update the albedo as more information becomes available over time. In such cases, as more observations are introduced, the albedo estimate converges to the true value.

The problem of video-based albedo estimation in a sequential mode can be formulated as follows. Given the estimate of albedo at frame/time $t-1$ characterized by its mean and covariance matrix, $\{\mu_{\rho,t-1}, \Sigma_{\rho,t-1}\}$, we want to update the posterior probability of the albedo $P(\rho|\mathcal{Z}^t, \Theta)$ as a new frame Z_t becomes available. Here, Z_t is the frame at time t, and $\mathcal{Z}^t = \{Z_1, ..., Z_t\}$. The parameter $\Theta = \{\theta_1, \ldots, \theta_t\}$, where θ_t denotes the surface pose at time t and, for now, is assumed to be known. Knowing this pose parameter, θ_t, an inverse warp of the 3D model of the face onto the image Z_t gives us a registered observation at time t as a d-dimensional intensity vector, $I_t = I_t(Z_t, \theta_t)$. Note that, we use a point-cloud model consisting of d points, each with a known normal $\mathbf{n}_i$ and an unknown albedo ρ_i.

Using Bayes' theorem, the posterior probability of the albedo can be sequentially updated as follows

$$P(\rho|\mathcal{Z}^t, \Theta) \propto P(Z_t|\rho, \mathcal{Z}^{t-1}, \Theta) P(\rho|\mathcal{Z}^{t-1}, \Theta) \qquad (8)$$

where the posterior probability at time $t-1$, $P(\rho|\mathcal{Z}^{t-1}, \Theta)$, acts as the prior probability at time t to recursively update $P(\rho|\mathcal{Z}^t, \Theta)$. From (5), the likelihood function $P(Z_t|\rho, \mathcal{Z}^{t-1}, \Theta)$ can be written as

$$P(Z_t|\rho, \mathcal{Z}^{t-1}, \Theta) = P(I_t|\rho, \mathcal{Z}^{t-1}, \Theta) = \mathcal{N}(I_t|H_t\rho, \Sigma_{v,t})$$

where the observation matrix, $H_t = diag(h_t)$, is a $d \times d$ diagonal matrix with entries $h_{ti} = Y^T(\mathbf{n}_i)L_t$ defined for the ith pixel of the face. The illumination vector, L_t, is approximated at each frame using the albedo estimates from previous frames. Finally, the posterior estimate for the albedo at t is given by the following Kalman filter update equations:

$$\mu_{\rho,t} = \mu_{\rho,t-1} + K_t(I_t - H_t\mu_{\rho,t-1}) \qquad (9)$$

$$\Sigma_{\rho,t} = (\mathbb{I} - K_t)\Sigma_{\rho,t-1} \qquad (10)$$

where $\mathbb{I}$ is the identity matrix of size d and the Kalman gain K_t is defined as $K_t = \Sigma_{\rho,t-1}(\Sigma_{\rho,t-1} + \Sigma_{v,t})^{-1}$.

Here, the prior albedo values (for faces), $\{\mu_{\rho,0}, \Sigma_{\rho,0}\}$, are estimated as the mean and covariance of the available training data in the Vetter dataset. Moreover, the initial observation noise covariance matrix, $\Sigma_{v,0}$, is learned using the training data when for each face the mean shape and mean albedo are used. We ignore the correlation among nearby pixels by defining Σ_ρ, Σ_v and therefore the Kalman gain, K, to be diagonal matrices. However, it should be noted that non-diagonal matrices can be used without significantly changing the fusion algorithm.

In order to make the albedo estimate robust against noisy pixels which are due to deviations from the Lambertian assumption, we update the observation noise covariance matrix, $\Sigma_{v,t}$, at each frame by assigning a very large value to those entries (corresponding to the pixels) whose observed intensities are above an upper threshold or below a lower threshold as well as to non-visible pixels. In this way we avoid the saturated pixels, pixels in cast shadows, pixels with specularity and occluded regions to affect the estimated value of the albedo.

As more knowledge about the albedo becomes available through new observations, the uncertainty in the static parameter ρ is updated. Equation (10) shows that the error covariance of the estimated ρ decreases over time (K_t's components are ≤ 1) and since the Kalman filter at each frame gives an unbiased MMSE estimate of ρ, a decrease in the error covariance indicates improvement in the estimated parameter over the time. In the ideal case where each pixel's intensity satisfies the Lambertian property in some frames over the sequence (i.e. not counting saturation or shadows), K_t and $\Sigma_{\rho,t}$ converge to zero and the final albedo estimate, μ_ρ, remains unchanged.

Figure 4 shows the result of applying the Kalman filter for sequential albedo estimation on a synthetic sequence. In this sequence, the head pose is fixed so as to focus on the performance of the Kalman filter for updating the state of a static parameter. It should be noted that while for a multiple image problem in which the pose is fixed (and so the correspondences are known) batch processing algorithms such as photometric stereo can be applied to get an accurate estimate of the object shape and albedo, our emphasis is on developing a sequential approach.

Figure 4 shows some frames of a sequence synthesized using the PIE illumination images [46]. The sequence starts with a face under harsh illumination conditions

Fig. 3. **(left to right)** Fifteen manually picked landmarks on the face, and the rendered 3D face in different views after registering the 3D model to the given face.

Fig. 4. The first row shows some frames of a synthetic sequence obtained from the PIE illumination dataset. The estimated albedo maps using faces up to these frames are shown along with their corresponding Kalman gains in the second and third row, respectively. Shown below the third row are the frame numbers. Note how the Kalman gain converges to zero when all parts of the face gets well lit.

and then the light source rotates in front of the face. The figure also shows the albedo map estimated for the face up to each frame along with the corresponding Kalman gain, K_t, at each frame. The Kalman gain images show how the algorithm assigns large weights to informative pixels in the current frame and reduces the weights of badly illuminated pixels as well as pixels that violate the Lambertian assumption. As the algorithm proceeds through the frames, the albedo estimate improves and finally stabilizes as K_t goes to zero.

4 Pose Tracking and Albedo Estimation

When the pose of the face is unknown, analytical inference of the albedo can still be done when the albedo posterior probability is conditioned by the head pose. This observation motivates the use of the Rao-Blackwellized particle filter (RBPF). Rao-Blackwellization of a particle filter involves splitting the state variables into two sets, such that analytical inference is possible on one set conditioned on the other [47]. Rao-Blackwellization leads to more accurate estimates of state parameters with fewer particles. It has been applied to various problems such as joint rigid and non-rigid face tracking [36] and joint face tracking and head pose estimation [47].

We characterize the head pose as a function of rotation and translation of the head, where the rotation and translation are described using 3-dimensional vectors $\mathbf{r} = \{r_1, r_2, r_3\}$ and $\mathbf{t} = \{t_1, t_2, t_3\}$, respectively. Using a 3D shape model registered to the face in the first frame, the goal is to obtain a trajectory of the pose parameter evolution $\theta_t = \{\mathbf{r}, \mathbf{t}\}_t$, over the frames as well as an estimate for albedo. To this end, at each time instant t, the RBPF uses the hybrid particle set $\{\theta_t^{(i)}, w_t^{(i)}, \mu_{\rho,t}^{(i)}, \Sigma_{\rho,t}^{(i)}\}$ to approximate the posterior $P(\theta_t, \rho|\mathcal{Z}^t)$ over the joint state vector $S_t = \{\theta_t, \rho\}$. Here $\{\theta_t^{(i)}, w_t^{(i)}\}$ approximate the posterior of pose parameters, $P(\theta_t|\mathcal{Z}^t)$, and $\{\mu_{\rho,t}^{(i)}, \Sigma_{\rho,t}^{(i)}\}$ form the analytical estimate (in terms of the mean and covariance of a Gaussian density) of the albedo associated with each pose particle, $\theta_t^{(i)}$, obtained using the Kalman filter described in Section 3.3. The posterior distribution $P(\theta_t, \rho|\mathcal{Z}^t)$ can be written as,

$$P(\theta_t, \rho|\mathcal{Z}^t) \propto P(Z_t|\theta_t, \rho) \times \int_{\theta_{t-1}} \int_{\mu_{\rho,t-1}} P(\theta_t, \rho|\theta_{t-1}, \mu_{\rho,t-1}) P(\theta_{t-1}, \mu_{\rho,t-1}|\mathcal{Z}^{t-1}) \tag{11}$$

By integrating out the albedo part of the state vector, we obtain a marginal filter for pose parameter, θ_t, as follows,

$$P(\theta_t|\mathcal{Z}^t) \propto \int_{\rho} P(Z_t|\theta_t, \rho) \times \int_{\theta_{t-1}} \int_{\mu_{\rho,t-1}} P(\theta_t, \rho|\theta_{t-1}, \mu_{\rho,t-1}) P(\theta_{t-1}, \mu_{\rho,t-1}|\mathcal{Z}^{t-1}) \tag{12}$$

Here, the posterior $P(\theta_{t-1}, \mu_{\rho,t-1}|\mathcal{Z}^{t-1})$ is approximated over the previous joint state by a set of particles $\{\theta_{t-1}^{(i)}, w_{t-1}^{(i)}, \mu_{\rho,t-1}^{(i)}, \Sigma_{\rho,t-1}^{(i)}\}$ as

$$\begin{aligned} P(\theta_{t-1}, \mu_{\rho,t-1}|\mathcal{Z}^{t-1}) &= P(\theta_{t-1}|\mathcal{Z}^{t-1}) P(\mu_{\rho,t-1}|\theta_{t-1}, \mathcal{Z}^{t-1}) \\ &\propto \sum_i w_{t-1}^{(i)} \delta(\theta_{t-1}^{(i)}) \alpha_{t-1}^{(i)}(\mu_{\rho,t-1}) \end{aligned} \tag{13}$$

where $\alpha_{t-1}^{(i)}(\mu_{\rho,t-1})$ is defined as the density on $\mu_{\rho,t-1}$ conditioned on the pose of the ith particle and the measurements $\mathcal{Z}^{t-1}$:

$$\alpha_{t-1}^{(i)}(\mu_{\rho,t-1}) = P(\mu_{\rho,t-1}|\theta_{t-1}^{(i)}, \mathcal{Z}^{t-1}) \tag{14}$$

Substituting (13) into the expression for the marginal filter (12) we obtain the following Monte-Carlo approximation to the exact marginal Bayes filter,

$$P(\theta_t|\mathcal{Z}^t) \propto \sum_i w_{t-1}^{(i)} \int_{\rho} P(Z_t|\theta_t, \rho) \times \int_{\mu_{\rho,t-1}} P(\rho|\theta_t, \theta_{t-1}^{(i)}, \mu_{\rho,t-1}) P(\theta_t|\theta_{t-1}^{(i)}, \mu_{\rho,t-1}) \alpha_{t-1}^{(i)}(\mu_{\rho,t-1}) \tag{15}$$

1- Initial model registration: Fifteen landmark points are manually selected on the face in the first frame using which the initial pose parameters, θ_0, are estimated. This step can be automated using face detection and facial component localization algorithms.

2- Initial parameter setting: We use the mean albedo and its variance over training data as the initial albedo, $\mu_{\rho,0}$, and the initial Kalman filter error covariance matrix, $\Sigma_{\rho,0}$, respectively. The 3DMM is calculated using the training data in the Vetter dataset (we mostly use the mean shape as the 3D model in our experiments). The initial illumination coefficients are obtained using the mean albedo and current observation.

3- RBPF iterations over frames: Starting from the posterior approximation $P(\theta_{t-1}, \rho|\mathcal{Z}^{t-1})$ estimated by a set of N weighted hybrid particles $\{\theta_{t-1}^{(i)}, w_{t-1}^{(i)}, \mu_{\rho,t-1}^{(i)}, \Sigma_{\rho,t-1}^{(i)}\}$, repeat for $j \leq N$:

1) Sample from the dynamic model $P(\theta_t|\theta_{t-1}^{(i)})$ for a chosen $\theta_{t-1}^{(i)}$ to obtain a predicted pose parameter, $\hat{\theta}_t^{(j)}$.
2) Get the observation vector $I_t^{(j)}$ through an inverse warp of the 3D model on the current frame using $\hat{\theta}_t^{(j)}$ as the head pose and then find the intensity at the model vertices.
3) Update $\mu_{\rho,t-1}^{(i)}$ and $\Sigma_{\rho,t-1}^{(i)}$ to $\mu_{\rho,t}^{(j)}$ and $\Sigma_{\rho,t}^{(j)}$ according to the Kalman filter equations (9,10).
4) Calculate the importance weight $w_t^{(j)}$ using (17).

Fig. 5. Algorithm Summary.

This approximation has a complicated form which makes it intractable in general. In theory, it is possible to directly sample from the approximation, but this is both computationally and analytically difficult. Hence, to obtain a practical algorithm we make one additional assumption that the head motion model for pose θ_t does not depend on the albedo value $\mu_{\rho,t-1}$ at time $t-1$, $P(\theta_t|\theta_{t-1}^{(i)}, \mu_{\rho,t-1}) = P(\theta_t|\theta_{t-1}^{(i)})$. So we can now move the motion model out of the integral in (15), yielding

$$P(\theta_t|\mathcal{Z}^t) \propto \sum_i w_{t-1}^{(i)} P(\theta_t|\theta_{t-1}^{(i)}) \int_\rho P(Z_t|\theta_t, \rho) \times \int_{\mu_{\rho,t-1}} P(\rho|\theta_t, \theta_{t-1}^{(i)}, \mu_{\rho,t-1}) \alpha_{t-1}^{(i)}(\mu_{\rho,t-1}) \quad (16)$$

Now we can perform importance sampling in the usual way, using the predictive density $\sum_i w_{t-1}^{(i)} P(\theta_t|\theta_{t-1}^{(i)})$ as the proposal density. We define the observation model of the particle filter as the projection error of the observed intensity vector $I_t^{(i)} = I_t(Z_t, \theta_t^{(i)})$ onto the subspace of spherical harmonic images. This makes the tracking component of the particle filter illumination-insensitive. Toward this end, the importance weights, $\{w_t^{(i)}\}$, are estimated as follows. For each pose particle $\theta_t^{(i)}$, first the spherical harmonic basis images (including albedo) are calculated as $B_t^{(i)} = diag(\mu_{\rho,t}^{(i)})\mathbf{Y}$, where $\mu_{\rho,t}^{(i)}$ is the analytical estimate of albedo obtained using (9). Here, $\mathbf{Y}$ is the $d \times 9$ matrix which, based on the discussion in Section 3.1, is fixed regardless of the head pose parameter $\theta_t^{(i)}$. By leveraging the Gaussian assumption in (5), the importance weights are defined as

$$w_t^{(i)} \propto \exp(-\frac{1}{2}\|I_t^{(i)} - B_t^{(i)} L_t^{(i)}\|_{\Sigma_v}^2), \text{where } L_t^{(i)} = (B_t^{(i)})^\dagger I_t^{(i)} \quad (17)$$

Here, $B_t^{(i)} L_t^{(i)}$ is the projection of the observation vector $I_t^{(i)}$ onto the basis $B_t^{(i)}$, and $(B_t^{(i)})^\dagger$ is the pseudo-inverse. Figure 5 has a summary of the proposed algorithm.

5 Experimental Results

In this section, we present experimental results on joint face tracking and albedo estimation. We report the results on some synthetic sequences generated using 3D Vetter dataset [41] and images in the PIE-illumination dataset [46]. The PIE-illumination dataset has several images of a subject taken under different illumination conditions. These images can be used to generate a synthetic video of the fixed head under desired illumination variations. We also show results on real sequences from the BU dataset [35]. This dataset has various sequences per subject in which significant illumination changes and 2D (in-plane) and 3D (out-of-plane) head rotations exist. The resolution of frames is 320×240 (non-interleaved) and the videos are collected at 30 fps.

To have examples with more extreme lighting conditions, we also collected some sequences of rotating heads in front of a fixed lighting source to evaluate the performance and limitations of our algorithm in situations where the Lambertian assumption for the face is largely violated. It should be noted that the ground truth albedo maps are available for the PIE and Vetter datasets. But for the BU dataset, there are sequences taken under uniform illumination which can be used to obtain the ground truth albedos with scale ambiguity for each subject. We should also mention that we did not perform shape correction in these experiments and just used the mean shape for the faces. We also evaluate the effect of using the mean shape instead of the true shape in the estimated albedo error.

5.1 Albedo Estimation

In this section, the goal is to evaluate the performance of our video-based albedo estimation algorithm using various synthetic and real sequences.

Synthetic sequences: We compare our results with the results of Biswas *et al.* [1] and Zhang *et al.* [5] using 68 synthetic PIE sequences with fixed frontal faces and rotating illumination source around them. To ensure a fair comparison, we apply their algorithms at each

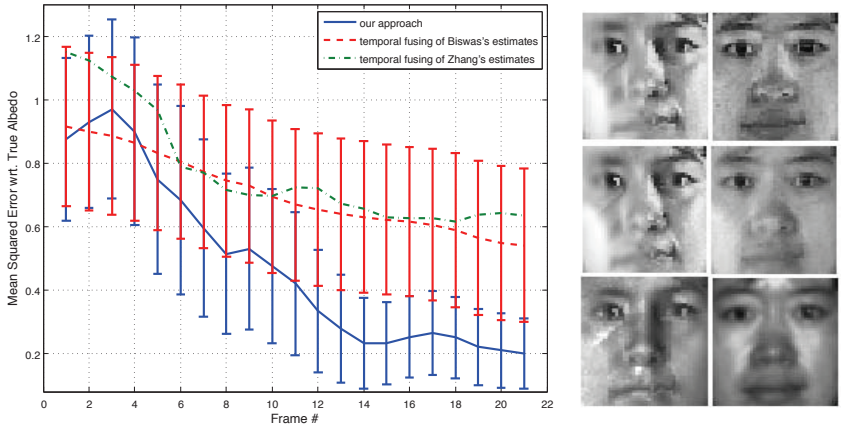


Fig. 6. The plot in the left column shows the MSE of estimated albedo for synthetic PIE sequences averaged over 68 subjects. We compared our results at each frame with those obtained from the temporal fusion of Biswas' [1] and Zhang's [5] estimates up to each frame. The right column shows the estimated albedos for a chosen subject obtained from (from top to bottom) our approach, [1], and [5]. The faces in the left column are the estimates obtained in the first frame and the right column shows the final albedo estimates. Both visualizations indicate the superior performance of the proposed approach.

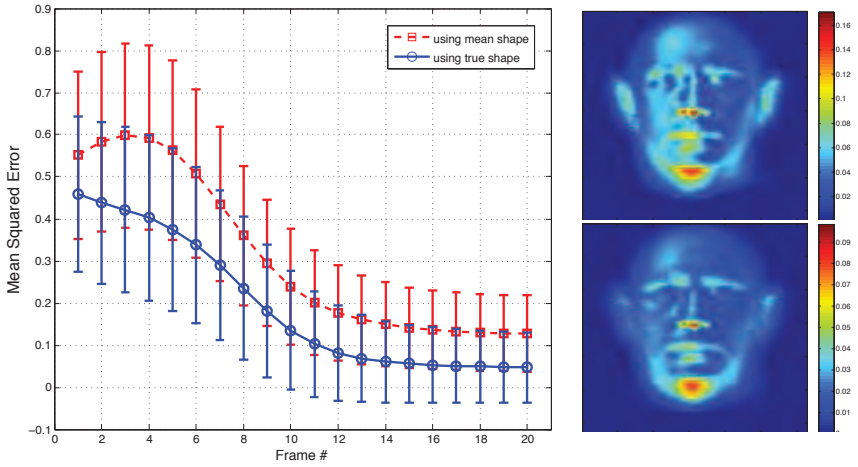


Fig. 7. The effect of using the mean shape instead of the true 3D shape on the estimated albedo error with respect to the ground truth. The estimates are obtained for the synthesized Vetter sequences (with 20 frames, fixed face and rotating illumination source around the face). The mean estimated albedo errors are also shown using color coded images **top:** when the mean shape is used for albedo estimation, and **bottom:** when the true subject specific shapes are used for albedo estimation. The figure is best viewed in color.

frame separately and then fuse the estimated albedos by temporal averaging over the frames (computing the mean of the estimated albedo maps up to each frame). In other words, at each frame we obtain the albedo map by averaging over the estimated albedos form all the previous frames up to (including) current frame. Figure 6 illustrates the error curves for the three methods. As can be seen, the proposed algorithm achieves the best final albedo estimates compared to other approaches. This result shows the impact of using the Kalman filter to fuse the information over the frames. [5] gives estimates with large error in the initial frames due to the harsh illumination condition in corresponding images. On the other hand, the algorithm proposed in [1] obtains better initial albedo because of incorporating the error statistics in the calculations, but the estimate is not improved through fusion.

While estimating the 3D shape of the subjects will increase the accuracy of the results, it is time consuming. On the other hand, in many cases using the mean shape of the face can produce results with an acceptable level of accuracy. To show the effect of using the mean shape instead of the true shape for each face, we use 3D faces in the Vetter dataset to synthesize sequences with the head fixed at a position and rotating illumination, consisting of 20 frames. The albedos are estimated using both the mean shape and the true subject specific shapes of the faces. Figure 7 illustrates the effect of using the mean shape instead of the true shapes on the mean squared error of the estimated albedos with respect to the ground truth for 100 synthetic sequences. The figure also shows the spatial distribution of the mean albedo errors for the two situations. As it can be seen, the albedo errors are comparable for the two cases even for the regions around eyes and nose. Moreover, in both cases the error in the final albedo is considerably smaller than the initial error. Therefore, although using the mean shape leads to the larger final error, the fact that there is no need to know or estimate the true 3D shape of the face compensates for that.

Fig. 9. Albedo maps estimated for three sequences in the BU dataset. The rows show some of the frames in each sequence followed by the estimated albedo using the proposed algorithm up to that frame. The last column shows the averaging of the estimated albedos obtained using [1] at each frame.

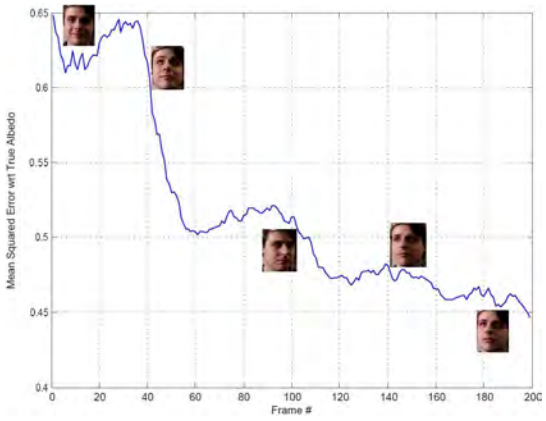


Fig. 8. The MSE of the albedo estimate versus frames for a sequence in the BU dataset along with the appearance of the face at some frames. Rotation of the face brings more information for albedo estimation and therefore reduces the estimate error.

Real sequences: Albedo estimation in real video sequences with changing pose is more challenging due to tracking errors as well as violations from the Lambertian assumption.[3] Figure 8 illustrates the decrease in the error of the estimated albedo map over the frames for a sequence from the BU dataset with varying head pose and illumination. The figure also shows the appearance of the face at those places where the error decreases drastically. As it can be seen, such sudden reductions in the albedo error happens when some previously shadowed parts of the face become illuminated, which means they bring new information for albedo estimation. It should also be noted that slight increases in errors are mainly due to tracking errors and errors created as a consequence in the albedo estimation process.

Figure 9 shows the albedo maps obtained for some

3. This source of error exists in the PIE sequences as well

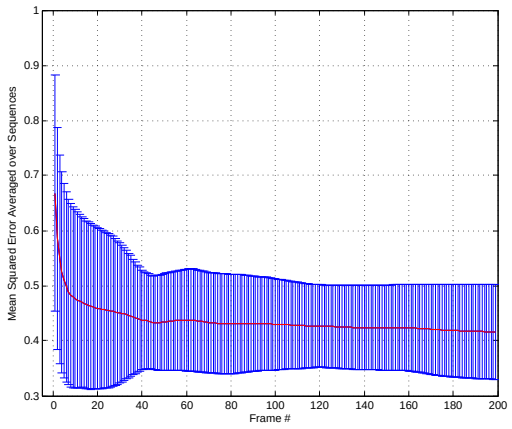


Fig. 10. The MSE in the final estimate of albedo with respect to the ground truth averaged over 80 sequences in BU dataset along with the standard deviation of the error at each frame.

of the subjects in the BU dataset. The rows show some frames of the selected sequences (including the very first and last frames), along with the estimated albedo maps using the proposed algorithm up to each frame. The selected sequences in this figure usually start with the face in a harsh illumination condition and then the motion of the face in subsequent frames brings more information regarding the reflectance properties of those shadow pixels and hence improves the albedo estimate obtained from the first frames. The last column result of the albedo rows shows the temporal averaging of the estimated albedo maps [1] over the frames. These results are blurry due to incorporating information from all the frames with equal weights.

The amount of reduction in the albedo error over the frames for various sequences depends on the illumination conditions throughout the frames as well as tracking accuracy. For some sequences the shape of the face is far from the average shape and also for some cases the illumination condition is not improving enough over the frames, so the final albedo estimate still has considerable error with respect to the ground truth, although it is less than the initial error. Figure 10 shows the average behavior of the albedo error with respect to the ground truth over 80 BU sequences along with the standard deviation of error at each frame.

Finally, to evaluate our algorithm for some extreme situations, we applied the algorithm on some sequences, collected indoors, where the face rotates in front of a fixed illumination source. Figure 11 shows some frames along with their estimated albedos for the two sequences one with a good lighting condition and the other one in a harsh illumination condition with saturated pixels and cast shadows. As can be seen, the albedo estimated using the second sequence is noisy which is mainly due to the saturated pixels. This example illustrates a limitation of our algorithm which occurs when a part of the object is not Lambertian. Since our algorithm excludes such pixels from the updating framework, their albedo estimates will not improve over the frames. But if specularity (or any other example of violation from Lambertian assumption) occurs in a limited number of frames and the surface parts that have such errors in some frames show their Lambertian properties in some other frames, our algorithm will be able to ignore the occurrences of such errors and get a good albedo estimate out of good frames in the sequence.

5.2 Illumination-Insensitive Tracking

Illumination-insensitive head pose tracking is an important part of the proposed algorithm. While pose estimation is necessary for our sequential albedo estimation algorithm, updating the albedo map at each frame also helps to have accurate pose tracking. We evaluate the performance of the tracking algorithm using sequences with both uniform and varying illumination in the BU dataset for which the ground truth pose information is available. The dataset with uniform illumination has 45 sequences for 5 subjects (9 sequences per subject) and the dataset with varying illumination has 27 sequence for 3 subjects, each sequence has 198 frames in which the face goes through several in-plane and out-of-plane rotations as well as translations.

Figure 12 shows some frames of two sequences with the tracked landmarks on the face. These examples show the ability of the tracker to maintain tracks in spite of illumination changes and large out-of-plane rotations. Figure 12 also presents a comparison between the rotation angles estimated for the left sequence, in the top row, and its ground truth. It can be seen that the tracker accurately estimates the pose of the face in almost all frames.

To have a quantitative evaluation of the precision of our tracking algorithm and evaluate its robustness against illumination variation, we use the metric introduced in [35] which is based on the Mahalanobis distance between the estimated and measured positions and orientations. Two normalized errors, position error and orientation error, are defined at each frame of the sequence. The precision of the tracker is then defined for each sequence as the root mean square error computed over the sequence up to the point where the track was lost. For this purpose we defined the track as lost when the position error at that frame exceeded a fixed threshold.

Table 1 shows the evaluation results for both subsets of the BU dataset with the uniform and varying illuminations. The percentage of tracked frames for both cases is 89.2 ± 4.003 and the averaged tracking time per frame is 2.05 ± 0.34 seconds (using Matlab software and on a 4GHz processor) where most of this time is spent for retrieving the observed intensities at the vertices of a 3D face model[4]. The timing of the joint pose and albedo estimation algorithm is the same as timing for tracking,

4. We can improve the tracking rate using a parallelized particle filter [48] as well as using a more powerful processor and by programming on C++.

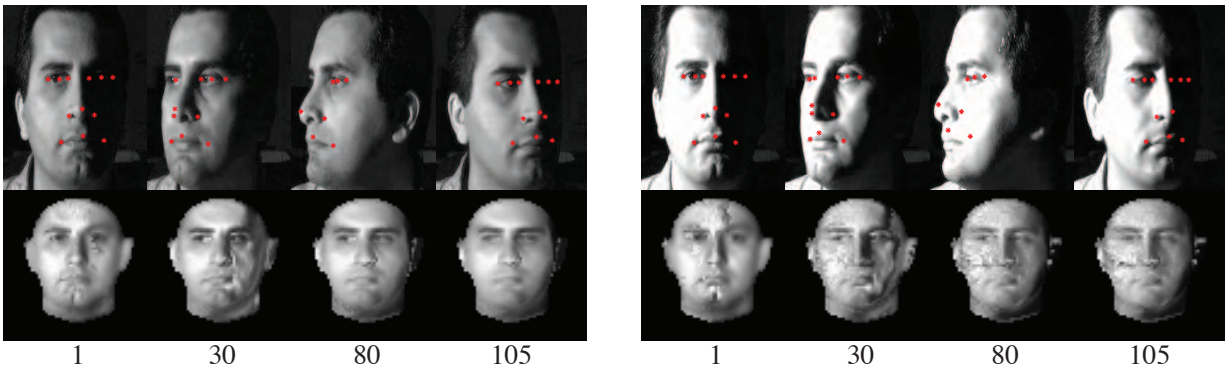


Fig. 11. Tracking results and albedo estimates for two sequences in **left:** good illumination and **right:** harsh illumination conditions. Shown below is the frame number of each image (out of 150 frames).

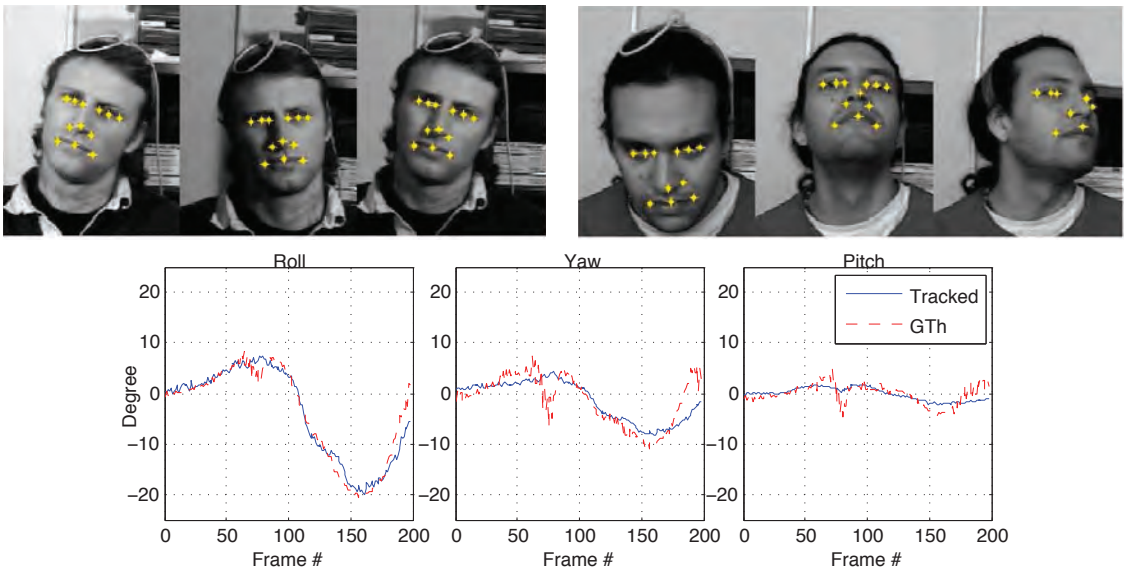


Fig. 12. Tracking results for two sequences under illumination changes and in/out-of-plane rotation. The **first row** shows faces with tracked landmarks on them, the **second row** presents a comparison between the estimated rotation angles (roll, yaw, pitch) with the ground truth for the sequence in the top-left column.

since the albedo estimation is based on just a linear operation. As the table shows, the proposed algorithm performs well for both orientation (rotation) and position (translation) estimation. Moreover, the comparable results on both datasets indicate that our algorithm is to a reasonable degree insensitive to illumination changes. However as it is expected, tracking on a dataset with varying illumination is more challenging.

	Uniform Illumination	**Varying Illumination**
Rotation	0.822 ± 0.44	1.01 ± 0.15
Translation	0.83 ± 0.42	0.95 ± 0.095

TABLE 1

Averaged tracking error in terms of both position (translation) and orientation (rotation) for two subsets of BU dataset one with uniform illumination and other one with varying illumination.

To show the importance of the albedo update step for pose tracking, we perform the pose tracking experiment using the particle filter framework, as discussed in Section 4 but without the albedo update using the Kalman filter step. So the albedo is estimated using the first frame of the sequence (we assume that the face has an almost frontal pose in the first frame of the sequence and it is partially shadowed) and the estimated spherical harmonic basis images B_t are therefore fixed throughout the pose tracking step. We compare the results from this algorithm with those of our algorithm (using RBPF) on two sequences. As Fig. 13 shows, the first algorithm looses the track whenever the face goes through an illumination change. This is because the estimated albedo and therefore the estimated spherical harmonic basis are not accurate and changing the appearance of the face due to illumination changes causes the face to not lie on the same spherical harmonic subspace. This emphasizes the importance of updating the albedo throughout the sequence so that the available information from multiple images is used to obtain a more accurate estimate of the albedo and spherical harmonic basis as a result.

Fig. 13. Comparing the pose tracking results with and without albedo updating step. First row for each sequence shows the tracking results of the particle filter without the albedo updating step and the second row has the results of the proposed algorithm. For the second sequence, we also show the estimated albedo map using the proposed algorithm at each frame. The frame numbers are shown below images.

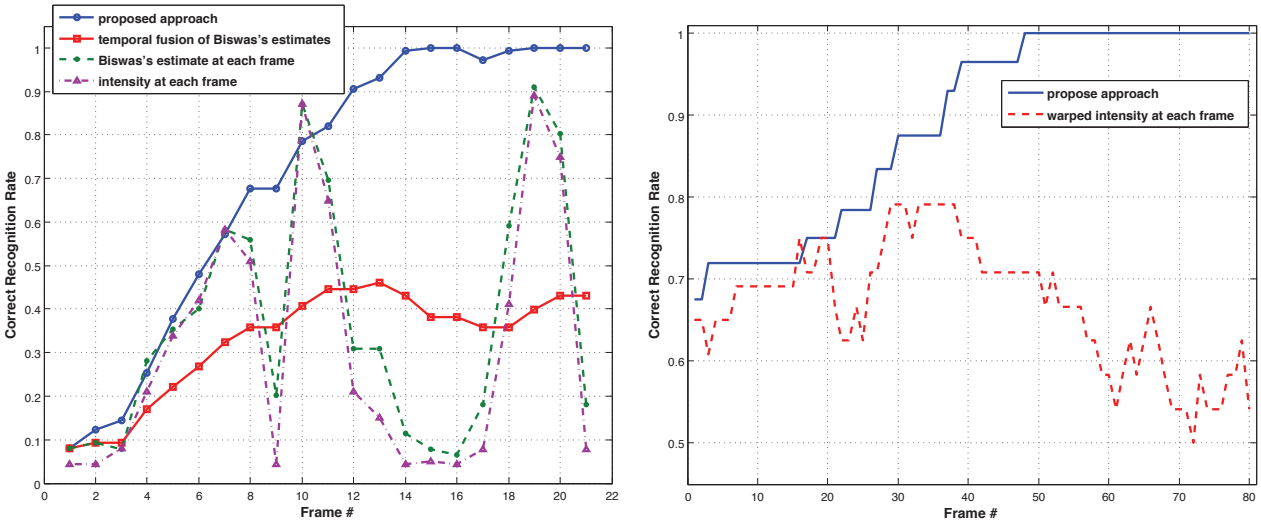


Fig. 14. Video-based face recognition rate versus number of frames used for albedo estimation. **Left:** comparing the recognition rate on the PIE synthetic sequences using the albedo maps resulting from [1] estimated at each single frame, temporal fusion of [1] estimates up to each frame, face intensity at each frame and the proposed algorithm; **Right:** recognition rate on the BU dataset using the albedo maps produced by the proposed algorithm compared to the recognition rate using the warped intensity at each frame.

5.3 Video-based Face Recognition

The resulting albedo maps provide signatures of faces that can be used as inputs to many existing 2D techniques for face recognition. Here our objective is to show the improvement in face recognition obtained due to video-based albedo estimation. To this end, we use the true albedo maps of the subjects as the gallery set and the probe set includes a number of videos per subject where the videos usually starts with the face partly in the shadow. We perform the face recognition experiment on the synthetic sequences from the PIE dataset as well as the BU sequences. In both cases we report the recognition rates averaged over all the sequences versus the number of frames used for albedo estimation. We expect the recognition rate to increase as more frames are used for albedo estimation, since the albedo maps improve over the frames.

Figure 14 illustrates the face recognition rate averaged over all the sequences versus frame number. For the PIE sequences (left column of the figure) we have 68 subjects and a sequence of 21 frames per subject in the test set. Each sequence starts with the face being partly in shadow and then the illumination rotates around the face. We apply the proposed algorithm to each sequence to obtain the estimated albedo maps up to each frame using which we perform the face recognition (blue solid curve). To show the importance of albedo estimation and proper fusion of information over frames for face recognition, we compare our results with the results from three other algorithms. First we obtain the albedo maps at each frame using Biswas' single image-based algorithm [1] and use them to perform face recognition at each frame (green dash curve). Then we temporally fuse these estimated albedo maps up to each frame (temporal averaging) and again perform recognition at each frame (red solid curve), and finally perform recognition using the face intensity at each frame (magenta dash-dotted curve).

As the plots show, recognition performance using the intensity at each frame as well as the albedo estimated from a single frame is completely dependent on the quality of the face at that frame and while it is around 90% for some frames, it decreases to below 10% for some other frames. The temporal fusion/averaging of the [1] estimates over the frames stabilizes the results but still the recognition rate is low due to the blurry albedo maps obtained through this process. But the proposed algorithm results in a considerable increase in the recognition rates over the sequences.

Similar results using the sequences in the BU dataset are presented in the right column of the figure. We have six subjects and an average of nine sequences per subject, each with 80 frames, using which we obtain the albedo maps up to each frame and perform face recognition. We compare the results from our algorithm with the case where we use the warped intensity at each frame for recognition. However since the pose of the head is changing throughout the sequence, to get the warped intensity at each frame we need to estimate the head pose. We can perform this step separately by tracking fiducial points using Kanade-Lucas-Tomasi (KLT) tracker [49], but since the illumination is changing considerably throughout the sequences, KLT fails to track the feature points. Therefore, we use the estimated poses from our algorithm (RBPF) and then obtain the warped intensity map at each frame for recognition. The results again show the superiority of having an accurate albedo map for face recognition compared to the intensity map at each frame.

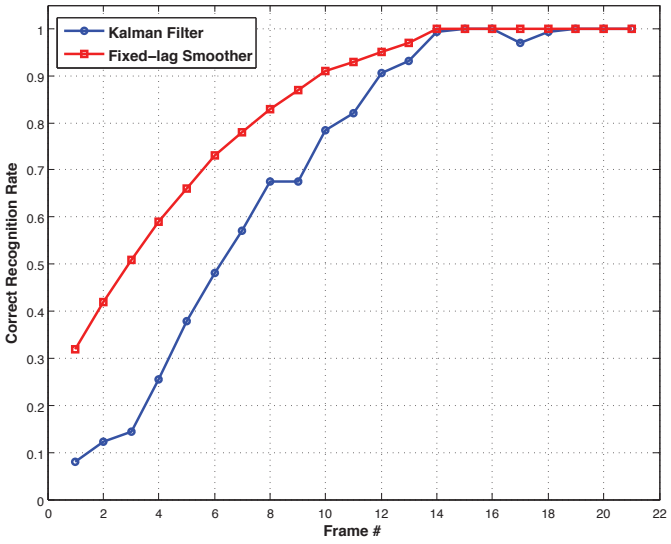


Fig. 15. Comparing the face recognition rates obtained using our approach (Kalman filter) with those of applying the fixed-lag smoother algorithm (with lag of 4 frames) on the synthetic PIE sequences.

5.4 Kalman Smoother

In the proposed algorithm, the estimate of ρ_t is made based on the noisy measurement set $\mathcal{Z}^t = \{Z_1, ..., Z_t\}$. But if a delay in the production of ρ_t be permitted, then more measurements become available during the delay interval and these new measurements can be used in producing the estimate of ρ_t. Thus a delay of N time units during which $Z_{t+1}, ..., Z_{t+N}$ appear allows estimation of ρ_t by

$$\mu_{\rho,t|t+N} = E[\rho_t | Z_1, ..., Z_{t+N}]$$

Such an estimate is called smooth estimate. Historically, three particular types of smoothing problems have been studied, each characterized by the particular subset of all possible smoothed estimates sought: Fixed-point smoothing, fixed-lag smoothing and fixed-interval smoothing. For our problem, the fixed-lag smoother [50] is the best since it allows "online" production of smoothed estimates.

Since more measurements are used in producing $\mu_{\rho,t|t+N}$ than in producing $\mu_{\rho,t|t}$, one expects the estimate to be more accurate, and generally, one expects smoothers to perform better than filters, although inherent in a smoother is a delay and, as it turns out, an

increase in estimator complexity. Further, the greater the delay, the greater the increase in complexity [50]. Thus depending on the delay and complexity that the system can tolerate, some improvements in the estimates can be obtained using smoothed estimates.

To investigate the trade-off between the delay and the improvement we obtain in the estimate of the albedo map we perform the sequential albedo estimation and face recognition experiments on the synthetic PIE sequences. We consider a lag of $N = 4$ frames and applied a fixed-lag smoother to the PIE sequences at each frame to estimate $\mu_{\rho,t|t+4}$ and then use these smoothed estimates to perform the face recognition experiment as explained in the section 5.3. Figure 15 shows the improvements that the smoothed estimates result compared to the causal estimates obtained using the Kalman filter. Increasing the value of N slightly increases the improvement.

It should be noted that the head poses are fixed in the synthetic PIE sequences, since if the head pose is varied then we need the pose information for the future frames as well which is not available at the current frame. Moreover, knowing that the delay at each frame in the proposed approach is due to the pose estimation step (the albedo estimation is performed in real-time), we need our sequential algorithm and the Kalman filter to jointly estimate the pose and albedo map at each frame.

5.5 Comparison with a Batch Processing Algorithm

We also compared our sequential albedo estimation algorithm with a batch processing method for albedo estimation. As we mentioned in the introduction, our algorithm has connections to the photometric stereo problem. Hence, we applied a photometric stereo algorithm (as a batch processing method)[5] to the synthetic PIE sequences to estimate the albedo (as well as the normal vectors to the surface) for each subject. Then we computed the average MSE of the estimated albedo and the ground truth albedo over 68 subjects. Note that while the photometric stereo algorithm optimizes the albedo globally (along with accounting for surface estimates etc.), our algorithm incrementally updates the albedo under a fixed shape model.

Table 2 shows the results for both the algorithms. As the table shows, using the batch processing method gives a slightly better error rate compared to our sequential algorithm. This result was expected since a batch processing method uses all the available information (including the future frames in our case) at once. However as we emphasized in the paper, our algorithm is applicable to a video in which frames come at a time and so the algorithm processes the information once it becomes available.

When pose variations are presented in the sequences, both the sequential and batch processing methods need correspondences across the images. We applied tracking algorithm to obtain these correspondences for the low resolution images we used for our experiments. These correspondences are not very accurate and it would be interesting to investigate the effects of these inaccurate correspondences on a batch processing method as a future work.

5. http://pages.cs.wisc.edu/ csverma/CS766_09/Stereo/stereo.html

TABLE 2
MSE error of the estimated albedos using a photometric stereo algorithm as well as our proposed sequential algorithm with respect to the ground truth averaged over 68 synthetic PIE sequences.

	Photometric Stereo	**Our approach**
Albedo MSE	0.166 ± 0.042	0.21 ± 0.08

6 Summary, Limitations and Future Work

We proposed a joint tracking and sequential albedo estimation framework using a Rao-Blackwellized particle filter. The tracking algorithm finds the best pose at each frame which minimizes the projection error of the observed appearance onto the spherical harmonic basis images. At the same time, a Kalman filter updates the albedo map estimated in the previous frame using current observations and by incorporating useful information regarding the albedo into the prior albedo estimate. Simultaneous pose and albedo estimation at each frame improves the final albedo map. Comparisons with the true poses and true albedo maps were shown to highlight the effectiveness of the algorithm. Moreover, the robustness of the algorithm against errors due to deviations from the Lambertian assumption has been evaluated. The albedo estimated using our approach can be used for video-based face recognition. The proposed algorithm has limitations and we briefly discuss some of them here.

Assumptions: The main assumption in this paper is the Lambertian assumption for the human face. As we discussed in the introduction, while this assumption is reasonable for the application in this paper, having a more accurate model for the face image formation leads to more precise results and enables improved inferences. We also make simplifying assumptions regarding independence of pixels in the face image. This assumption can be removed by adding a prior albedo model and also having full covariance matrices in the Kalman filter.

Robustness: In this work we achieved some robustness against non-Lambertian effects (e.g. cast shadow, saturation and specularity) by updating the Kalman filter observation noise matrix at each frame in an ad-hoc manner. It might be interesting to actually use a proper robust distribution instead of Gaussian as the noise model.

Speed: Sequential estimation of albedo has a lot of advantages in the context of real-time implementations.

For one, we need a small buffer to store frames before they are processed. Further, even when the processing algorithm is not real-time, we can continue processing frames using strategies for carefully dropping frames while maintaining a finite buffer. This is especially relevant in our setting. While the albedo estimation can be performed in real-time, the pose estimation step which is based on the particle filter implementation in Matlab is slow which makes the whole algorithm non-real time. However it should be noted that our choice of particle filters — in addition to providing powerful inference capabilities — also allows parallel implementation. There has been a significant body of work on parallel and pipelined implementations of particle filters [48], [51]. A basic premise of this body of work is that particle filters are extremely parallelizable and linear speedup in the number of computing nodes is very much possible. A GPU implementation with modern GPUs that have 100-1000s of computing nodes, for example, has the capability of achieving real-time performance.

The delay in estimating the albedo at each frame can be used to interpret the problem as a smoothing problem, as we discussed in the paper. But the interpretation as a fixed-lag smoothing problem is currently applicable only for the constant pose case. Extension to variable pose case would be interesting.

Deformation: We do not account for non-rigid deformation in our models. However since at each frame we only update the albedo for those pixels whose current intensity is in a reasonable range with respect to their intensity in the previous frame (we assume small changes between two frames), our algorithm can handle non-rigid deformation up to some degree. This approach can also be generalized for other types of objects which are rigid or can only have small non-rigid deformations.

ACKNOWLEDGMENTS

This work was partially supported by a MURI from the Office of Naval Research under the grant N00014-08-1-0638. The authors thank Soma Biswas for valuable and insightful discussions.

REFERENCES

[1] S. Biswas, G. Aggarwal, and R. Chellappa, "Robust estimation of albedo for illumination-invariant matching and shape recovery," *TPAMI*, vol. 31, no. 5, pp. 884–899, 2009.

[2] R. Ramamoorthi, "Modeling illumination variation with spherical harmonics," in *Face Processing: Advanced Modeling Methods*, 2006.

[3] B. K. P. Horn and M. J. Brooks, *Shape from Shading*. Cambridge Massachusetts: MIT Press, 1989.

[4] R. Zhang, P. sing Tsai, J. E. Cryer, and M. Shah, "Shape from shading: A survey," *TPAMI*, vol. 21, no. 8, pp. 690–706, 1999.

[5] L. Zhang and D. Samaras, "Face recognition from a single training image under arbitrary unknown lighting using spherical harmonics," *TPAMI*, vol. 28, no. 3, pp. 351–363, 2006.

[6] S. Zhou, G. Aggarwal, R. Chellappa, and D. Jacobs, "Appearance characterization of linear lambertian objects, generalized photometric stereo and illumination-invariant face recognition," *TPAMI*, vol. 29, no. 2, pp. 230–245, 2007.

[7] A. Lakdawalla and A. Hertzmann, "Shape from video: Dense shape, texture, motion and lighting from monocular image streams," in *Photometric Analysis for Comp. Vision.*, 2007.

[8] Y. Xu and A. Roy-Chowdhury, "Integrating motion, illumination, and structure in video with applications in illumination-invariant tracking," *TPAMI*, vol. 29, no. 5, pp. 793–806, 2007.

[9] R. Basri and D. Jacobs, "Lambertian reflectance and linear subspaces," *TPAMI*, vol. 25, no. 2, pp. 218–233, 2003.

[10] Z. Wen, Z. Liu, and T. Huang, "Face relighting with radiance environment maps," in *CVPR*, vol. 2, 2003.

[11] Y. Xu, A. Roy-Chowdhury, and K. Patel, "Pose and illumination invariant face recognition in video," in *CVPR*, 2007.

[12] A. Barmpoutis, R. Kumar, B. C. Vemuri, and A. Banerjee, "Beyond the lambertian assumption: A generative model for ABRDF fields of faces using anti-symmetric tensor splines," in *CVPR*, 2008.

[13] T. Weyrich, W. Matusik, H. Pfister, B. Bickel, C. Donner, C. Tu, J. Mcandless, J. Lee, A. Ngan, H. Wann, and J. M. Gross, "Analysis of human faces using a measurement-based skin reflectance model," *ACM Trans. Graphics*, vol. 25, no. 3, pp. 1013–1024, 2006.

[14] Y. Wang, Z. Liu, G. Hua, Z. Wen, Z. Zhang, and D. Samaras, "Face re-lighting from a single image under harsh lighting conditions," in *CVPR*, 2007.

[15] N. Joshi and D. J. Kriegman, "Shape from varying illumination and viewpoint," in *ICCV*, 2007.

[16] C. H. Esteban, G. Vogiatzis, and R. Cipolla, "Multiview photometric stereo," *TPAMI*, vol. 30, no. 3, pp. 548–554, 2008.

[17] J. Lim, J. Ho, M. hsuan Yang, and D. Kriegman, "Passive photometric stereo from motion," in *ICCV*, 2005.

[18] L. Zhang, B. Curless, A. Hertzmann, and S. M. Seitz, "Shape and motion under varying illumination: Unifying structure from motion, photometric stereo, and multi-view stereo," in *ICCV*, 2003.

[19] D. Simakov, D. Frolova, and R. Basri, "Dense shape reconstruction of a moving object under arbitrary, unknown lighting," in *ICCV*, 2003, pp. 1202–1209.

[20] R. J. Woodham, "Photometric method for determining surface orientation from multiple images," *Optical Engineerings*, vol. 19, no. 1, pp. 139–144, 1980.

[21] K. Ikeuchi, "Determining surface orientation of specular surfaces by using the photometric stereo method," *TPAMI*, vol. 3, no. 6, pp. 661–669, 1981.

[22] W. Y. Zhao and R. Chellappa, "Symmetric shape-from-shading using self-ratio image," *Intl. J. Computer Vision*, vol. 45, no. 1, pp. 55–75, 2001.

[23] J. Atick, P. Griffin, and A. Redlich, "Statistical approach to SFS: Reconstruction of 3D face surfaces from single 2D images," *Neural Computation*, vol. 8, pp. 1321–1340, 1996.

[24] R. Dovgard and R. Basri, "Statistical symmetric shape from shading for 3D structure recovery of faces," in *ECCV*, 2004.

[25] W. A. P. Smith and E. R. Hancock, "Recovering facial shape using a statistical model of surface normal direction," *TPAMI*, vol. 28, no. 12, pp. 1914–1930, 2006.

[26] B. K. P. Horn, "Determining lightness from an image," *Computer Graphics, Image Processing*, vol. 3, pp. 277–299, 1974.

[27] E. H. Land and J. J. McCann, "Lightness and retinex theory," *J. of the Optical Society of America*, vol. 61, no. 1, pp. 1–11, 1971.

[28] M. F. Tappen, E. H. Adelson, and W. T. Freeman, "Estimating intrinsic component images using non-linear regression," in *CVPR*, 2006.

[29] V. Blanz and T. Vetter, "Face recognition based on fitting a 3D morphable model," *TPAMI*, vol. 25, no. 9, pp. 1063–1074, 2003.

[30] Y. Wang, L. Zhang, Z. Liu, G. Hua, Z. Wen, Z. Zhang, and D. Samaras, "Face relighting from a single image under arbitrary unknown lighting conditions," *TPAMI*, vol. 31, no. 11, pp. 1968–1984, 2009.

[31] S. Biswas and R. Chellappa, "Pose-robust albedo estimation from a single image," in *CVPR*, 2010, pp. 2683 – 2690.

[32] T. Yu, N. Xu, and N. Ahuja, "Recovering shape and reflectance model of non-Lambertian objects from multiple views," in *CVPR*, 2004.

[33] G. Aggarwal, A. Veeraraghavan, and R. Chellappa, "3D facial pose tracking in uncalibrated videos," in *PReMI*, 2005.

[34] S. Ba and J. Odobez, "A probabilistic head pose tracking evaluation in single and multiple camera setups," in *CLEAR, Evaluation and Workshop*, 2007.

[35] M. L. Cascia, S. Sclaroff, and V. Athitsos, "Fast, reliable head tracking under varying illumination: An approach based on registration of texture-mapped 3D models," *TPAMI*, vol. 22, no. 4, pp. 322–336, 2000.

[36] T. K. Marks, J. R. Hershey, and J. R. Movellan, "Tracking motion, deformation, and texture using conditionally gaussian processes," *TPAMI*, vol. 32, no. 2, pp. 348–363, 2010.

[37] E. Murphy-Chutorian and M. Trivedi, "Head pose estimation in computer vision: A survey," *TPAMI*, vol. 31, no. 4, pp. 607–626, 2008.

[38] Y. Tanabe, T. Inui, and Y. Onodera, *Group Theory and Its Applications in Physics*. Springer, 1990.

[39] Z. Yue, W. Zhao, and R. Chellappa, "Pose-encoded spherical harmonics for face recognition and synthesis using a single image," *EURASIP Journal on Advances in Signal Processing*, 2008.

[40] V. Blanz and T. Vetter, "A morphable model for the synthesis of 3d faces," in *SIGGRAPH*, 1999.

[41] P. Paysan, R. Knothe, B. Amberg, S. Romdhani, and T. Vetter, "A 3D face model for pose and illumination invariant face recognition," in *IEEE Intl. Conf. AVSS*, 2009.

[42] D. F. DeMenthon and L. S. Davis, "Model-based object pose in 25 lines of code," *Intl. J. Computer Vision*, vol. 15, no. 1-2, pp. 123–141, 1995.

[43] R. B. Altman, "A probabilistic algorithm for calculating structure: Borrowing from simulated annealing," Stanford University, Tech. Rep., 1990.

[44] R. B. Altman and J. F. Brinkley, "Probabilistic constraint satisfaction with structural models," in *Symposium on Computer Applications in Medical Care*, 1993.

[45] P. S. Maybeck, *Stochastic models estimation and control*. Academic Press, 1979.

[46] T. Sim, S. Baker, and M. Bsat, "The CMU pose, illumination, and expression database," *TPAMI*, vol. 25, no. 12, pp. 1615–1618, 2003.

[47] S. Ba and J. Odobez, "A Rao-blackwellized mixed state particle filter for head pose tracking in meetings," in *ACM MMMP*, 2005.

[48] A. C. Sankaranarayanan, A. Srivastava, and R. Chellappa, "Algorithmic and architectural optimizations for computationally efficient particle filtering," *IEEE Trans. on Image Processing*, vol. 17, no. 5, pp. 737–748, 2008.

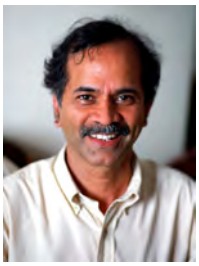
[49] J. Shi and C. Tomasi, "Good features to track," in *CVPR*, 1994.

[50] B. Anderson and J. Moore, *Optimal Filtering*. Prentice Hall, 1979.

[51] S. Hong, S. S. Chin, P. M. Djurić, and M. Bolić, "Design and implementation of flexible resampling mechanism for high-speed parallel particle filters," *J. VLSI Signal Processing*, vol. 44, no. 1-2, pp. 47–62, 2006.

Sima Taheri received her B.Sc and M.Sc degrees in Electrical Engineering from the Sharif University of Technology, Iran in 2003 and 2005, respectively and the M.Eng degree from the Department of Electrical and Computer Engineering at the National University of Singapore in 2007. She is currently a Ph.D candidate in the Department of Computer Science at the University of Maryland, College Park. Her research interests lie in the areas of computer vision, machine learning and medical image analysis.

Aswin C. Sankaranarayanan is a research scientist in the Department of Electrical and Computer Engineering at Rice University, Houston, TX. His research interests lie in the areas of computer vision, signal processing, and image and video acquisition. Dr. Sankaranarayanan received his B.Tech in Electrical Engineering from the Indian Institute of Technology, Madras in 2003 and MSc and PhD degrees from the Department of Electrical and Computer Engineering at the University of Maryland, College Park in 2007 and 2009, respectively. He was awarded the distinguished dissertation fellowship by the ECE department at the University of Maryland in 2009.



Rama Chellappa received the B.E. (Hons.) degree from the University of Madras, India, in 1975 and the M.E. (Distinction) degree from the Indian Institute of Science, Bangalore, in 1977. He received the M.S.E.E. and Ph.D. Degrees in Electrical Engineering from Purdue University, West Lafayette, IN, in 1978 and 1981 respectively. Since 1991, he has been a Professor of Electrical Engineering and an affiliate Professor of Computer Science at University of Maryland, College Park. He is also affiliated with the Center for Automation Research and the Institute for Advanced Computer Studies (Permanent Member). In 2005, he was named a Minta Martin Professor of Engineering. Prior to joining the University of Maryland, he was an Assistant (1981-1986) and Associate Professor (1986-1991) and Director of the Signal and Image Processing Institute (1988-1990) at University of Southern California, Los Angeles. Over the last 31 years, he has published numerous book chapters, peer-reviewed journal and conference papers in image processing, computer vision and pattern recognition. He has co-authored and edited books on MRFs, face and gait recognition and collected works on image processing and analysis. His current research interests are face and gait analysis, markerless motion capture, 3D modeling from video, image and video-based recognition and exploitation, compressive sensing, sparse representations and domain adaptation methods. Prof. Chellappa served as the associate editor of four IEEE Transactions, as a Co-Editor-in-Chief of Graphical Models and Image Processing and as the Editor-in-Chief of IEEE Transactions on Pattern Analysis and Machine Intelligence. He served as a member of the IEEE Signal Processing Society Board of Governors and as its Vice President of Awards and Membership. Recently, he completed a two-year term as the President of IEEE Biometrics Council. He has received several awards, including an NSF Presidential Young Investigator Award, four IBM Faculty Development Awards, an Excellence in Teaching Award from the School of Engineering at USC, two paper awards from the International Association of Pattern Recognition, the Society, Technical Achievement and Meritorious Service Awards from the IEEE Signal Processing Society and the Technical achievement and Meritorious Service Awards from the IEEE Computer Society. He has been selected to receive the K.S. Fu prize from the International Association of Pattern Recognition. At University of Maryland, he was elected as a Distinguished Faculty Research Fellow, as a Distinguished Scholar-Teacher, received the Outstanding Faculty Research Award from the College of Engineering, an Outstanding Innovator Award from the Office of Technology Commercialization, the Outstanding GEMSTONE Mentor Award and the Poole and Kent Teaching Award for Senior Faculty. He is a Fellow of IEEE, the International Association for Pattern Recognition, the Optical Society of America and the Association for the Advancement of Science. In 2010, he received the Outstanding ECE Award from Purdue University. He has served as a General and Technical Program Chair for several IEEE international and national conferences and workshops. He is a Golden Core Member of IEEE Computer Society and served a two-year term as a Distinguished Lecturer of the IEEE Signal Processing Society.