Vehicle Matching and Recognition under Large Variations of Pose and Illumination

Tingbo Hou
Computer Science Department
Stony Brook University
thou@cs.sunysb.edu

Sen Wang
Kodak Research Laboratories
Eastman Kodak Company
sen.wang@kodak.com

Hong Qin
Computer Science Department
Stony Brook University
qin@cs.sunysb.edu

Abstract

Matching vehicles subject to both large pose transformations and extreme illumination variations remains a technically challenging problem in computer vision. In this paper, we develop a new and robust framework toward matching and recognizing vehicles with both highly varying poses and drastically changing illumination conditions. 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. We compare the relit image and the re-rendered target image to match vehicles in the original reference 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. 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.

1. Introduction

Object matching and recognition remain an important and long-term task with continuing interest from computer vision with significant applications in security, surveillance, and robotics. Many types of representations have been employed 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 outdoor conditions such as highly varying pose and severely changing illumination, the problem becomes extremely challenging. As shown in Fig. 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 as shape and texture are faint in this case.

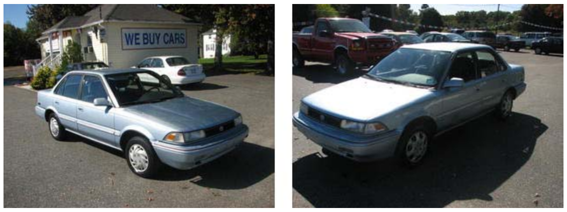

Figure 1. Images of the same vehicle taken from different viewpoints and lightings.

Currently, the most popular approaches in object recognition focus on the appearance-based methods [11, 4] and the model-based methods [5, 18]. 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 Principal Component Analysis (PCA) subspace or feature subspace. In practice, technical issues arise from appearance variation due to different poses 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: (1) the 3D model can precisely fit to the input images; (2) pose estimation is accurate enough. To estimate the appearance of objects, global and local clues have been used to simulate the 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 Fig. 1.

The primary contribution of this paper lies in a novel and robust framework toward vehicle matching and recognition, which can handle large pose transformations and illumination variations simultaneously. Our vehicle matching framework is shown in Fig. 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 albedos 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

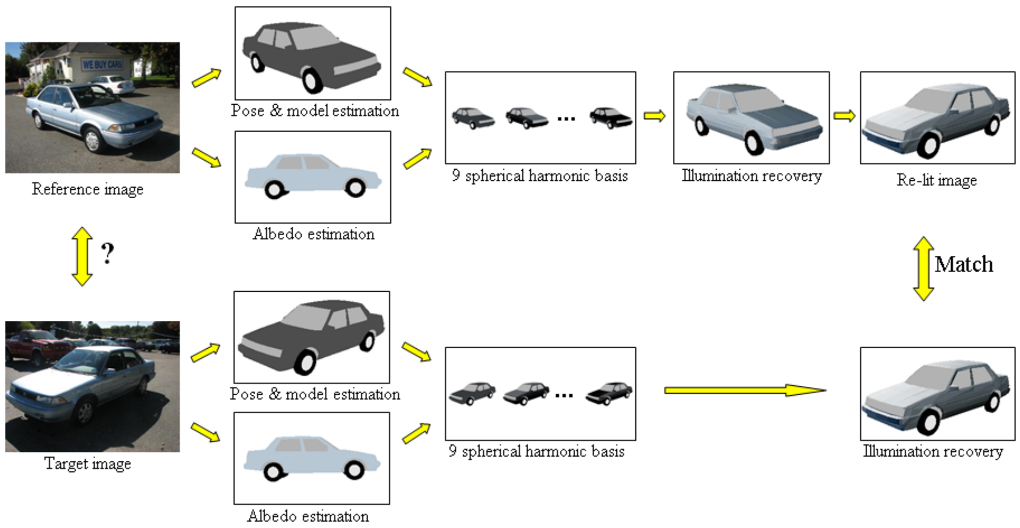

Figure 2. Our vehicle matching framework.

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. Although our current object domain in this paper is vehicles, this computational framework can also be generalized to other types of objects.

2. Previous Work

Vehicle matching has been studied in many areas of computer vision, with many different purposes such as detection, tracking, and recognition. Koller et al. [10] represent vehicles by a general 3D model parameterized by 12 length parameters. Their method needs to calibrate a moving plane from video sequences. Some other work on vehicle detection [9] and tracking [16] is proposed by using block matching, template matching, or a simple sedan model. They are all based on video sequences, not one single image. In [17], 3D computer graphic models are employed to recognize vehicles. It performs well for one vehicle class, but is limited for several similar vehicle classes. The Hidden Markov Models (HMM) has been employed in [8] to classify shadows from moving vehicles, which obstructs robust visual tracking. For vehicle identification, Shan et al. [14] use 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 extended work was done in [7, 15] for vehicle matching. However, their methods do not consider the texture and illumination of vehicles. Guo et al. [6] propose a model-based approach to match vehicles. They use approximate 3D models to handle pose transformation and a piecewise MRF model to guess texture of occluded parts. However, their method has limitations on sensitive model fitting and varying illumination.

For illumination recovery, recent research [1, 12] shows that any image under arbitrary illumination conditions can be approximately represented by a linear combination of the spherical harmonic basis. Spherical harmonics are a set of functions that form an orthonormal basis for the set of all square-integrable functions defined on the unit sphere. Zhang et al. [18] successfully applied spherical harmonics in face recognition and synthesis with a 3D morphable model, which is difficult to build.

3. Vehicle Matching Framework

In this section, we will introduce our framework of vehicle matching under various poses and lightings.

3.1. 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 large truck. Some 3D vehicle models are shown in Fig. 3. Unlike the approach in [6], which requires that 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 albedo for each vehicle; (2) the removed parts may have complex cases, which have no benefits for matching. 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.

Figure 3. Some 3D vehicle models.

Considering the fact that pose estimation is easily trapped into local minimum in the searching space, we select 3 different initial poses for each vehicle model with reasonable projection. For each model, we compute edge maps under these 3 initial fittings and use chamfer distance [14] to measure the similarity with edge maps of the original input image, as shown in Fig. 4. Finally, we select the top 2 matched models as candidates for the next step.

3.2. Pose Recovery

Pose recovery has been extensively studied in recent years, with many effective methods proposed. Among them, the Iterative Closest Point (ICP) algorithm is a widely used method [2, 13]. During fitting a 3D model to a 2D image, the transformation is not linear or approximately linear. Thus, we cannot use ICP directly in our framework. Here, we design new strategies in fitting as follows:

1) For each of 2 candidate vehicle models, we search for the next better pose in 3D translation and rotation, respectively, by discrete samplings with adaptive distance $d \cdot s$, where d is the average closest point distance in ICP, and s is a fixed distance. More specifically, 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. Thus, we can control the speed of searching 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.

2) We will stop searching when the average closest point distance d reaches our 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, we let the sampling distance jump to $D \cdot s$, where D is a large factor to pull the searching out of the local minimum.

3) The best 3D object model is selected with minimal average closest point distance from candidate models.

Figure 4 shows an example of pose recovery, where green edges are from the input image and red edges are from the projected edge map of the 3D vehicle model.

3.3. Estimation of Albedo

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 [3]. In previous work of applying spherical harmonics [18], 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 3 channels, respectively.

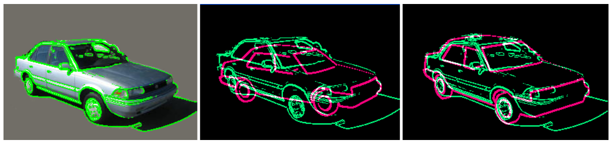

Figure 4. Pose estimation. Edge detection in the original image is shown on the left. Initial fitting between 3D vehicle model and the original image is shown in the middle. The fitting result is shown on the right. Green edges come from the original input image and red edges come from the 3D vehicle model.

For Lambertian objects, the diffused component of the surface reflection satisfies Lambert's Cosine Law, given by $I = \rho max(n^T s, 0)$, where I is the pixel intensity, s is the light source direction, ρ is 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 Lambert's 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: $n^T \rho s = I$. 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): $N \rho_r s = I_r$, 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. 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 between 0 and 1, and s is the normalized direction vector whose length equals 1, we can compute ρ_r by

$$\rho_r = \frac{|\rho_r s|}{|s|} = |\rho_r s|. \quad (1)$$

Similarly, we can compute albedo in green channel ρ_g and albedo in blue channel ρ_b. Figure 5 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.

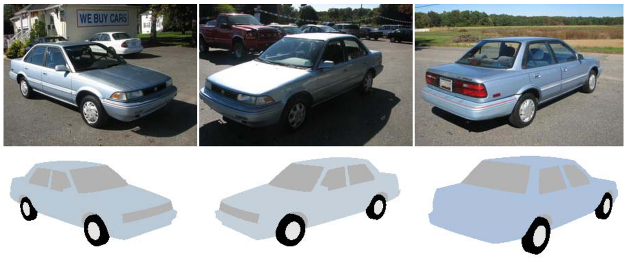

Figure 5. Albedo estimation. The original input images are shown in the first row and estimated albedos are shown in the second row. The two left images come from the same car and the right-most image comes from another car.

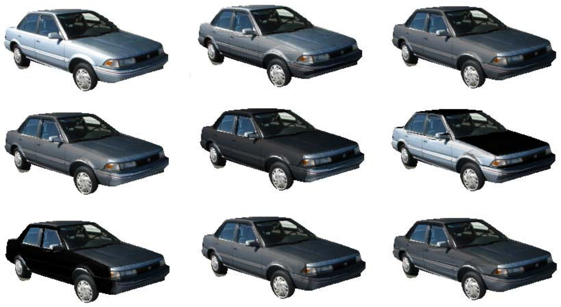

Figure 6. An example of the first 9 spherical harmonic basis images with RGB 3 channels. Light colors represent positive values and darker colors represent negative values.

3.4. Illumination Recovery

As described in [1, 12], any image under arbitrary illumination conditions can be approximately represented by a linear combination of spherical harmonic basis as

$$I \approx bl, \tag{2}$$

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. They are the sphere analog of the Fourier basis on the line or circle. The first 9 spherical harmonic basis can be computed by using normals and albedo in each vertex of the objects [1, 12].

In our framework, we use unified albedo for the body of the vehicle model that is estimated in Section 3.3. 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 images first, and then estimate the illumination coefficients l by using the method of least squares in Eq. 2. Figure 6 shows an example of the first 9 spherical harmonic basis images with RGB 3 channels where light colors represent positive values and darker colors represent negative values.

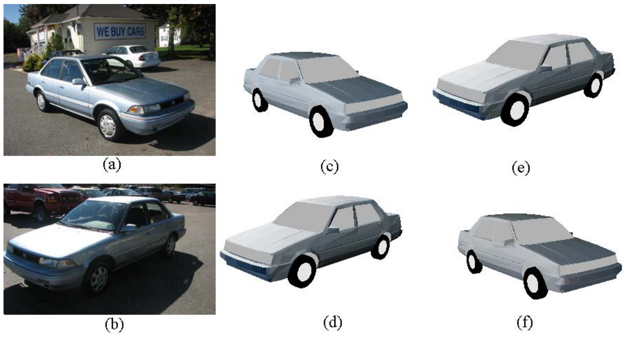

Figure 7. Examples of illumination recovery and re-lighting. (a) and (b) are two input images. (c) and (d) are re-rendered images of (a) and (b) after illumination recovery. (e) is the relit image by transferring both pose information and illumination effects from (b) to the 3D vehicle model estimated from (a). (f) is the relit image by transferring both pose information and illumination effects from (a) to the 3D vehicle model estimated from (b). The relit images (e) and (f) are very similar to the re-rendered images (d) and (c), respectively.

3.5. 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. In our framework, we use this technique to render the reference object under illumination conditions of the target image. By Eq. 2, we obtain two illumination representations of both the reference image and the target image: $I_r \approx b_r l_r$, $I_t \approx b_t l_t$, 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:

$$I_{relit} \approx b_r l_t, \tag{3}$$

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 7 shows examples of illumination recovery and re-lighting. From the results, we can see that the relit image Fig. 7(e) is very similar to the re-rendered image Fig. 7(d) and the relit image Fig. 7(f) is very similar to the re-rendered image Fig. 7(c). 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.

3.6. Vehicle Matching

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

$$NMD = \frac{\Sigma_{i=0}^{n} \| I_{relit}^{i} - I_t^{i} \|}{\Sigma_{j=0}^{n} I_t^{j}}, \tag{4}$$

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 effect 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) Recover pose information and estimate 3D vehicle models in both the reference image and target image as described in Section 3.1 and 3.2.

2) Estimate albedos from two input images by Eq. 1.

3) Compute the spherical harmonic basis in each input image, and then estimate illumination coefficients by Eq. 2.

4) Re-render the target object by recovered shape, albedo, and illumination parameters and re-lighting the reference object by Eq. 3.

5) Compare the relit image and the re-rendered image by computing the normalized matching distance in Eq. 4 to match vehicles in the reference image and target image.

4. Experimental Results

In this section, we will evaluate our framework using both synthetic and real data subject to various pose and illumination conditions and compare our methods with the method without illumination recovery [6].

4.1. 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 three new 3D car models to synthesize 6 vehicle images rendered by OpenGL with one diffuse light source and global ambient light, as shown in Fig. 8 (the left-most image and other 5 images in the first row). Images 1, 2, and 3 are synthesized by different car models with different texture. Images 1, 4, 5, and 6 are synthesized by the same car model with different pose and lightings. We match image 1 (as the probe image) to all other 5 images. The matching performances of our method and the method without illumination recovery are shown in the second and third rows. From experimental results, we can observe that our method can successfully match image 1 to images 4, 5, and 6 with perfect normalized matching distances. However, the method without illumination recovery fails in this experiment due to the effect of illumination. It matches image 1 with image 2, while image 2 comes from a different vehicle. 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 9 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 [6] are shown in the third row. According to the experimental results, our method successfully matches the reference image 2 and the target image while the method without illumination recovery fails due to extreme variations of pose and illumination.

4.2. Recognition Experiments

Our real data consists of 24 image galleries captured from 24 vehicles. Each gallery has 7 images under various viewpoints and lightings. The image resolution is from 310×233 to 640×480. Our experiment is conducted by the same schema as we did on synthetic data.

Figure 10 shows the recognition results of the following methods: the method without illumination recovery, the method using average texture as albedo, and our framework with illumination recovery under different ranks. 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.

5. Conclusion

We have detailed a new framework to match vehicles subject to large variations of both pose and illumination. By estimating pose and albedo map, the illumination condition can be accurately recovered by using spherical harmonics representation. This will also allow us to re-light the reference object under any target condition such as pose and illumination. Experimental results demonstrate that our framework has significantly improved the matching and recognition performance, especially when objects are under both large pose and illumination variations. For the future, we will further improve our illumination model with specularity and shadow components and apply our method to additional objects such as the human face and body.

References

[1] R. Basri and D. Jacobs. Lambertian reflectance and linear subspaces. *PAMI*, 25(2):218–233, 2003.

[2] P. Besl and N. McKey. A method for registration of 3d. *PAMI*, 14(2):239–256, 1992.

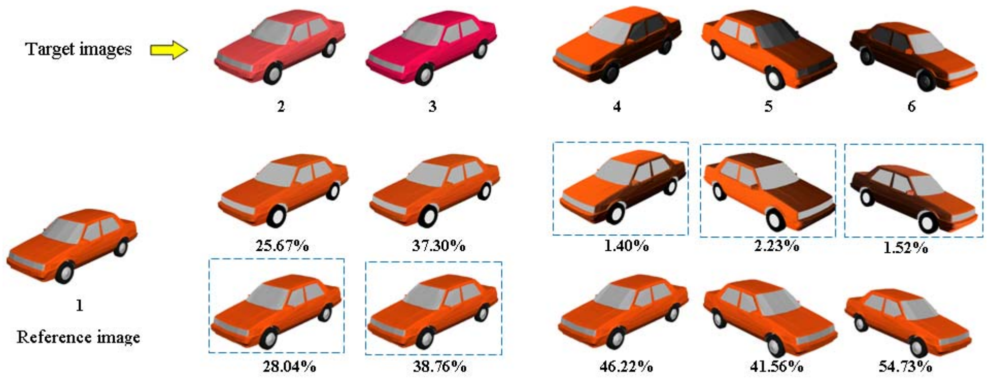

Figure 8. Comparison on synthetic data. 1 is the reference image and 2, 3, 4, 5, 6 are target images (1, 4, 5, and 6 are synthesized from the same car model and 2, 3 are from another two car models). In the second row, there are results of our method by transferring the pose information and illumination conditions from target images (2, 3, 4, 5, and 6) to the car model in the reference image 1 with their matching scores, which are computed between the relit images in the second row to the target images in the first row. In the third row, there are results of the method without illumination recovery in [6]. The percentage numbers are the normalized matching distance (NMD) defined in Eq. 4 and a smaller number stands for higher similarity between the reference image and the target image, and vice versa.

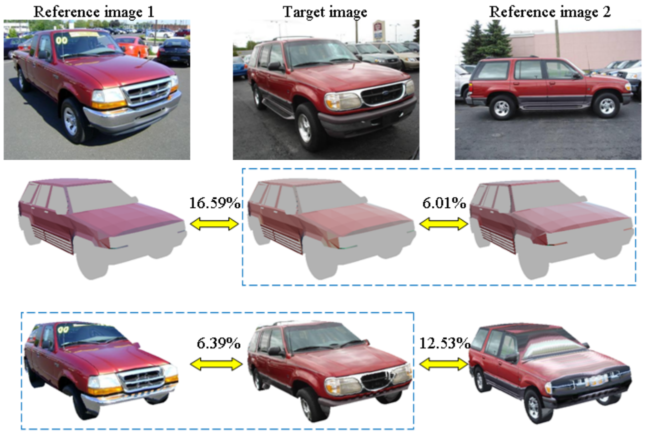

Figure 9. Comparison on real data. Original images are shown in the first row. The right image (reference image 2) and the middle image (target image) are from the same SUV, while the left image (reference image 1) is from another vehicle. The results of our method are shown in the second row and results of the method without illumination recovery are shown in the third row. The numbers are the normalized matching distance defined in Eq. 4.

[3] S. Biswas, G. Aggarwal, and R. Chellappa. Robust estimation of albedo for illumination-invariant matching and shape recovery. In *ICCV*, 2001.

[4] R. Fergus, P. Perona, and A. Zisserman. Object class recognition by unsupervised scale-invariant learning. In *CVPR*, 2006.

[5] W. Gardner and D. Lawton. Interactive model-based vehicle tracking. *PAMI*, 18(11):1115–1121, 1996.

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

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

[8] J. Kato, T. Watanabe, S. Joga, J. Rittscher, and A. Blake. An hmm-based segmentation method for traffic monitoring movies. *PAMI*, 24(9):1291–1296, 2002.

[9] Z. Kim and J. Malik. Fast vehicle detection with probabilistic feature grouping and its application to vehicle tracking. In *ICCV*, 2003.

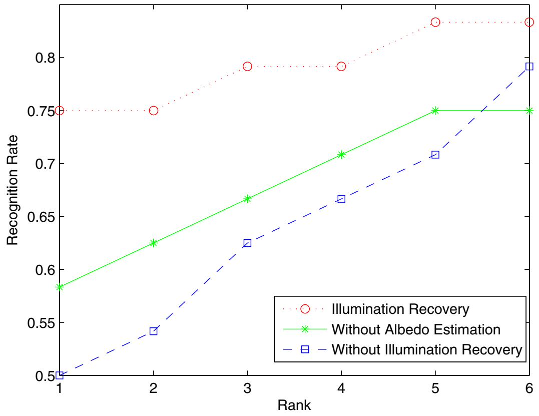

Figure 10. Recognition results on real data.

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

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

[12] R. Ramamoorthi and P. Hanrahan. An efficient representation for irradiance environment maps. In *SIGGRAPH*, pages 497–500, 2001.

[13] S. Rusinkiewicz, O. Hall-Holt, and M. Levoy. Real-time 3d model acquisition. *ACM Trans. Graph.*, 21(3):438–446, 2002.

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

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

[16] L. Stefano and E. Viarani. Vehicle detection and tracking using the block matching algorithm. In *IEEE Int. Multiconference on Circuits, Systems, Communications and Computer*, 1999.

[17] M. Stevens and J. Beveridge. Using multisensory occlusion reasoning in object recognition. In *CVPR*, 1997.

[18] L. Zhang, S. Wang, and D. Samaras. Face synthesis and recognition from a single image under arbitrary unknown lighting using a spherical harmonic basis morphable model. In *CVPR*, 2005.