Chapter 1

3D MODEL-DRIVEN VEHICLE MATCHING AND RECOGNITION

Tingbo Hou[1], Sen Wang[2], Hong Qin[1]

[1] *Department of Computer Science*
Stony Brook University
Stony Brook, NY 11794, USA
thou@cs.sunysb.edu
qin@cs.sunysb.edu

[2] *Kodak Research Laboratories*
Eastman Kodak Company
Rochester, NY 14650, USA
sen.wang@kodak.com

Abstract Matching vehicles subject to both large pose transformations and extreme illumination variations remains a technically challenging problem in computer vision. In this chapter, we first investigate the state-of-the-art studies on vehicle matching, inverse rendering by which illumination can be factorized from the light reflectance field, and applications of the Near-IR illumination in computer vision. Then a 3D model-driven framework is developed, towards matching and recognizing vehicles with varying pose and (visible or Near-IR) illumination conditions. We adopt a compact set of 3D models to represent basic types of vehicle. The pose transformation is estimated by using approximated vehicle models that can effectively match objects under large viewpoint changes and partial occlusions. Second, with the estimation of surface reflectance property, illumination conditions are approximated by a low-dimensional linear subspace using spherical harmonics representation. By estimated pose and illumination conditions, we can re-render vehicles in the reference image to generate the relit image with the same pose and illumination conditions as the target image. Finally, we compare the relit image and the re-rendered target image to match vehicles in the original ref-

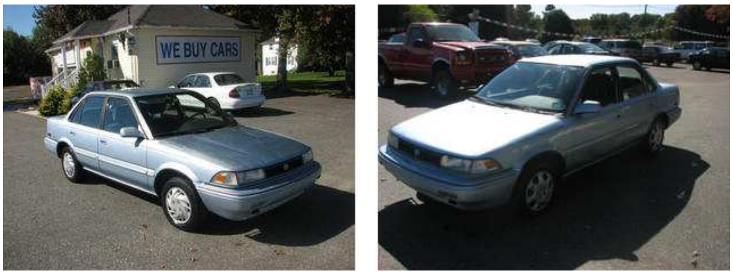

Figure 1.1. Images of the same vehicle taken from different viewpoints and lightings, subject to large pose and illumination variations.

erence image and target image. Furthermore, no training is needed in our framework and re-rendered vehicle images in any other viewpoints and illumination conditions can be obtained from just one single input image. In our experiments, both synthetic data and real data are used. Experimental results demonstrate the robustness and efficacy of our framework, with a potential to generalize our current method from vehicles to handle other types of objects.

Keywords: vehicle matching, 3D model-driven method, inverse rendering, spherical harmonics, near-IR illumination.

1. Introduction

Object matching and recognition remain an important and long-term task with continuing interest from computer vision and various applications in security, surveillance, and robotics. Many types of representations have been exploited to match and recognize objects by a set of low-dimensional parameters, such as shape, texture, structure, and other specific feature patterns. However, when it comes to unconstrained conditions such as highly varying pose and severely changing illumination, the problem becomes extremely challenging. As shown in Figure 1.1, object appearance may be tremendously different with varying pose and illumination conditions. Although the texture of a vehicle is consistent, its appearance indeed varies a lot under different lightings. Thus, such clues like shape and texture are weak in this case.

Currently, popular approaches in object recognition focus on two trends: the appearance-based methods (Murase and Nayar, 1995; Fergus et al., 2006) and the model-based methods (Gardner and Lawton, 1996; Romdhani et al., 2002). In appearance-based methods, objects are typically represented by a group of feature vectors, and a set of positive and negative examples is adopted to train a classifier spanning on the

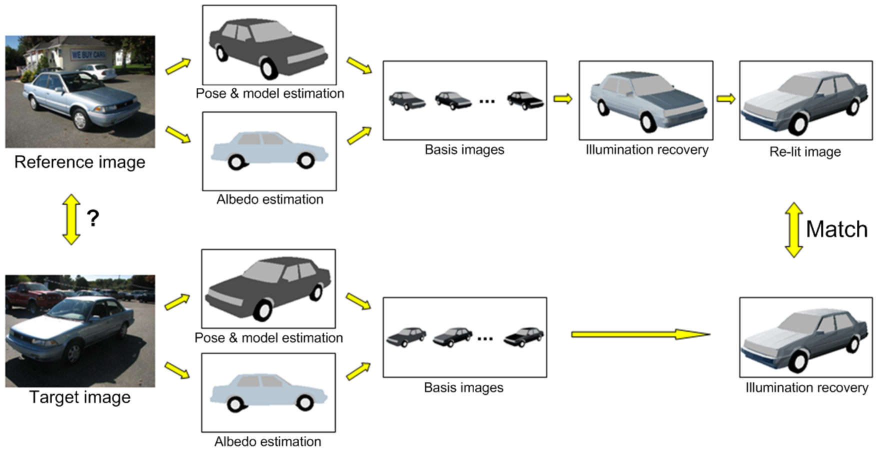

Figure 1.2. Our vehicle matching framework. Given input images, the model and pose transformation are first determined. Then we estimate reflectance property (albedo) of objects. The illumination is recovered by spanning given image to basis images, and the re-lit image is generated by transferring illumination. Finally, objects are compared in a shared domain with recovered illumination.

principle component analysis (PCA) subspace or feature subspace. In practice, technical issues arise from appearance variation due to different pose and lightings. Model-based methods require a set of 3D models to provide geometric constraints. Ideally, when object domain is known, the explicit utilization of 3D models can largely alleviate the problem of feature matching. However, it stands on two basic assumptions: first, the 3D model can precisely fit to the input images; second, pose estimation is accurate enough. To estimate appearance of objects, global and local clues have been used to simulate texture of the 3D model. Despite the progress, it still has limited success in illumination variations, since illumination conditions can dramatically affect appearances as shown in Figure 1.1.

The union of model and illumination is appealing, since appearances can be decomposed and reassembled by them. It provides a pose and illumination invariant view to examine the problem of matching and recognition. However for general objects, it is hard to obtain their 3D models from single image. To alleviate this restriction, we choose vehicle as object domain with simple geometric structure. The illumination can be visible spectrum or Near-Infrared (IR) spectrum. The Near-IR light can also be reflected by objects since it is close to the visible light in spectrum. One benefit of Near-IR illumination is to allow our method work in low-luminance environment with active Near-IR light sources. Thus the primary contribution of this chapter lies in a 3D model-driven framework towards vehicle matching and recognition working under visible or Near-IR illumination, which can handle large pose transformations and illumination variations simultaneously. Our vehicle matching framework is shown in Figure 1.2. Given original input images, the pose transformation is first estimated by using approximated 3D vehicle models that can effectively match objects under large viewpoint changes and partial occlusions. Second, we estimate reflectance property of objects, taking advantage of the fact that the body of a vehicle has unified color and material. After that, we compute their spherical harmonic basis and recover illumination conditions both in the reference image and target image. By effectively estimating both pose and illumination conditions, we can re-render vehicles in the reference image to generate the relit image with the same pose and illumination conditions as the target image. Finally, we make comparisons between the relit image and the re-rendered target image to match vehicles in the original reference image and target image.

2. Previous Work

In this section, we will investigate previous related work in vehicle matching, inverse rendering and Near-IR illumination.

Vehicle Matching

Vehicle matching has been studied in many areas of computer vision, with different purposes such as detection, identification, tracking, and recognition. Appearance-based methods are well applied on vehicles, with no difference with other objects. Recently, Shan et al., 2005 exploited an embedding vector to represent each vehicle image by exemplars of vehicles within the same camera. Each component of this vector is a non-metric distance computed by oriented edge maps. The measurement they defined describes the appearance-based same-different probabilities of two vehicles. The extended work was done by Guo et al., 2007; Shan et al., 2008 for vehicle matching. Here, we pay more attention on model-based methods, since 3D model can connect appearances from multiple views. And thus large pose variation can be easily handled. A vehicle has concise shape that can be easily represented by a simple 3D model. Koller et al., 1993 represented vehicles by a general 3D model parameterized by 12 length parameters. Their method needs to calibrate a moving plane from video sequences. Kim and Malik, 2003 used a simple sedan model to detect vehicle, and used probabilistic feature grouping for vehicle tracking. Guo et al., 2008 proposed a model-based approach to match vehicles. They used approximate 3D models to handle pose transformation, and a piecewise Markov Random Field (MRF) model to guess texture of occluded parts. However, their method has limitations on sensitive model fitting and varying illumination. Another benefit for model-based methods is that illumination as a higher dimension is possible analyzed when the 3D shape is known. In the work of Hou et al., 2009, a vehicle matching framework was proposed using a compact set of vehicle models and spherical harmonics representation of illumination.

Inverse Rendering

Illumination can be interpreted as one of the attributes of light reflectance field. Its analysis and manipulation can be fulfilled by factorizing illumination from images, which is named “inverse rendering”. Inverse rendering which measures rendering attributes: lighting, texture, and bidirectional reflectance distribution function (BRDF) from photographs, continues to be an active research area with interest from both computer vision and computer graphics. In previous work, tremendous progress has been made in the recovery of these three rendering attributes with one or two unknowns (Sato et al., 1997; Yu et al., 1999; Debevec et al., 2000). In general cases where lighting, texture, and BRDF are all unknown, this problem becomes ill-conditioned until strong assumptions and requirements on input data have been made. Ramamoorthi and Hanrahan, 2001 presented a signal processing framework for inverse rendering with known geometry and isotropic BRDFs. In their work, the reflected light field was expressed as a convolution of the lighting and BRDF using spherical harmonics. As a frequency-space convolution, spherical harmonics has been used as a tool to represent lighting. In the work of Basri and Jacobs, 2003, it is shown that the reflected light field from a Lambertian surface can be characterized using only its first 9 spherical harmonic coefficients, where geometry is assumed to be known. Later, Zhang et al., 2005; Zhang and Samars, 2006 integrated the spherical harmonic illumination representation into the Morphable Model approach, by modulating the texture component with the spherical harmonic bases. They used PCA to initialize geometry and texture from a large set of training data, and estimate lighting and basis images independently through iteration. To alleviate the strong requirements on geometry and texture, Wang et al., 2007 proposed a subregion based framework that uses a MRF to model the statistical distribution and spatial coherence of texture. Though lighting in a small region is more homogeneous, it is hard to segment an image into homogeneous regions. Their method still needs training data to compute PCA texture model.

Near-IR Illumination

Low-cost infrared cameras make it possible to address computer vision problems in a larger range of the electromagnetic spectrum (Morris et al., 2005). Here, we only focus on the near-infrared illumination with wavelength varying from $0.7\mu m$ to $1\mu m$. It is very close to the visible spectrum, and thus it can be reflected by objects, generating IR images similar with images under visible spectrum. Novotny and Ferrier, 1999 used active IR to measure distance. They proposed a method of determining the reflectance property of a surface under infrared illumination using Phong model, since IR LEDs are well approximated as a point light source. Ji and Yang, 2002 studied real time 3D face pose discrimination based on active IR illumination. The IR is adopted since pupils in IR images are more clear and stable than images under visible illumination. In the work of Zhu et al., 2002; Zhao and Grigat, 2006, active Near-IR illumination was employed in eye detection and eye tracking. Zou et al., 2005 used active Near-IR illumination projected by LED light source to illumination invariant face recognition. The Near-IR light source can provide constant illumination, and produce images with higher quality than images under ambient illumination. More work on face recognition using active Near-IR illumination can be found in the work of Pan et al., 2007; Li et al., 2007. Wang et al., 2008 presented a method for relighting faces for reducing the effects of uneven lighting and color in video conference. Their setup consists of a compact lighting rig and two cameras. The IR camera is 8 times (120fps/15fps) faster than the color camera. They used active IR lights to obtain an illumination bases of the scene, and thus they can image relighting. In the work of Fredembach and Süsstrunk, 2009, illuminant was detected and estimated in Near-IR images by simply looking at the ratios of two images: a standard RGB image and a Near-IR only image. As the differences between illuminants are amplified in the near-infrared, this estimation proves to be more reliable than using only the visible band.

3. Vehicle Matching Framework

In this section, we will introduce our framework of vehicle matching under various pose and illumination conditions.

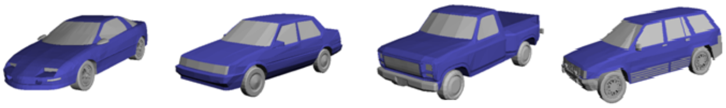

Figure 1.3. Some 3D vehicle models adopted in our framework. Models are selected from the Princeton Shape Benchmark.

Model Determination

Our dataset contains 5 representative models that stand for 5 different categories of vehicles including compact-size car, full-size sedan, small pickup truck, SUV, and big truck. These 3D models are selected from the Princeton Shape Benchmark [1], with some of them shown in Figure 1.3. Unlike the approach by Guo et al., 2008, which requires each vertex in 3D model has its semantic ownership, we take the body of a vehicle as an object and ignore some parts (windows, wheels, and lights) for such reasons: (1) typically, the body has uniform color and material, which leads to uniform reflectance property to illumination; (2) the removed parts have different patterns and properties. For example, windows could have mirror reflection, wheels may turn right or left with the same pose of the body, and lights could be on or off.

For each input image, we will first determine which model best represents the vehicle that appears in the image. Considering the fact that pose estimation is easily trapped into local minimum in the searching space, we select three different initial poses for each vehicle model with reasonable projection. For each model, we compute edge maps under these three initial fittings and use chamfer distance (Shan et al., 2005) to measure the similarity with edge maps of the original input image, as shown in Figure 1.5. Finally, we select the top two matched models as candidates for the next step.

Pose Recovery

Here, pose recovery refers to aligning a 3D model to a object in the image. This task is easier to perform if the 3D model and the image have similar features. A few correspondences will be enough to perform the alignment, since the objects are rigid. However, the visual contents of geometric models are unknown at this stage. We only have their geometric information, i.e. 3D coordinates and normals. So it comes to a simple question: How to compare geometry with texture?

An intuitive idea is to utilize the geometric edges, i.e., silhouettes and intersecting lines of two smooth surfaces, which happen to appear in the edge maps of images. We employ an approach inspired by the one

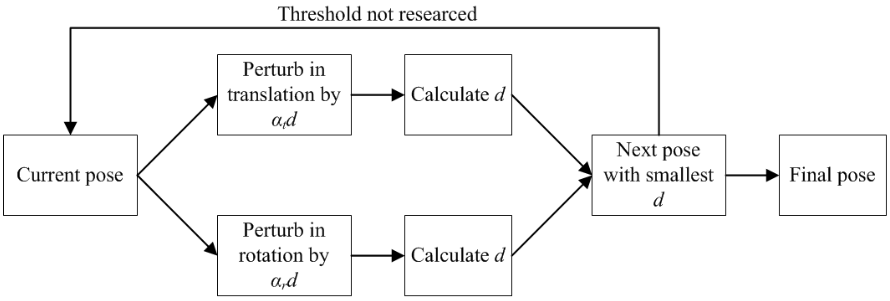
Threshold not researced
Perturb in translation by $\alpha_t d$
Calculate d
Current pose
Next pose with smallest d
Final pose
Perturb in rotation by $\alpha_r d$
Calculate d

Figure 1.4. Flow chart of pose recovery. Please refer to texts for more details.

in Guo et al., 2008 that used an Iterative Closet Point (ICP)-directed search to iteratively align the geometric edges to the image edges. The searching space is spanned by six independent components: three translation elements and three rotation angles along three axes, by isometric sampling. By making this simplification, we assume that the intrinsic parameters of cameras are fixed.

The flow chart of our pose recovery is shown in Figure 1.4. For a current pose of a candidate model, we search for the next better pose with minimal average closet point distance d, given by

$$d = \frac{1}{N} \sum_{i=1}^{N} d_i, \tag{1.1}$$

where d_i denotes the distance between pixel i in the geometric edges and its closet pixels in the image edges. We sample the searching space of translation and rotation respectively by perturbing a sampling distance. In 3D translation, we use 3 samplings for each direction, with positive, zero, and negative distance, that is, 27 samplings, and similarly in rotation, 27 samplings in 3 angles along 3 axes. The scale of geometric edges can be adjusted through the translation in the direction of depth. Furthermore, we employ adaptive sampling distances $\alpha_t d$ and $\alpha_r d$, where α_t and α_r are scaling parameters of translation and rotation. Thus, the speed of searching can be controlled in the way that when it is getting close to the minimum distance, the sampling distance is getting smaller to achieve a more precise search.

The searching will stop when the average closest point distance d reaches a threshold. However, when the searching gets stuck at some point, which means it keeps choosing zero sampling distance, while the threshold has not been reached, the sampling distance will jump to Ds,

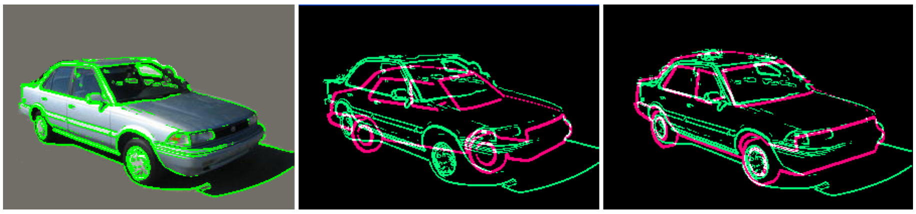

Figure 1.5. Pose estimation. Edge detection in the original image is shown on the left. Initial fitting between 3D vehicle model and the original image is shown in the middle. The alignment is shown on the right. Green lines are image edges and red lines are geometric edges of the 3D vehicle model.

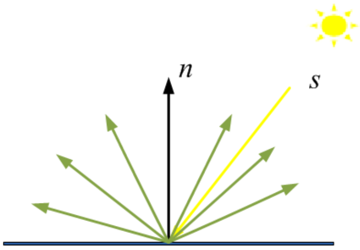

Figure 1.6. Lambertian reflectance with point light source: Light (green) is scattered equally to every direction regardless of the observer's angle of view, and the intensity is subject to the Lambertian Cosine Law.

where D is a large factor to pull the searching out of the local minimum. Finally, the best 3D object model is selected with minimal average closest point distance from candidate models. Figure 1.5 shows an example of pose recovery, where green lines are image edges and red edges are geometric edges of the 3D vehicle model.

Estimation of Reflectance Fraction

Albedo is the fraction of light that a surface point reflects when it is illuminated, which is an intrinsic property that depends on materials of the surface. There are some approaches in literature to estimate albedo from a single image (Biswas et al., 2007). In previous work of applying spherical harmonics (Zhang et al., 2005), the brightness of a pixel is taken as albedo. In our framework, taking the observation that the body of a vehicle has uniform texture and materials, we estimate albedo in RGB channels for visible spectrum.

For Lambertian objects, the diffused component of the surface reflection satisfies Lambertian Cosine Law (as shown in Figure 1.6), given by

$$I = \rho max(n^T s, 0), \tag{1.2}$$

where I is the pixel intensity, s is the light source direction, ρ is the reflectance fraction of the surface (albedo), and n is the surface normal of the corresponding 3D points. The expression implicitly assumes a single dominant light source placed at infinity, which is the most common case where vehicle images are taken. Note that Lambertian law in its pure form is nonlinear due to the *max* function, which accounts for the formation of attached shadows. Shadows and specularities do not reveal any information about their reflectivity. Thus they should not be included in the computation of estimation. In most cases, vehicle images are taken outside where the primary light source is the sun, and thus the estimation is realistic.

By collecting 3D points with positive $(n^T s)$ and the corresponding image pixels excluding shadows and specularities, we can obtain a reflective equation for each point in the 3D model, written as:

$$n^T \rho s = I. \tag{1.3}$$

Note that s is almost the same for each point in the 3D model, since the only dominant light source is placed at infinity. Therefore, we can get a formula for all reflective equations as (for example in the red channel):

$$N \rho_r s = I_r, \tag{1.4}$$

where N is the $n \times 3$ matrix that consists of surface normals of n points, ρ_r is the albedo in the red channel, and I_r is intensity value of the red component of n corresponding pixels in the image. And so are the green and blue channels. We then take $\rho_r s$ together as a variable and estimate it by the method of least squares. Since ρ_r is a positive fraction in the range $[0, 1]$, and s is the normalized direction vector whose length equals 1, we can compute ρ_r by

$$\rho_r = \frac{|\rho_r s|}{|s|} = |\rho_r s|. \tag{1.5}$$

Similarly, we can compute albedo in green channel ρ_g and albedo in blue channel ρ_b. Figure 1.7 shows that albedo maps in the second row are estimated from 3 input images in the first row. The two left images are taken from the same car and the right-most image is from another car. Despite varying illuminations, the albedo estimation is accurate and robust.

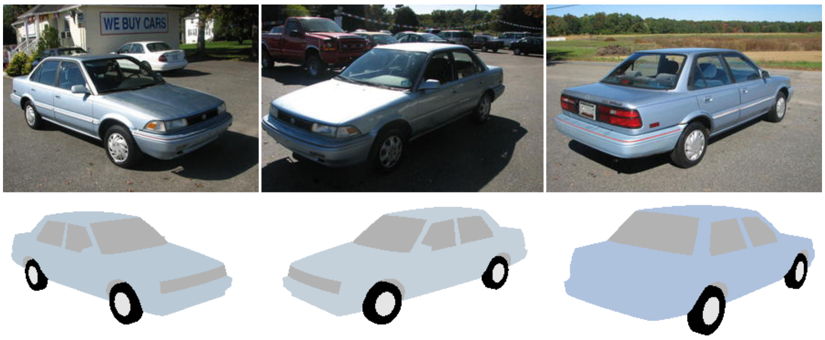

Figure 1.7. Reflectance fraction (albedo) estimation. The input images are shown in the first row and estimated albedos are shown in the second row. The two left images come from the same car and the right-most image comes from another car.

Illumination Recovery

As described by Basri and Jacobs, 2003; Ramamoorthi and Hanrahan, 2001, any image under arbitrary illumination conditions can be approximately represented by a linear combination of spherical harmonic basis as:

$$I \approx bl, \tag{1.6}$$

where b is the spherical harmonic basis and l is the vector of illumination coefficients. The set of images of a convex Lambertian object obtained under a wide variety of lighting conditions can be approximated accurately by a 9-dimensional linear subspace (Basri and Jacobs, 2003; Ramamoorthi and Hanrahan, 2001; Zhang and Samars, 2006). They are the sphere analog of the Fourier basis on the line or circle. The first 9 spherical harmonic basis images of an object can be computed by:

$$\begin{aligned}
&b_{00} = \tfrac{1}{\sqrt{4\pi}}\lambda, && b_{10}^{e} = \sqrt{\tfrac{3}{4\pi}}\lambda.*n_z,\\
&b_{11}^{o} = \sqrt{\tfrac{3}{4\pi}}\lambda.*n_y, && b_{11}^{e} = \sqrt{\tfrac{3}{4\pi}}\lambda.*n_x,\\
&b_{20} = \tfrac{1}{2}\sqrt{\tfrac{3}{4\pi}}\lambda.*(2n_{z^2} - n_{x^2} - n_{y^2}),\\
&b_{21}^{o} = 3\sqrt{\tfrac{5}{12\pi}}\lambda.*n_{yz}, && b_{21}^{e} = 3\sqrt{\tfrac{5}{12\pi}}\lambda.*n_{xz},\\
&b_{22}^{o} = 3\sqrt{\tfrac{5}{12\pi}}\lambda.*n_{xy}, && b_{22}^{e} = \tfrac{3}{2}\sqrt{\tfrac{5}{12\pi}}\lambda.*(n_{x^2} - n_{y^2}),
\end{aligned} \tag{1.7}$$

where the superscripts e and o denote the odd and the even components of the harmonics, respectively, λ is the vector of the object's albedo, n_x, n_y, n_z are three vectors of the same length that contain the x, y, and z components of the surface normals. Further, n_{xy} is a vector such that

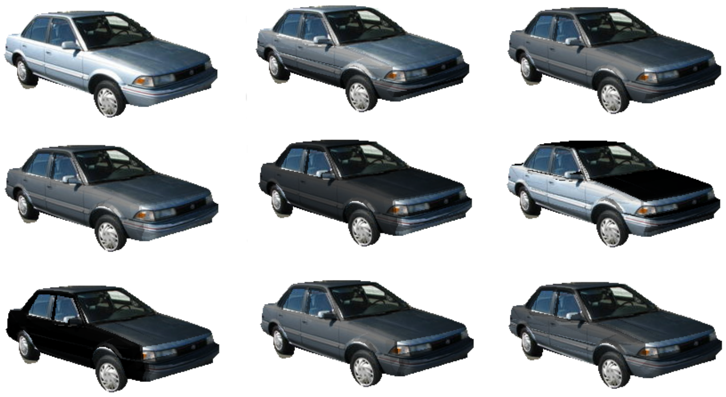

Figure 1.8. An example of the first 9 spherical harmonic basis images with RGB channels. Light colors represent positive values and darker colors represent negative values.

the ith element $n_{xy,i} = n_{x,i}n_{y,i}$, and $\lambda. * v$ denote the component-wise product of λ with any vector v.

In our framework, we use unified estimated albedo for the body of the vehicle model. The visible part of a 3D vehicle model, which is projected to the input image due to recovered pose, provides us normal vectors, estimated albedo, and appearances with illumination effects for each visible 3D point associated with corresponding 2D pixels. Therefore, we can compute the first 9 spherical harmonic basis using Equation (1.7), and estimate the illumination coefficients l by using the method of least squares in Equation (1.6). Figure 1.8 shows an example of the first 9 spherical harmonic basis images with RGB channels where light colors represent positive values and darker colors represent negative values.

Re-lighting

Re-lighting is used to generate new images of the object from the reference image by transferring illumination effects in the target images (Wang et al., 2007; Wen et al., 2003; Zhang et al., 2005). In our framework, we use this technique to render the reference object under illumination conditions of the target image. In the work of Wang et al., 2008, re-lighting was constructed on basis images obtained under various active IR illumination. Our basis images are from the spherical harmonic bases. By Equation (1.6), we obtain two illumination representations of both the reference image and the target image:

$$I_r \approx b_r l_r, \;\; I_t \approx b_t l_t, \tag{1.8}$$

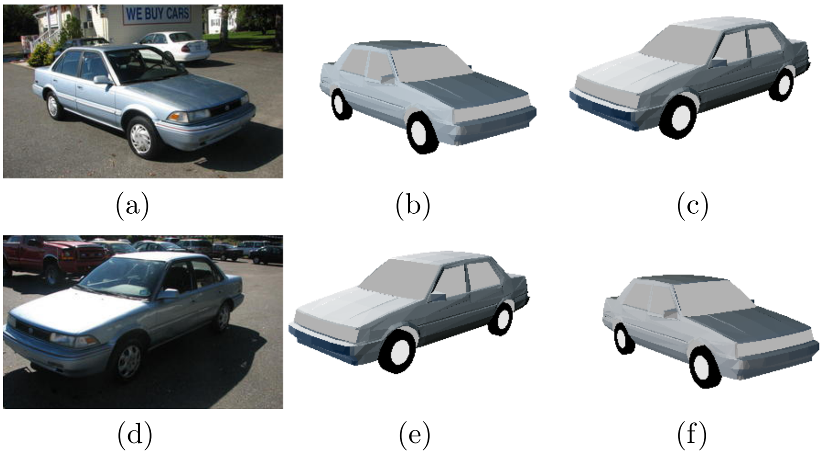

Figure 1.9. Examples of illumination recovery and re-lighting. (a) and (d) are two input images. (b) and (e) are re-rendered images of (a) and (d) after illumination recovery. (c) is the relit image by transferring both pose information and illumination effects from (d) to the 3D vehicle model estimated from (a). (f) is the relit image by transferring both pose information and illumination effects from (a) to the 3D vehicle model estimated from (d). The relit images (c) and (f) are very similar to the re-rendered images (e) and (b), respectively.

where the subscript r denotes the reference object, and subscript t denotes the target object. By re-lighting, we can transfer the illumination effects from the target image to the reference object if they are subject to the same pose:

$$I_{relit} \approx b_r l_t, \tag{1.9}$$

where I_{relit} is the relit images of the reference object with the illumination conditions of the target image.

With this re-lighting technique, we can render an object under any pose and illumination conditions associated with one single input image. Figure 1.9 shows examples of illumination recovery and re-lighting. From the results, we can see that the relit image Figure 1.9(c) is very similar to the re-rendered image Figure 1.9(e) and The relit image Figure 1.9(f) is very similar to the re-rendered image Figure 1.9(b). Therefore, we just compare the relit image with the re-rendered target image to match vehicles in the original reference image and target image despite large variations of pose and illumination.

Vehicle Matching

In order to match two images, we use the normalized matching distance (NMD), defined as

$$NMD = \frac{\Sigma_{i=0}^{n} \| I_{relit}^{i} - I_{t}^{i} \|}{\Sigma_{j=0}^{n} I_{t}^{j}}, \tag{1.10}$$

where I_{relit} is the relit image and I_t is the re-rendered image of target objects. NMD describes the difference between the reference object and the target object, despite the affect of pose and illumination variations. A smaller distance stands for higher similarity, and vice versa.

The vehicle matching algorithm in our framework can be summarized as follows:

1) Determine 3D vehicle models and recover their poses in both the reference image and target image.

2) Estimate reflectance fractions (albedos) from two input images by Equation (1.5).

3) Compute the spherical harmonic basis and illumination coefficients for each input image, respectively, by Equation (1.7) and Equation (1.6).

4) Re-render the target object by Equation (1.8) and re-lighting the reference object by Equation (1.9).

5) Compare the relit image and the re-rendered image by computing the normalized matching distance by Equation (1.10) to match vehicles in the original reference image and target image.

4. Near-IR Illumination

Surfaces of objects have similar reflectance property under active Near-IR illumination and visible illumination. Therefore, the framework proposed above can also be applied on images under Near-IR illumination. One benefit of using Near-IR illumination is that it can provide constant illumination, and work in low-luminance environment without conspicuous light. Besides, specular reflection, which is not considered by our illumination model, is significantly reduced in Near-IR image. The disadvantage of Near-IR illumination is that it does not have any color information. It is not significant to objects like human face, but is important to vehicle, since vehicle has plentiful colors that can be discriminated easily in visible illumination. Figure 1.10 shows two image pairs of Near-IR image and color image of the same scenes, obtained from the work of Fredembach and Süsstrunk, 2009. Near-IR images are very close to gray images under visible spectrum.

Objects may appear unnatural under IR illumination, since many materials do not have the same reflectance fraction under visible or Near-IR

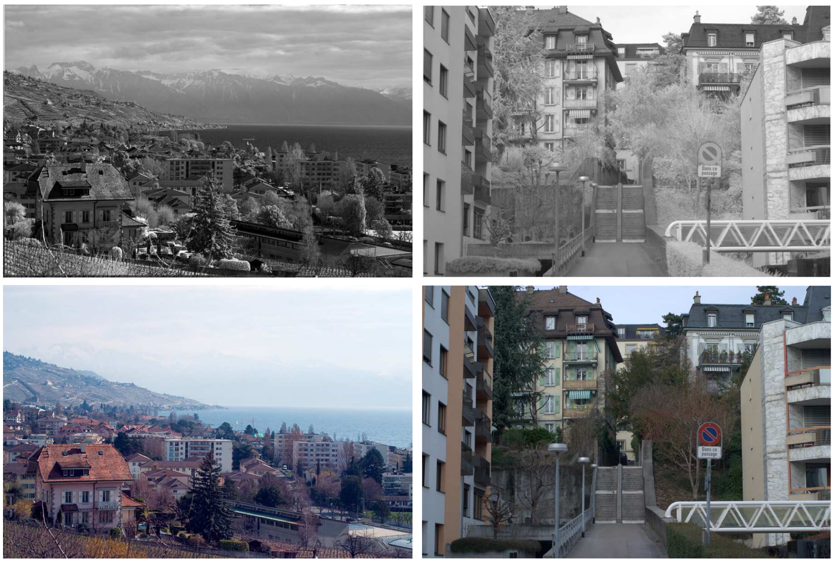

Figure 1.10. Near-IR and color image pairs of the same scenes. Images are from the data in Fredembach and Süsstrunk, 2008.

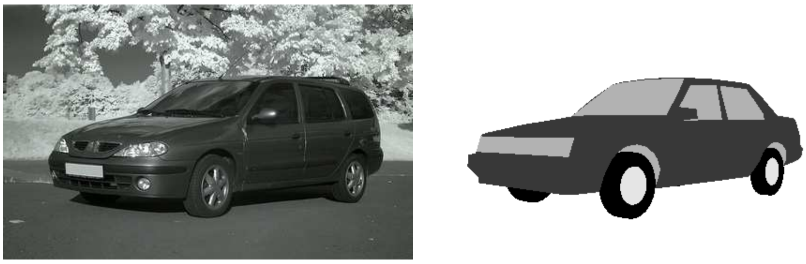

Figure 1.11. An example of Near-IR vehicle image and its estimated reflectance property.

spectrum, e.g., a surface with green color becomes brighter in Near-IR illumination than in the visible spectrum. Therefore, we do not compare images across spectral bands, which means we only match images both under visible illumination, or both under Near-IR illumination. We also assume the diffuse reflectance of Lambertian surface has a constant ratio (i.e., albedo in visible spectrum) in the same spectral band. In visible illumination, the albedo of surface has 3 channels, while in Near-IR illumination, the reflectance property has only one channel. Figure 1.11 shows an example of Near-IR vehicle image and its estimated reflectance property. There is no much difference with image under visible illumination.

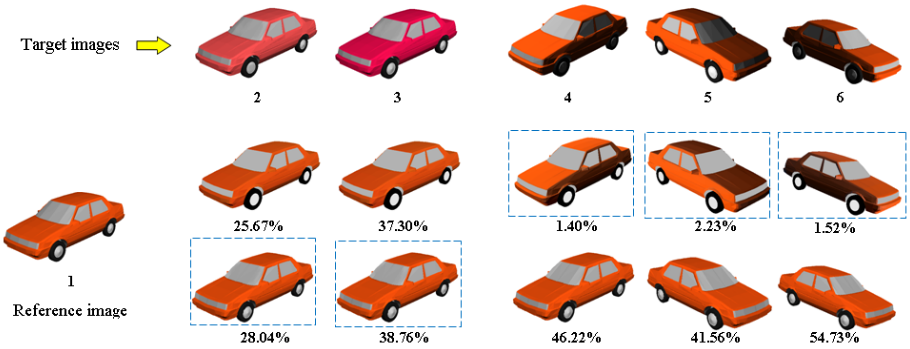

Figure 1.12. Comparison on synthetic data. 1 is the reference image and 2, 3, 4, 5, 6 are target images (1, 4, 5, 6 are from the same car and 2, 3 are from another car). The second row are relit images by our method with their matching distances shown below. The third row are relit images by the method without illumination recovery by Guo et al., 2008. Our method matches the reference image with the correct target images (4, 5, 6) with lower matching distances, while the method without illumination recovery matches it to the wrong target images (2, 3).

5. Experimental Results

In this section, we will evaluate our framework using both synthetic and real data subject to various pose and illumination conditions. The dataset contains N galleries (2 to 7 images in each gallery) of vehicle images. Images in the same gallery are obtained from the same vehicle under different pose and varying lightings. The evaluation schema is to take one probe image to recognize which gallery (object) it matches. We also compare our methods with the method without illumination recovery by Guo et al., 2008 both on synthetic data and real data.

Matching Experiments

Before our recognition experiments, we conduct matching experiments on both synthetic data and real data to show how illumination conditions will affect matching and recognition results. First, we use our 3D car models to synthesize 6 vehicle images rendered by OpenGL with one diffuse light source and global ambient light, as shown in Figure 1.12. Image 1, 2, 3 are rendered by different cars with the same pose. Images 1, 4, 5, 6 are rendered by the same car with different pose and lightings. We match image 1 (as the probe image) to the other 5 images. The matching performances of our method and the method without illumination recovery are shown in the second row and the third row with their matching distances, respectively. From experimental results, we can ob-

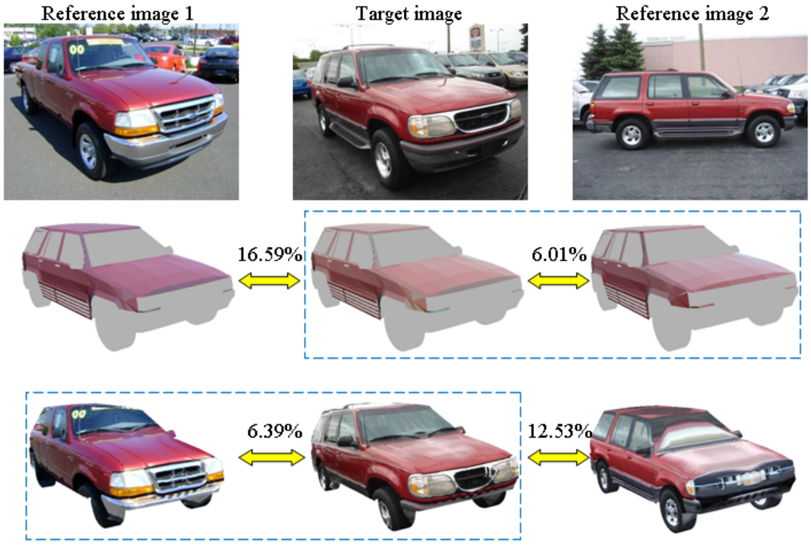

Figure 1.13. Comparison on real data. Original images are shown in the first row. The right image (reference image 2) and the middle image (target image) are from the same SUV, while the left image (reference image 1) is from another vehicle. The results of our method are shown in the second row, and results of the method without illumination recovery are shown in the third row. Numbers above yellow arrows connecting two images are their matching distances.

serve that our method can correctly match image 1 to target images (4, 5, 6) with lower normalized matching distances. However, the method without illumination recovery matches the reference image to wrong target images (2 and 3) due to the effect of illumination. Even in the same illumination condition, there is still a mismatch due to viewpoint variations. For example, images 1 and 5 are under the same illumination condition but taken from different viewpoints. The method without illumination recovery uses symmetry to guess the texture of vehicles. This is not correct because one side of the car is illuminated while the other side is shaded.

Figure 1.13 shows matching experiments on real data. Three input images are in the first row (2 reference images are in the left and right, 1 target image is in the middle). The right image (reference image 2) and the middle image (target image) are from the same SUV, while the left image (reference image 1) is from another vehicle. The matching results of our method are shown in the second row, and the results of the method without illumination recovery are shown in the third row. According to their matching distances experimental results, our method

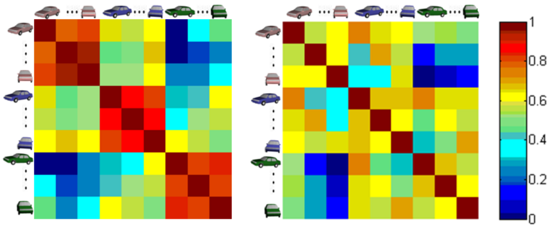

Figure 1.14. Similarity matrices for our method (left) and the method without illumination recovery (right). The x and y coordinates are these 9 images. The value of each entry is illustrated by a color, which can be specified in the color index bar: 1.0 indicates highest similarity and 0.0 indicates lowest similarity. The diagonal has similarity 1.0 where an image matches itself.

correctly matches the reference image 2 and the target image, while the method without illumination recovery fails due to extreme variations of pose and illumination.

Recognition Experiments

Synthetic Data. Our synthetic data for recognition contains 9 image galleries synthesized from 9 vehicle models, each of which consists of 6 images under large variations of pose and lighting. First, in order to test the robustness of our framework, we randomly pick up 9 images belonging to 3 different vehicles from our synthetic dataset, and compute a similarity matrix among these 9 images. Figure 1.14 shows results by our method (left) and the method without illumination recovery (right), where the x- and y-coordinates are these 9 images. We take the first image in each row as probe image and match it to the other images. Each entry of the matrix stands for a similarity between the probe images (y-coordinate) and the target images (x-coordinate). The value of each entry is illustrated by a color, which can be specified in the color index bar: 1.0 indicates highest similarity and 0.0 indicates lowest similarity. The diagonal has similarity 1.0 where an image matches itself. An ideal similarity matrix would have a block diagonal structure with consistently high scores on the main diagonal blocks and consistently low scores elsewhere. From results, we can see that our method provides more distinguishable bands of rows and columns between different vehicles, indicating that it has a better capability to recognize objects subject to various pose and illumination conditions. However, there is no distinct diagonal block in the right matrix, which clearly suffers from the variation of illumination.

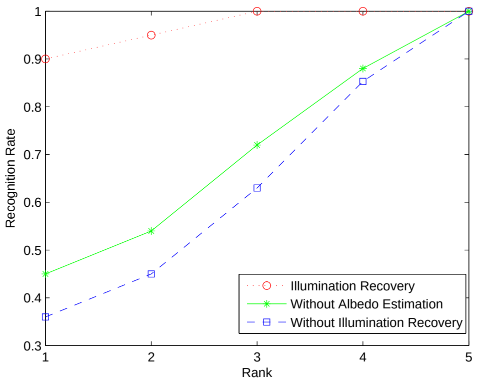

Figure 1.15. Recognition results on synthetic data: The x-coordinate denotes the size of each gallery from 1 to 5, and the y-coordinate denotes the recognition rates.

For recognition, we conduct our recognition experiments on 9 image galleries (6 images in each gallery) with rank from 1 to 5 (the number of images in each gallery is from 1 to 5). Here, we test the following algorithms: (1) the method without illumination recovery by Guo et al., 2008, (2) the method using average texture of vehicle body as albedo (no albedo estimation), and (3) our framework. Figure 1.15 shows the recognition results, where the x-coordinate denotes the size of each gallery from 1 to 5, and the y-coordinate denotes the recognition rates. From the results, we can see that our framework always achieves the highest recognition rates among these methods. Besides, the performance of our framework is robust to the size of each gallery while the method without illumination recovery does not perform well when the size of the gallery is very small. This is because they do not consider the illumination variations, and thus it needs many more images to discern illumination changes. Furthermore, their method is also more restricted under some extreme poses, for example, an image taken from the front of a vehicle can never provide texture information on two sides and the back. However, our framework tremendously improves the recognition performance in these aspects, by which we can still recognize vehicles under limited inputs with unconstrained pose and illumination conditions.

Real Data. Our real data consists of 30 image galleries captured from 30 vehicles (24 under visible illumination, and 6 under Near-IR illumination). For visible illumination, each gallery has 7 images under various viewpoints and lightings, while for IR illumination, there are 5 images in each gallery. All 6 galleries in Near-IR illumination are from

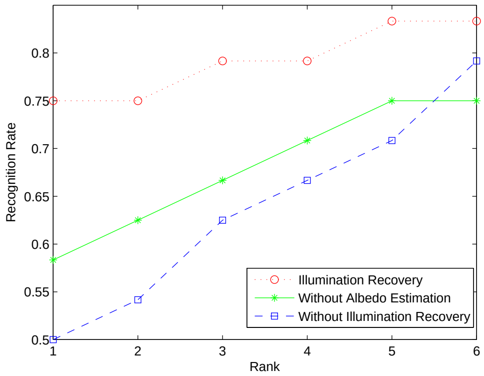

Figure 1.16. Recognition result on real data (visible spectrum).

the same category of vehicle, since we have a small-size dataset. The image resolution is from 310 × 233 to 640 × 480. Our experiment is conducted by the same schema as we did on synthetic data, and we do not mix up color images and Near-IR images since they apparently different.

Figure 1.16 shows the recognition result of real data under visible illumination by the following methods: the method without illumination recovery, the method using average texture as albedo, and our framework with illumination recovery under different ranks. And Figure 1.17 shows the Near-IR part. The x-coordinate denotes the rank (number of images involved per gallery), and the y-coordinate denotes the recognition rates. From results, we can see that by illumination recovery, the recognition results of our methods are significantly improved and stable when the number of images involved per gallery changes. However, the other two methods use semantic ownership of vehicle model and the symmetry of vehicle body to represent texture information. These are not accurate due to the effect of illumination conditions, especially when the size of the gallery is small.

6. Conclusion

We have detailed a 3D model-driven framework to match vehicles subject to large variations of both pose and lightings in visible or Near-IR illumination. By estimated pose and albedo, the illumination condition can be approximately recovered by using spherical harmonics representation. This will also allow us to re-light the reference object under any target condition of pose and illumination. Based on algorithmic compo-

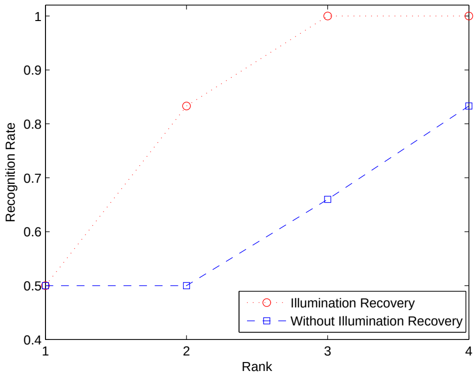

Figure 1.17. Recognition result on real data (Near-IR spectrum).

nents, matching between two input images is conducted in a common domain by computing the distance from the re-rendered images. Experimental results demonstrate that our framework has improved the matching and recognition performance, especially when objects are under both large pose and illumination variations. Besides vehicles, our framework can also be generalized to handle other types of objects.

There are also some limitations in our framework. When the approximated fitting is coarse and inaccurate due to non-standard types of vehicles and camera distortion, the recognition suffers and reduces to vehicles subject to the same category. Besides, the real illumination condition is much more complicated than the current assumption, i.e., a dominating light source in infinity. For example, in indoor auto-shows sometimes the highlight area can affect the recognition results. The first limitation can be further improved by the morphable model, which has been successfully applied in face recognition; while the second limitation requires more techniques in the illumination model, which are the future tasks we expect to undertake.

Acknowledgments

Hong Qin and Tingbo Hou (Stony Brook University)'s research reported in this chapter is partially supported by NSF grants: IIS-0949467, IIS-0710819, and IIS-0830183.

Notes

1. http://shape.cs.princeton.edu/benchmark/

References

Basri, R. and Jacobs, D. W. (2003). Lambertian reflectance and linear subspaces. *TPAMI*, 25(2):218–233.

Biswas, S., Agrawal, G., and Chellappa, R. (2007). Robust estimation of albedo for illumination-invariant matching and shape recovery. In *ICCV*.

Debevec, P., Hawkins, T., Tchou, C., Duiker, H.-P., Sarokin, W., and Sagar, M. (2000). Acquiring the reflectance field of a human face. In *SIGGRAPH*, pages 145–156.

Fergus, R., Perona, P., and Zisserman, A. (2006). Object class recognition by unsupervised scale-invariant learning. In *CVPR*, pages 264–271.

Fredembach, C. and Süsstrunk, S. (2008). Colouring the near infrared. In *Proceedings of the IS&T/SID 16th Color Imaging Conference*, pages 176–182.

Fredembach, C. and Süsstrunk, S. (2009). Illuminat estimation and detection using near infrared. In *SPIE/IS&T Electronic Imaging*, volume 7250.

Gardner, W. F. and Lawton, D. T. (1996). Interactive model-based vehicle tracking. *TPAMI*, 18(11):1115–1121.

Guo, Y., Rao, C., Samarasekera, S., Kim, J., Kumar, R., and Sawhney, H. (2008). Matching vehicles under large pose transformations using approximate 3d models and piecewise mrf model. In *CVPR*.

Guo, Y., Shan, Y., Sawhney, H., and Kumar, R. (2007). Peet: prototype embedding and embedding transition for matching vehicles over disparate viewpoints. In *CVPR*.

Hou, T., Wang, S., and Qin, H. (2009). Vehicle matching and recognition under large variations of pose and illumination. In *CVPR Workshop on Object Tracking and Classification Beyond and in the Visible Spectrum*, pages 24–29.

Ji, Q. and Yang, X. (2002). Real time 3d face pose discrimination based on active ir illumination. In *ICPR*, volume 4, pages 310–313.

Kim, Z. and Malik, J. (2003). Fast vehicle detection with probabilistic feature grouping and its application to vehicle tracking. In *ICCV*, pages 524–531.

Koller, D., Daniilidis, K., and Nagel, H.-H. (1993). Model-based object tracking in monocular image sequences of road traffic scenes. *IJCV*, 10(3):257–281.

Li, S. Z., Chu, R., Liao, S., and Zhang, L. (2007). Illumination invariant face recognition using near-infrared images. *TPAMI*, 29(4):627–639.

Morris, N. J. W., Avidan, S., Matusik, W., and Pfister, H. (2005). Statistics of infrared images. In *CVPR*.

Murase, H. and Nayar, S. K. (1995). Visual learning and recognition of 3-d objects from appearance. *IJCV*, 14(1):5–24.

Novotny, P. M. and Ferrier, N. J. (1999). Using infrared sensors and the phong illumination model to measure distances. In *ICRA*, pages 1644–1649.

Pan, K., Liao, S., Zhang, Z., Li, S. Z., and Zhang, P. (2007). Part-based face recognition using near infrared images. In *CVPR*, pages 1–6.

Ramamoorthi, R. and Hanrahan, P. (2001). A signal-processing framework for inverse rendering. In *SIGGRAPH*, pages 117–128.

Romdhani, S., Blanz, V., and Vetter, T. (2002). Face identification by fitting a 3d morphable model using linear shape and texture error functions. In *ECCV*, pages 3–19.

Sato, Y., Wheeler, M. D., and Ikeuchi, K. (1997). Object shape and reflectance modeling from observation. In *SIGGRAPH*, pages 379–388.

Shan, Y., Sawhney, H. S., and Kumar, R. (2005). Vehicle identification between non-overlapping cameras without direct feature matching. In *ICCV*, pages 378–385.

Shan, Y., Sawhney, H. S., and Kumar, R. (2008). Unsupervised learning of discriminative edge measures for vehicle matching between non-overlapping cameras. *TPAMI*, 30(4):700–711.

Wang, O., Davis, J., Chuang, E., Rickard, I., de Mesa, K., and Dave, C. (2008). Video relighting using infrared illumination. *Computer Graphics Forum (Proceedings Eurographics)*, 27(2).

Wang, Y., Liu, Z. C., Hua, G., Wen, Z., Zhang, Z. Y., and Samaras, D. (2007). Face re-lighting from a single image under harsh lighting conditions. In *CVPR*.

Wen, Z., Liu, Z., and Huang, T. S. (2003). Face relighting with radiance environment maps. In *CVPR*, pages 158–165.

Yu, Y., Debevec, P., Malik, J., and Hawkins, T. (1999). Inverse global illumination: Recovering reflectance models of real scenes from photographs. In *SIGGRAPH*, pages 215–224.

Zhang, L. and Samars, S. (2006). Face recognition from a single training image under arbitrary unknown lighting using spherical harmonics. *TPAMI*, 28(3):351–363.

Zhang, L., Wang, S., and Samaras, D. (2005). Face synthesis and recognition from a single image under arbitrary unknown lighting using a spherical harmonic basis morphable model. In *CVPR*, pages 209–216.

Zhao, S. and Grigat, R.-R. (2006). Robust eye detection under active infrared illumination. In *ICPR*, volume 4, pages 481–484.

Zhu, Z., Ji, Q., Fujimura, K., and Lee, K. (2002). Combining kalman filtering and mean shift for real time eye tracking under active ir illumination. In *ICPR*, volume 4, pages 318–321.

Zou, X., Kittler, J., and Messer, K. (2005). Face recognition using active near-ir illumination. In *BMVC*.