A Joint Illumination and Shape Model for Visual Tracking

Amit Kale and Christopher Jaynes*
Ctr. for Visualization and Virtual Environments and Department of Computer Science
University of Kentucky
{amit,jaynes}@cs.uky.edu

Abstract

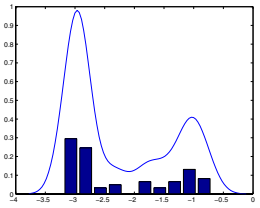
Visual tracking involves generating an inference about the motion of an object from measured image locations in a video sequence. In this paper we present a unified framework that incorporates shape and illumination in the context of visual tracking. The contribution of the work is twofold. First, we introduce a a multiplicative, low dimensional model of illumination that is defined by a linear combination of a set of smoothly changing basis functions. Secondly, we show that a small number of centroids in this new space can be used to represent the illumination conditions existing in the scene. These centroids can be learned from ground truth and are shown to generalize well to other objects of the same class for the scene. Finally we show how this illumination model can be combined with shape in a probabilistic sampling framework. Results of the joint shape-illumination model are demonstrated in the context of vehicle and face tracking in challenging conditions.

1. Introduction

Visual tracking involves generating an inference about the motion of an object from measured image locations in a video sequence. Unfortunately, this goal is confounded by sources of image appearance change that are only partly related to the position of the object in the scene. For example, unknown deformations, changes in pose of the object, or changes in illumination can cause a template to change appearance over time and lead to tracking failure.

Shape change for rigid objects can be captured by a low-dimensional shape space under a weak perspective assumption. Thus tracking can be considered as the statistical inference of this low-dimensional shape vector. This interpretation forms the basis of several tracking algorithms including the well-known Condensation algorithm [7] and its variants. A similarly concise model is required if we are to robustly estimate illumination changes in a statistical tracking framework while avoiding undue increase in the dimensionality of the problem. This is the topic of this paper.

(a)



(b)

Figure 1. Tracking a car across drastic illumination change. (a) A template constructed for the vehicle in sunlight will change appearance as it enters shadow and traditional shape-tracking fails. (b) The histogram of the zeroth coefficient of our illumination model. This work shows how the modes of these distributions are sufficient to accurately track though both shape and illumination change.

The study of appearance change as a function of illumination is a widely studied area in computer vision [2, 11, 1]. These methods focus on accurate models of appearance under varying illumination and their utility for object recognition. However they typically require an explicit 3-D model of the object which somewhat limits their application to surveillance applications. A general yet low-dimensional parameterization of illumination has thus far been elusive in a tracking context.

In this work we focus on the problem of tracking objects through simultaneous illumination and shape change. Examples include monitoring vehicles that move in and out of shadow or tracking a face as it moves thorough different lighting conditions in an indoor environment. The approach is intended for use in traditional video surveillance and monitoring tasks where a large number of illumination samples of each object to be tracked are unavailable [6] and features that are considered to be invariant to illumination are known to be unreliable [2].

The contribution of the work is twofold. First, we intro-

*This work was funded by NSF CAREER Award IIS-0092874 and by Department of Homeland Security

duce a a multiplicative, low dimensional model of illumination that is computed as a linear combination of a set of Legendre functions. Such a multiplicative model can be interpreted as an approximation of illumination image as discussed in Weiss [12]. Although the model is not intended to be applied for recognition tasks under differing illumination, it is sufficient to capturing appearance variability for improved tracking. The Legendre coefficients together with the shape vectors define a joint shape-illumination space. Our approach then is to estimate the vector in this joint space that best transforms the template to the current frame. This is in contrast to approaches that adapt the template over time by modifying a continuously varying density [3, 13]. Direct adaption of the template requires careful selection of adaption parameters to avoid problems of drift [10].

In alternative formulation of the problem Freedman and Turek [4] introduce an illumination invariant approach to computing optic-flow that can be used to localize object templates. The method was shown to be quite robust at tracking objects through shadows. However it is computationally expensive and it is unclear how known system dynamics can be integrated within the approach. We do not seek illumination invariance but instead estimate the illumination changes using our model as part of the tracking process. However, use of illumination invariant optic flow as a low-level primitive could be used in combination with the work here to inform the shape space sampling distributions and is the subject of future work.

When using this joint shape illumination space for tracking, it is no longer obvious how this space should be sampled. For example, Figure 1a shows a vehicle that moves from bright sunlight to shadow. Because this transition can occur instantaneously between frames, the smoothness assumptions that are used to derive the sampling distribution for shape cannot are often violated for the illumination component. Furthermore, the additional degrees-of-freedom that it are required to model illumination can lead to decreased robustness at runtime or require an inordinate number of tracking samples in each frame. However, we discover the surprising result that a small number of centroids extracted from the underlying distributions of our illumination coefficients are often adequate to represent the influence of most of the illumination conditions existing in the scene. Figure 1b shows a distribution of the zeroth order coefficient in our model for the car moving through the scene in Figure 1a. In Section 3.1 we discuss how important modes of these distributions are extracted and used to track through drastic illumination changes such as these.

2. A Multiplicative Model of Appearance Change due to Illumination

The image template throughout the tracking sequence can be expressed as:

$$U_t(x,y) = L_t(x,y)R(x,y) \tag{1}$$

where $L_t(x,y)$ denotes the illumination image in frame t and $R(x,y)$ denotes a fixed reflectance image [12]. Thus if the reflectance image of the object is known, tracking becomes the problem of estimating the illumination image and a shape-vector.

Of course, the reflectance image is typically unavailable and the illumination image can only be computed modulo the illumination contained in the image template shown in Equation 2.

$$L_t = \tilde{L}_t L_0 R(x,y) \tag{2}$$

where L_0 is the initial illumination image and $\tilde{L}_t$ is the unknown illumination image for frame t.

Our proposed model of appearance change, then, is simply the product of the input image with a function $f_t(x,y)$ that approximates L_t and is defined over the image domain, $P \times Q$. A naive way of compensating for appearance change then is to allow each $f(x,y), x = 1, \cdots, P, y = 1, \cdots, Q$ to vary independently. However, it is known that for a convex Lambertian object, the change in appearance of neighboring pixels is not independent and the excessive additional degrees-of-freedom can make the tracking problem intractable.

Instead we construct the illumination compensation image f from a linear combination of a far lower dimensional set of n basis functions. In order to be useful, the basis functions must be both both orthogonal in the 2D image domain and straightforward to compute. Furthermore they must be capable of spanning most of the appearance changes in the template due to illumination. For the work here we utilize the Legendre polynomial basis although any other polynomial basis will suffice. To give an idea about the type of variation the basis supports, Figure 2 shows the Legendre basis of order three.

Let $p_n(x)$ denote the nth Legendre basis function. Then, for a given set of coefficients $\Lambda = [\lambda_0, \cdots, \lambda_{2n}]^T$, the scaled intensity value at a pixel is computed as:

$$\begin{aligned}\hat{U}(x,y) = (\frac{1}{2n+1}(\lambda_0 + \lambda_1 p_1(x) + \cdots + \lambda_n p_n(x) + \\ \lambda_{n+1} p_1(y) + \cdots + \lambda_{2n} p_n(y)) + 1) U(x,y)\end{aligned} \tag{3}$$

For purposes of notation, we will denote the effect of Λ on the image as

$$\mathbf{\Delta}\Lambda U \equiv U \otimes \mathbf{P}\Lambda + U \tag{4}$$

where

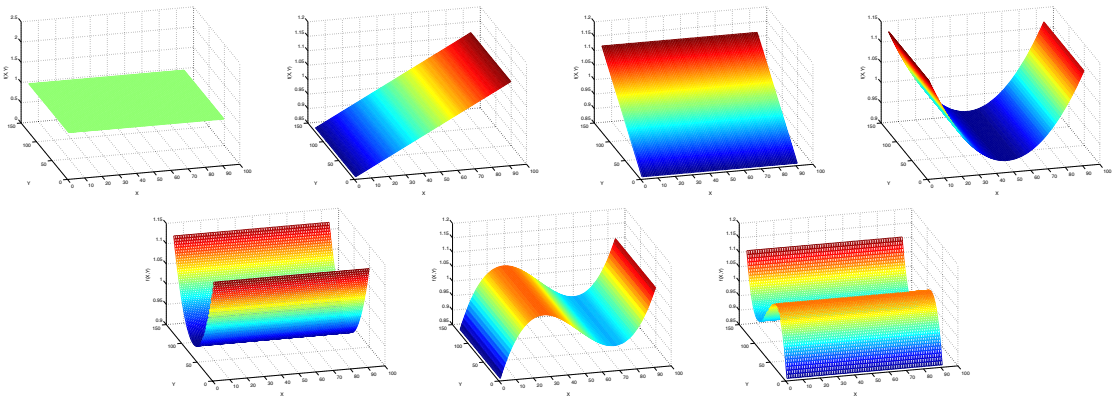


Figure 2. First seven Legendre basis functions used to track illumination change in an image template.

$$\mathbf{P} = \begin{bmatrix} \frac{1}{2n+1}p_0 & \cdots & \frac{1}{2n+1}p_n(y_1) \\ \vdots & \vdots & \vdots \\ \frac{1}{2n+1}p_0 & \cdots & \frac{1}{2n+1}p_n(y_{PQ}) \end{bmatrix}. \quad (5)$$

We define $\otimes$ as an operator that scales the rows of P with the corresponding element of U written as a vector. Given an input template T and an image U, the Legendre coefficients that minimize the error between $\mathbf{\Delta}.\Lambda U$ and T can be computed by solving the least squares problem,

$$U \otimes P\Lambda \approx T - U. \quad (6)$$

Each of the basis functions is scaled by a particular choice of Λ_i and then linearly combined using Equation 4 to derive a illumination image.

Figure 3 demonstrates how how this low-dimensional set of Legendre polynomials can accommodate illumination change. Figure 3a is an input template and Figure 3b is the same image template relit from a different direction. Using an a least squares fit for Λ, a new image that is more similar in appearance to the target image is generated (see Figure 3c).

(a) (b) (c)

Figure 3. An example of illumination compensation using a low-dimensional multiplicative model. (a) Input template. (b) Input image under new illumination. (c) Synthesized image that is the product of illumination basis functions with the input.For this example a third order Legendre polynomial was used and the Legendre coefficients were computed using (6).

3. A Joint-space of Illumination and Shape for Tracking

For the sake of generality, we assume an N_S-dimensional shape space and a N_λ-dimensional illumination space that results in a joint space $\mathcal{A}_\mathcal{L} = \mathcal{L}(\mathbf{W}, T, \mathbf{\Delta})$. that maps a joint shape and appearance vector $X^A \in \mathbb{R}^{N_S+N_\lambda}$:

$$X_A = \begin{bmatrix} X \\ \Lambda \end{bmatrix} \quad (7)$$

to a deformed and relit template, $U \in \mathbb{R}^{N_T}$:

$$U = [\mathbf{\Delta}\Lambda]\,[I\,(\mathbf{W}X + T)]\,. \quad (8)$$

$\mathbf{W}$ denotes a $N_T \times N_S$ shape matrix. The constant offset T denotes the template against which shape variations are measured. No such offset is required for the illumination component. $I(\cdot)$ simply refers to the image intensities measured on the shape grid implied by the shape component of X^A.

The proposed joint shape-illumination space can be sampled sequentially to track objects through a range of shape and illumination changes. This is best accomplished in a robust way using a particle filter framework. Particle filters (PF) are very widely studied in computer vision and different variants of its implementation exist [13, 9]. Two important components of a PF include a state evolution model $p(X_t^A|X_{t-1}^A)$ and an observation model $p(Y_t|X_t^A)$. The PF tracker approximates the posterior density, $p(X_t^A|Y_{1:t})$ with a set of weighted particles $\{(X^A)_t^j, w_t^j\}$ with $\sum_{j=1}^M w_t^j = 1$. The likelihood $p(Y_t|X_t^A)$ of a particular hypothesis and in the case of the joint shape-appearance model is computed using the transformed image and the template. A likelihood measure on the joint shape-illumination hypothesis X_i^A is

computed as the sum of absolute difference (SAD) between U and T.

The other component of PF tracking is the specification of $p(X_t^A|X_{t-1}^A)$. Typically a Gauss-Markov model is assumed, whereby $X_{t+1}^A \sim \mathcal{N}(X_t^A, V)$. In the absence of any knowledge about the expected range of motion and illumination change, a brute force approach is required and the variance on the normal distribution of each component in X^A is set to a high value. This necessitates an unreasonable increase in the number of particles in order to maintain reliable tracking and such an approach is now more likely to suffer from local minima. With the additional dimensions that the new model implies, the problem can be even more formidable than traditional shape tracking where recent work has studied how more informed sampling distributions for shape tracking can be derived [8]. In the following section we outline how meaningful sampling densities for illumination can be learned from a few examples and show that these densities can in fact are degenerate. As a result, the new model can be represented by several centroids in the Legendre basis.

3.1. Learning Sampling Distributions for Illumination and Shape

We assume that we have a static camera acquiring images of a scene and that the illumination conditions, although variable within the scene, do not change significantly over time. Ground truth video sequences consisting of a starting template T and its location and shape in subsequent frames, $\{U_1, \cdots, U_N\}$ are used to compute shape-vectors $\{X_1, \cdots, X_N\}$ corresponding to this motion. Furthermore, a set of Legendre coefficents $\{\Lambda_1, \cdots, \Lambda_N\}$ that best map $\{U_1, \cdots, U_N\}$ to T are computed via standard least squares fitting (6).

The shape sampling distribution $h(X)$ must model the incremental motion between frames. For smooth motions, shape distributions can be computed from shape difference vectors $\{X_2 - X_1, \cdots, X_N - X_{N-1}\}$. Standard kernel-density methods can then be used for estimating a sampling distribution from using these differences. Alternatively a uniform density $U(a, b)$ corresponding to the maximal ranges of state components can be used as a simple approximation of $h(X)$.

In the case of our new illumination model, sampling distributions in the Legendre space must be estimated. It is natural to consider whether a differential model similar to the one used for X is suitable in this regard. Figure 6 illustrate the problem with such an approach for the illumination space. Although components of shape space are more or less smoothly monotonic (Figure 4c), this is not the case for the illumination coefficents. For example, the first coefficents, λ_0, changes dramatically as the subject moves through differing illumination. The result is a trajectory that cannot be modeled by considering discrete differences (Figure 4d).

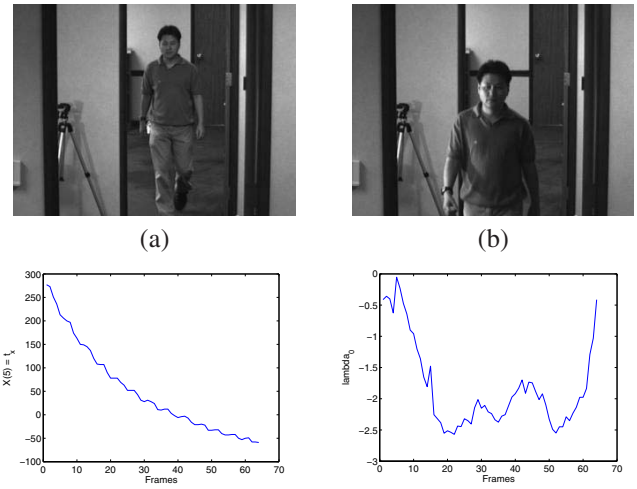


Figure 4. Difficulty of using a differential model for building a sampling distribution for illumination (a) and (b) show images of a person walking in a hallway towards the camera (c) shows the y-translation component of X and (d) shows the λ_0 coefficient of Λ as a function of time. As can be seen even for smooth motions, the illumination component displays discontinuities.

Figure 5. Our approach is motivated by the fact that certain dominant illumination conditions can be quantized into a few centroids in the illumination space. For example, in this scene some of the salient illumination conditions are: (a) subject is diffusely lit from above (b) subject passes thorough shadow, (c) subject strongly lit from the side, and (d) subject in darker region of room near camera.

One approach to this problem is to identify subregions of monotonicity and then build a mixture of distributions using discrete differences that are particular to each. However

one direct consequence of using these distributions is that the number of particles needed to span the corresponding regions in illumination space will be extremely large adding an additional computational burden on top of the traditional shape space sampling. Clearly a more efficient way of sampling the illumination space must be found if the resulting algorithm is to be useful.

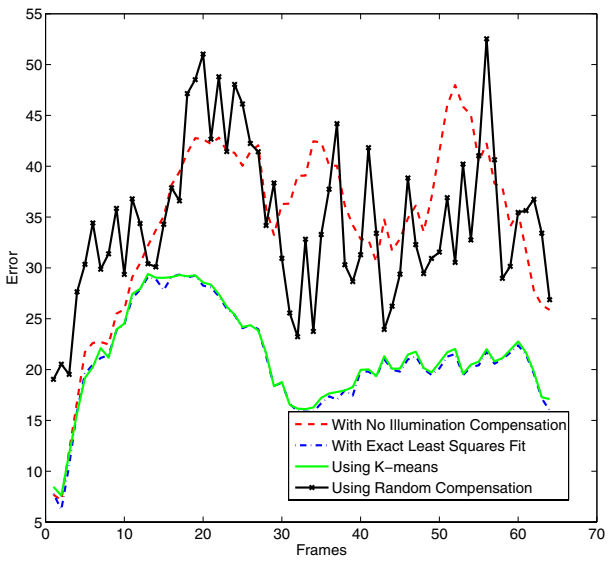


Figure 6. A plot of the SAD error as a function of time. The red dashed line represents the situation with no illumination compensation. The blue dash-dotted line represents the compensation with the least squares fit for Λ The green solid line represents compensation with the vector quantized values of the least square fits. The black crossed line represents compensation with a random Λ.

Although the underlying distribution of Λ is of course continuous, we can discard much of this information in favor of tracking robustness by seeking the most important illumination modes that are present in the distribution. This step is motivated by the observation that a scene is typically composed of a discrete set of illumination conditions. For example, the underlying illumination distribution for the scene shown in Figure 4 arises from certain salient illumination conditions in the scene as shown in Figure 5.

In order to achieve an efficient sampling of the illumination space we perform a $k-$means clustering $\{\Lambda_1, \cdots, \Lambda_N\}$ and use the k centroids $c_1, \cdots, c_k$ as a representation of the illumination space.

To demonstrate that clustering in this way does not degrade our ability to track, we studied many face tracking examples under different illumination conditions. The results support the claim that only a few modes are needed instead of the entire distribution. For example, Figure 6 show the SAD score achieved for a typical face tracking process us-

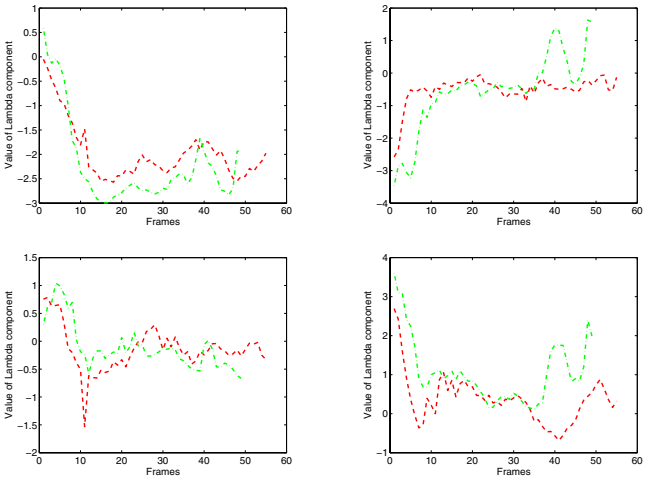


Figure 7. The range of Λ as a function of time shown for two individuals for a 2nd order Legendre polynomial fit. The similarity in the range of values taken by different components of Λ can be seen from the plot. As a consequence the centroids describing the illumination conditions in a scene for different people are close.

ing several different approaches. Both random selection of Legendre coefficents and no compensation lead to high error. More importantly, the plot shows nearly no difference between exact least-squares fits of a second order Legendre that utilizes only six centroids discovered via $k-$means clustering process.

This result is typical of most situations and a rate-distortion study found that $k = 6$ is adequate to represent the variability of Λ for our indoor surveillance senario. Figure 7 shows the result of least-squares fits to the first four Legendre coefficents for two different subjects. Note that the range of variability is nearly the same for both subjects justifying our use of the same centroids to represent several subjects from the same scene.

These results allows us to coarsely sample the illumination space with minimal impact on the tracking results while retaining the ability to generalize to previously unseen objects within the same class. This requires only minor modification to the standard particle filter that incorporates the k illumination clusters. Specifically for every particle j, drawn from $h(X)$, we sample i from $\{1, \cdots, k\}$ with probability $\frac{1}{k}$ and compute

$$U = [\mathbf{\Delta}\Lambda_{ci}] \left[I \left(\mathbf{W} X_t^j + T \right) \right] \tag{9}$$

before measuring the SAD distance. The new algorithm, then, combines traditional shape tracking with our multiplicative model of illumination compensation. Table 1 summarizes the joint shape-illumination tracking algorithm.

Given: an estimate of shape sampling distributions, $h(X)$ and k cluster centers, $c_1, ..., c_k$ in the illumination basis.

1. Initialize sample set $\mathcal{X} = \left\{X_0^j, 1/M\right\}$
  2. For $t = 1, \cdots, T$
    3. For $j = 1, \cdots, M$
    4. Generate $X_t'^j$ from X_{t-1}^j using $h(X)$
      5. Compute transformed image regions in accordance with shape vectors $X_t'^j$
      6. Pick an i from $\{1, \cdots, k\}$ with probability $\frac{1}{k}$
      7. Compute U using (9)
      8. Compute likelihood $p(Y_t|X_t'^j)$ by measuring the SAD distance between U and T
    9. End
    10. Importance resample $\{X_t'^j\}$ based on $\{p(Y_t|X_t'^j)\}$ to get $\{X_t^j\}$
11. End

Table 1. The Particle Filter using the new shape-illumination space.

4. Experimental Results

We now demonstrate the utility of the joint shape-illumination model in two different scenarios. The results discussed here are indicative of results the system achieved for many such sequences. For example, in the car sequence twenty cars were successfully tracked over a period of two hours[1] In each case we follow the procedure described in Section 3.1 to establish sampling distributions in the joint-space over some set of training samples. Tracking was then performed using 200 particles on new objects using the algorithm in Table 1.

The car dataset was generated from a camera observing a road from above as cars approach an intersection and move in and out of shadow. Two sequences were used to acquire the sampling distributions. Training involved marking locations of the moving car in successive frames. Using these locations the corresponding shape-vector was computed. We used a 3-D shape space that spans scaling and translations in X and Y. Using the maximal values of the shape difference vectors, a uniform distribution over the corresponding range was computed for each shape component. Using the least squares method (see Section 6) we fit different orders of Legendre polynomials and computed the resulting SAD error. We found that a first order Legendre polynomial was adequate to capture the illumination change in this case where the object is more or less planar. The $k-$means clustering process yields two centers $\{c_1, c_2\}$ that were then used to represent the discretized illumination space.

Figure 8 shows tracking results for a car using the joint shape-illumination tracker. The white square corresponds to the MAP estimate for that frame. The new tracking algorithm is compared to a traditional particle filter that does not encompass illumination change (Figure 8 bottom row). The same shape sampling distributions were used by both algorithms.

The particle filter tracks the template well as long as the illumination conditions that existed when the template was captured remain unchanged. However, at the shadow boundary the traditional tracker fails. On the other hand, the new illumination model captures this appearance change and the joint shape-illumination likelihoods remain high for the correct estimate via the additional degree-of-freedom afforded by the illumination model.

A second dataset contained several different subjects moving through different illuminations in an indoor environment. The illumination conditions in this case were significantly more complex than the vehicle tracking dataset. Sunlight through window and different light sources (i.e. fluorescent overhead lamps and incandescent desk lights) persist throughout the space making the dataset very challenging. In fact, to test the algorithm a strong diffuser lamp was placed in a room to generate strong side lighting (see Figure 9). Ground truth was again generated from two different sequences. A second order Legendre polynomial was chosen for the illumination component. Using rate-distortion studies as discussed in Section 3.1, we found that around six clusters were required to capture the variability in the scene. Here we discuss tracking results when six clusters were used. Using more centroids does not lead to degradation of the results, however it requires an additional number of particles.

Figure 9 shows two different subjects moving through various illumination conditions as they approach a surveillance camera. These sequences are typical for this setup and only three frames are shown in the interest of space.

Figure 10 shows the initial template for each subject and the illumination image generated by the illumination centroid associated with the MAP estimate. This illumination image was multiplied to the grid indicated by the shape vector in the frames shown in Figure 9. As can be seen these illumination images are able to compensate for the illumination changes in the sequence.

5. Conclusions and Future Work

In this paper we presented an approach to track across shape and illumination change. We introduced a low-dimensional multiplicative model of illumination change that is expressed as a linear combination of a Legendre basis. We demonstrated how this new model is capable of

[1]The cars were arbitrarily picked in the sequence and initial locations of the cars were hand extracted and passed on to the tracker

Figure 8. Example of tracking a car through drastic illumination changes. The bottom row shows the result using a conventional particle filter while the top row shows the result using our algorithm.

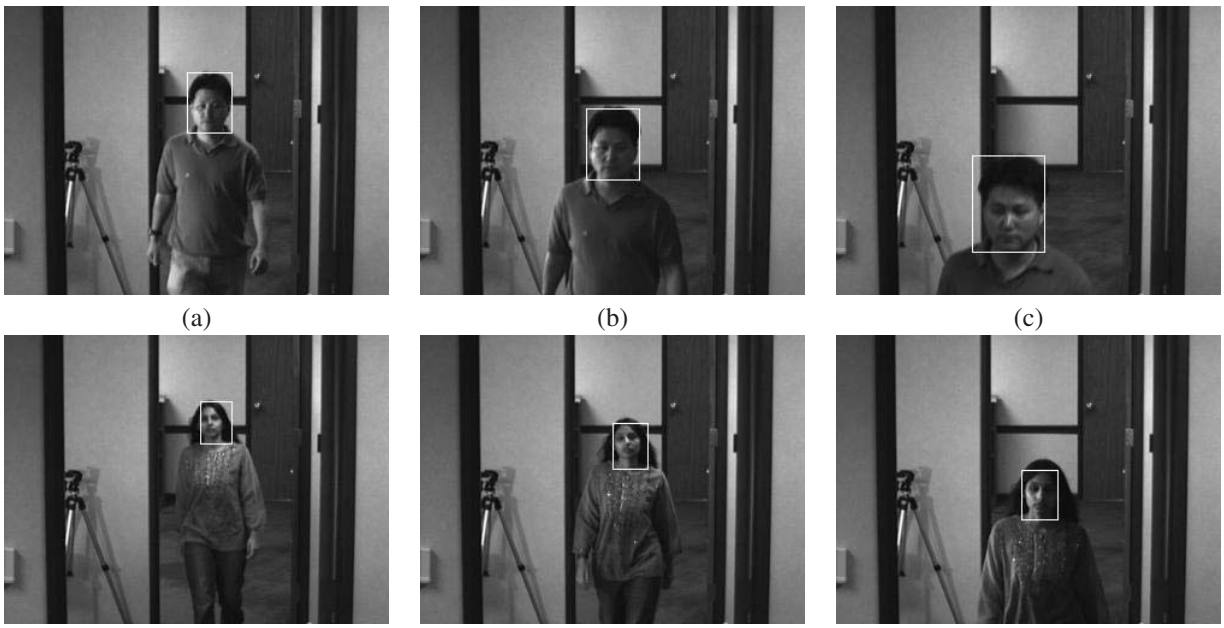


Figure 9. Example of tracking faces in an indoor setting. The illumination conditions existing in this scenario are significantly more complex than those in the vehicle tracking situaion.

capturing appearance change in the tracked template. We showed how the Legendre coefficients can be combined with the shape vector to define a new shape-illumination space. We discovered that in this new illumination space, a small number of centroids suffice to capture illumination changes in particular scenario. We showed how to estimate these centroids and incorporate them in the particle filtering framework at run time without adding excessive computational burden. We demonstrated the utility of our approach for both vehicle and face tracking scenario. One of the assumptions in our work is that the initial template in the training and testing sequences are acquired under sim-

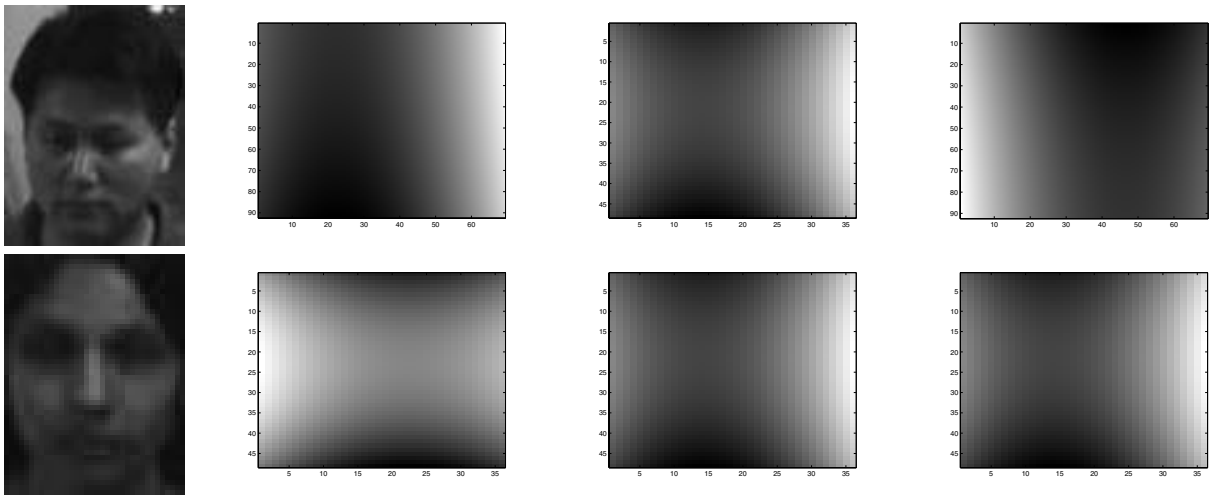


Figure 10. Initial template and illumination images constructed from the Legendre basis that were used to model appearance change in the sequence shown in Figure 9.

ilar illumination conditions. We expect to incorporate the bilinear style-content factorization of Freeman and Tenenbaum [5] to overcome this drawback. Finally, more sophisticated studies involving the stability of the learned distributions over time and slow illumination changes are underway. Initial results indicate that the distributions can be quite stable but may need to be re-learned over some period over time. For example, distributions learned at dawn no longer apply at dusk.

References

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

[2] P. Belhumeur and D.J.Kriegman. What is the set of images of an object under all possible illumination conditions. *IJCV*, 28(3):1–16, 1998. 1

[3] B.Han and L. Davis. On-line density-based appearance modeling for object tracking. *Proceedings of ICCV*, 2005. 2

[4] D. Freedman and M. Turek. Illumination-invariant tracking via graph cuts. *IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR)*, 2:10–17, 2005. 2

[5] W. Freeman and J. Tenenbaum. Learning bilinear models for two factor problems in vision. *Proceedings of IEEE CVPR*, 1997. 7

[6] G. Hager and P. Belhumeur. Efficient region tracking with parametric models of geometry and illumination. *IEEE Trans. PAMI*, 20(10):1025–1039, 1998. 1

[7] M. Isard and A. Blake. Condensation - conditional density propogation for visual tracking. *IJCV*, 21(1):695–709, 1998. 1

[8] A. Kale and C. Jaynes. Shape space sampling distributions and their impact on visual tracking. *IEEE International Conference on Image Processing*, 2005. 4

[9] L. Lu, X.Dai, and G. Hager. A particle filter without dynamics for robust 3d face tracking. *Proc. of FPIV*, 2004. 3

[10] I. Matthews, T. Ishikawa, and S. Baker. The template update problem. In *Proceedings of the British Machine Vision Conference*, September 2003. 2

[11] R. Ramamoorthi. Analytic pca construction for theoretical analysis of lighting variability in images of lambertian object. *IEEE Trans. PAMI*, 24(10):1–12, 2002. 1

[12] Y. Weiss. Deriving intrinsic images from image sequences. *Proc of ICCV*, 2001. 2

[13] S. Zhou, R. Chellappa, and B. Moghaddam. Visual tracking and recognition using appearance-adaptive models in particle filters. *IEEE Trans. on Image Processing*, November 2004. 2, 3