MODELING AND TRACKING OF FACES IN REAL-LIFE ILLUMINATION CONDITIONS

Rahul Thota, Archana Kalyansundar, Amit Kale

Siemens Corporate Technology India
Siemens Information Systems Limited,
No.84 Keonics Electronics City, Bangalore

ABSTRACT

In this paper, we address the problem of face tracking across illumination changes and occlusions. The method is based on leveraging the strengths of both Adaboost to deal with clutter and the image based parametric illumination model proposed by Kale and Jaynes. We show that a simple non-linear transformation of the Adaboost score multiplied with the illumination compensated likelihood leads to a fast robust tracking paradigm. We demonstrate the ability of our method to detect occlusions at the same time ensuring that mis-assignments between the occluder and the occluded does not occur. We present experimental results of our method on low resolution surveillance indoor and outdoor videos using an off the shelf DSP. We also demonstrate the power of the parametric illumination model for pose constrained face recognition when matching across known illumination conditions.

Index Terms: Surveillance, Tracking, Face recognition.

1. INTRODUCTION

This paper addresses the problems that arise for face tracking in the presence of illumination changes and occlusion. Solving this problem can provide situational awareness in establishments like smart buildings e.g. saying who went where. In contrast to approaches that use whole body appearance models for individuals, face is person specific and can be acquired non-intrusively. Our approach is intended for face tracking using a single camera set up at vantage points e.g. along a hallway so that an approximate frontal view of a face is available.

An important contribution to the problem of face processing in video has been the Adaboost based face detection algorithm of Viola and Jones [1]. Several researchers have tried to extend their framework to the problem of visual tracking. An example of this is ensemble tracking [2]. In their paper, tracking is treated as a binary classification problem, where an ensemble of weak classifiers is trained on-line to distinguish between pixels belonging to object and the background. One of the drawbacks of this approach is that recomputation of the weak classifiers at every frame can lead to drift in the presence of sudden appearance changes due to illumination/pose changes. More recently there has been a shift towards directly using detection scores for tracking [3]. Specifically Thierry et al [3] transforms Adaboost and SVM detection scores to output probabilities and use them in a particle filter framework. One of the difficulties of using this approach for face based tracking is that it can fail in the presence of occlusions and illumination changes. Particle filters [4](PF) have been a popular tracking paradigm for the last decade. PFs rely on a motion model on shape space in the prediction step which prescribes the search windows on which likelihood is computed. One of the problems in realistic surveillance scenarios is that objects do not always obey the motion model. This problem is especially serious in the case of low frame rate cameras. Alternatively adaptive appearance models [5, 6] can be used. Such models do not however preclude the possibility of drift. 3-D model based methods are also quite popular [7, 8] for modeling illumination effects. In surveillance type scenarios where resolution of the face may go as low as 15×15(see Figure 6), registration and 3D motion estimation for using 3D models is daunting. A better alternative is to use scene-specific illumination priors. For example, [9]proposed a joint shape illumination model in a PF based visual tracking framework which relies on coarse quantization of the illumination space modeled by linear combination of Legendre polynomials. As discussed above, however, the prediction step can produce candidate regions on non-face clutter, which in cases of low frame rate surveillance imagery or dimly lit backgrounds can lead to tracking failure when using a small particle budget. In this paper, we address the above problems using a novel likelihood function which leverages the strengths of both Adaboost and the illumination model of Kale and Jaynes [9]. Product fusion is an appropriate decision fusion rule for these two independent evidences on a given image region. We show that a simple non-linear transformation of the Adaboost score multiplied with the illumination compensated likelihood leads to robust tracking. We show how the form of this new likelihood leads to a fast algorithm, combining the ability of Adaboost to deal with background clutter with a model for illumination and identity. Additionally we also show how the illumination model improves pose constrained face recognition and compares our approach with an alternative illumination model based on style content factorization [10]. We demonstrate how this aspect of our likelihood function provides robustness to recovery from occlusion especially in the case of occlusion by other faces. We have implemented a real-time system on the Texas Instuments DSP TMS320DM642 that works on live images. We present results of our method in challenging indoor and outdoor conditions including illumination changes and occlusions.

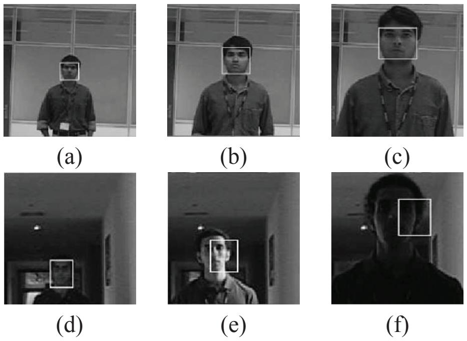
2. TECHNICAL DETAILS

2.1. Detection based tracking

Viola and Jones [1] pioneered the use of Adaboost for face detection. The method involves computing a sequence of “weak” classifiers each of which is a thresholded spatial filter output, on the image subwindow considered. The final strong classifier output score $H(U)$ is a weighted combination of weak classifiers $H(U) = \sum_{t=1}^{T} \alpha_t h_t(U)$. The weight α_t of a selected weak classifier h_t is inversely proportional to its classification error.

Given the face location in the first frame, a simple way to track people is to exhaustively sample subwindows $U_{k,t}$ in a reasonably sized fixed region around the previous location of the face (at mul-

978-1-4244-2354-5/09/$25.00 ©2009 IEEE



(a) (b) (c)

(d) (e) (f)

Fig. 1. Detection based tracking results. (a)(b)(c) show the results in a simple scenario without many illumination changes and (d)(e)(f) show the tracking failure in the presence of strong illumination changes.

tiple scales) and compute Adaboost score over the collection $\mathcal{U}_t$ of these subwindows. A reasonable estimate of the face postion will correspond to the location where the Adaboost score peaks. Taking this location as the current track, the next set of samples $\mathcal{U}_{t+1}$ can be generated and the procedure repeated. The results of this tracker are shown in Figures 1(a)(b)(c). We trained the Adaboost classifier using close to 5000 faces (used originally by Viola and Jones) and 8000 non-faces: the database is generic enough and encompasses various illumination conditions. Despite this, the detection based tracking fails to track across illumination changes (see Figures 1(d),(e),(f)). They clearly show how the confidence map peak does not correspond to the ground truth in adverse lighting conditions. Furthermore, these detection based tracking approaches do not provide any identity characterization of the objects being tracked leading to difficulties when occlusions by objects of the same class occur.

2.2. Illumination modeling

Illumination changes affect the appearance of the object being tracked. Note that the standard adaboost face detector handles uniform illumination changes using variance normalization. However it is apparent that such a normalization is incapable of handling uneven illumination conditions such as those shown in Figure 1(d)(e)(f). Ideally illumination compensation will account for appearance changes due to varied lighting more accurately than illumination invariant feature based algorithms. Recently Kale and Jaynes [9] introduced a low dimensional model of appearance change to deal with such situations. We briefly review their model here. The image template T_t in the tracking sequence can be expressed as:

$$T_t(x,y) = L_t(x,y)R(x,y) = \tilde{L_t}(x,y)T_0(x,y) \qquad (1)$$

where $L_t(x,y)$ denotes the illumination image in frame t and $R(x,y)$ denotes a fixed reflectance image [11]. In absence of knowledge of R, the problem of estimating the illumination image reduces to estimating $\tilde{L_t}$ w.r.t to the illumination contained in the image template $T_0 = L_0.*R$. Kale and Jaynes [9] model $\tilde{L_t}$ as a linear combination of a set of N_Λ Legendre basis functions. Denoting $p_k(x)$ as the k th Legendre basis function for $N_\Lambda = 2k+1$, $\Lambda = [\lambda_0, \cdots, \lambda_{N_\Lambda}]^T$, the scaled intensity value at a pixel of the template T_t is computed as:$T_t(x,y) = T_0(x,y) + T_0(x,y)\mathbf{P}(x,y)\Lambda$ where

$$\mathbf{P}(x,y) = \frac{1}{2k+1}[1\ p_1(x)\cdots p_k(x)\ p_1(y)\cdots p_k(y)] \qquad (2)$$

Rewriting T_0 and T_t as vectors we get $[T_t]_{vec} = [T_0]_{vec} + [T_0]_{vec} \otimes \mathbf{P}\Lambda$ so that when $\Lambda \equiv 0, T_t = T_0$. Operator $\otimes$ refers to multiplying each row of $\mathbf{P}$ by $T_0(x,y)$. Given T_t and T_0, the Legendre coefficients that relight T_t to resemble T_0 can be computed by solving the least squares problem:

$$A_{T_t}\Lambda_t \approx [T_t - T_0]_{vec} \qquad (3)$$

where

$$A_{T_t} \triangleq [T_t]_{vec} \otimes \mathbf{P} \qquad (4)$$

so that $A_{T_t} \in \mathbb{R}^{N_\Lambda + M}$ Given the ground truth of template locations in successive frames (3) can be used to find the illumination coefficients $\{\Lambda_1, \cdots, \Lambda_N\}$. Although the underlying distribution of these Λ_ts is continuous [9] shows that much of this information can be condensed down to a discrete number of important illumination modes $\{c_1, \cdots, c_k\}$ via $k-$means clustering, without sacrificing tracking accuracy. For example if a corridor has predominantly three different lighting conditions, we find that the Legendre coefficient vectors also cluster into three groups.

2.2.1. Face Recognition Performance Analysis

The proposed illumination model increases discriminability between objects of the same class, apart from providing a simple illumination compensation. We illustrate this using a simple experiment on the CMU-PIE dataset. We divided the frontal images of the dataset into training and testing sets of 34 people each.

For each person, the face image under illumination condition 19 was chosen as her/his template T_j. For each face of a different illumination condition $U_{i,j}, i \neq 19, j = 1:34$ in the training we compute the $\Lambda_{i,j}$ which relights it to resemble $T_j, j = 1:34$ using (3). In order to test the generalizability of the centroids, we compute the average $\bar{\Lambda}_i = \frac{1}{34}\sum_{j=1}^{34}\Lambda_{i,j}$. We apply $\bar{\Lambda}_i$ to each person in the test set to derive illumination compensated versions of the raw uncompensated inputs $U_{i,j}, j = 35:68, i \neq 19$ computed as $\hat{T}_{i,j} = U_{i,j} + U_{i,j} \otimes \mathbf{P}\bar{\Lambda}_i$. We compare the recognition rates obtained when using the raw uncompensated images $U_{i,j}, j = 35:68, i \neq 19$ and these illumination compensated versions $\hat{T}_{i,j}$ with the gallery comprising of $T_j, j = 35:68$ using similarity matrices and cumulative match characteristics (CMC)(see Figure 2(b)). As can be seen, the top match corresponds to the actual person 78% of times with illumination compensation compared to 34% when using no compensation. Thus, illumination compensation by appropriate centroids learned from the training data significantly improves the discriminability without using a simple model. This fact has implications for tracking where the goal is to ensure that the identities are maintained in the event of an occlusion. Using the above illumination model we can derive a simple metric to encompass identity and illumination change.

$$d_{illum}(U_t, T_0) = \min_{\{c_1,\cdots,c_k\}} d(T_0, U_t + U_t \otimes \mathbf{P}c_k) \qquad (5)$$

where d denotes SAD (sum of absolute differences) distance.

Appearance changes due to illumination can also be modeled by separating style (illumination) from content (people). Such a separation has been successfully deployed in [10], which uses a bilinear model to fit a training set of observed images. Bilinear models are two-factor models which are separable: the outputs are linear in either factor when the other is held constant. As in (6) the observation image, in a vector form $U^{persons,illum}$ can be represented as the weighted combination of each of the basis vectors $W_{i,j}$

$$U^{illum,person} = \sum_{i,j} W_{i,j} a_i^{illum} b_j^{person} \qquad (6)$$

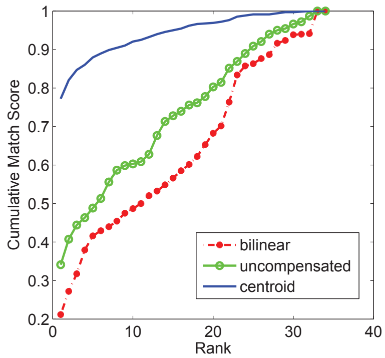


Fig. 2. (a)Cumulative Match characteristics (over PIE dataset): Clearly with our illumination compensation, the recognition rate improves compared to SCF

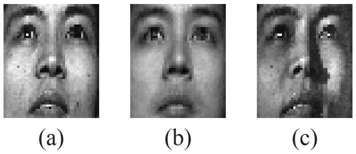

(a) (b) (c)

Fig. 3. (a) Reference image (b) SCF-based rendition of the frontal lighting (c) Illumination compensation using the centroids model.

where i and j vary over the number of illuminations ($I = 20$) and persons ($J = 34$) used in the training respectively. Here the illumination information is encoded in the I dimensional parameter vector a_i^{illum} and the person information is embedded in the J dimensional parameter vector b_j^{person}. So we can render an image of a particular person in a particular illumination by summing the basis images $W_{i,j}$ weighted by the corresponding parameter vectors. Although style content factorization (SCF) leads to better re-lighting of faces visually (Figure 3), it does not automatically translate to better recognition performance (Figure 2). As it can be seen from that plot, recognition probability using this method drops to 21%. We believe that this drop in performance is due to loss of high frequency/person dependent features.

2.3. Algorithm Description

One way to combine the benefits of both Adaboost and the illumination model is to consider a decision fusion of the two. Treating the two cues as being independent, a plausible approach to leverage their strengths is to consider their product fusion [12] and using the peak of the result. Such a product fusion does not yield satisfactory results however. The reason is that many locations such as dimly lit background clutter with even low Adaboost scores can yield low values of d_{illum}. A possible solution to is to apply a nonlinear transformation to the orthogonal evidences before performing a product fusion.

An examination of the failure modes of simple Adaboost based tracking reveals that (a) For certain illumination conditions, the Adaboost scores on the face region can in fact be lower than certain regions which partially overlap the face while for both these cases the scores are high (> 0.85). (b) The Adaboost scores on dimly lit clutter are lower (around 0.7). On the other hand, the values for d_{illum} for regions which partially overlap the face are usually much higher than those for the true face regions. These facts suggest a simple way of getting the best of both worlds: If we know the lower bound τ_{min} of all Adaboost scores on face regions, all regions with scores above τ_{min} can be considered to be candidates for face locations. However, only the true face will yield a low value for d_{illum} . Thus, the face tracking problem can be reduced to the problem of maximizing the likelihood function:

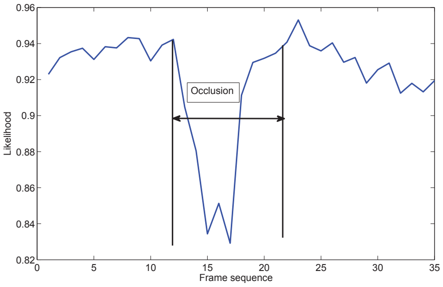


Fig. 4. Likelihood plot during when tracking faces across occlusions corresponding to the top 3 rows of Figure 6(d). As we can see (7) drops by a large value when the face gets occluded.

$$L(U_{k,t}, T_0) = exp^{-d^2_{illum}(U_{k,t}, T_0)} I(H(U_{k,t}) > \tau_{min}) \qquad (7)$$

where $I(.)$ is an indicator function. It is clear that d_{illum}, needs to be evaluated only on a subset of windows for which $I(H(U_{k,t}) > \tau_{min}) = 1$. This also achieves a considerable speed up in the algorithm at run time besides improving its robustness.

Occlusion Handling The likelihood measure (7) also provides a powerful means for handling occlusion. From our experiments, we observed a sharp drop in the likelihood function just at the onset of occlusion (see figure 4). Thus if $\hat{L}_t - \hat{L}_{t-2} > \theta$ then tracking is stopped owing to occlusion. Furthermore, since the drop in the likelihood observed due to occlusion is generally much higher compared to its variance caused by illumination changes, the threshold θ is easy to pick and the performance is not overly sensitive to its choice. We chose $\theta = 0.04$. Under the assumption that the occluder occludes only for a short duration it is reasonable to expect that after the end of occlusion the likelihood value returns to the value before occlusion $\hat{L}_{last} = L(\hat{U}_{t_{last}}, T_0)$. Thus to determine the occlusion status of the object it is enough to examine a fixed size perimeter around the last location $\hat{U}_{t_{last}}$, the condition $L(\mathcal{U}^{t>t_{last}}, T_0) = \hat{L}_{last}$. But, in the event of longer occlusions such an assumption may not hold for reasons like centroid switching due to change in the lighting etc. In order to account for such small variations in the likelihood we setup a tolerance range ($\hat{L}_{t_{last}} - \sigma, \hat{L}_{t_{last}} + \sigma$), where σ is simply the standard deviation of the illumination compensated likelihood on the ground-truth.

3. RESULTS

We captured a surveillance database with people walking indoors and outdoors. For the proposed method we computed both τ_{min} and centroids (c_m's) using k-means over 3 persons' ground truth data. The tracker was initialized using Adaboost on a small ROI at the entry point. Testing was done on people different from the ones

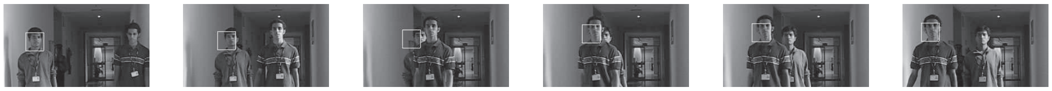


Fig. 5. Tracking result when using meanshift [13]. As we can see, since face-histograms are not distinctive enough to distinguish occluder and the occluded, the track gets swapped to the wrong face after occlusion.

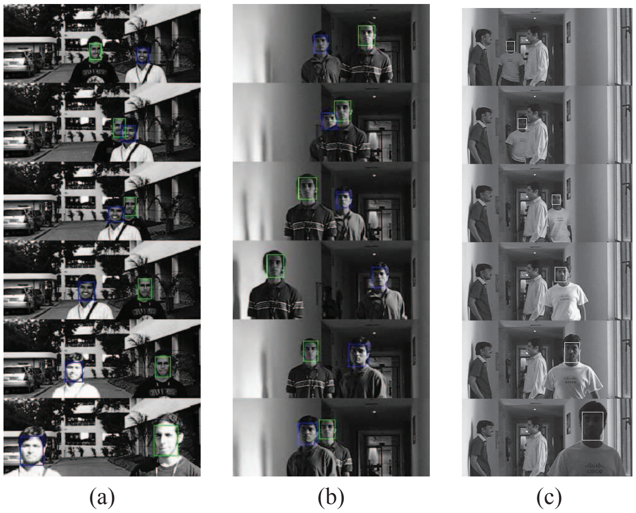


Fig. 6. Face tracking results (a)Outdoor tracking and (b)Indoor tracking with occlusion; (c) Tracking across natural occlusion in our corridor.

used for ground truth. Typical values of τ_{min} and M evaluated in the indoor environment are 0.92 and 3 respectively.

In order to test the robustness of our algorithm, we set up a camera facing a corridor in our establishment to monitor the flow of people. Most of the time, complicated occlusions do not occur in the course of natural flow of people in the corridor. Figure 6(c) shows one of the harder cases where a person is occluded by two other people conversing in the corridor. The occlusion in this case, is more pronounced and yet we do not lose track of the person when using our algorithm. In our experiments, we mainly dealt with a corridor/hallway scenario since it can be thought of as a basic building block in a larger surveillance system which can track people over a wider area within an establishment.

We also compared the performance of mean shift (MS) tracking [13] on some of the test sequences. The results are shown in Figure 5. We see that the tracker results in mis-assignment of identity, which is a consequence of face histograms not being distinctive enough.

The algorithm was implemented on TI's TMS320DM642 DSP platform clocked at 720MHz. The hardware platform was interfaced with a camera and a VGA monitor. We optimized the algorithm with negligible loss in precision to obtain a performance of 48 frames per second for the case of two-centroid environment (sunlight and shadow scenario - installed outdoor) and 10 fps in a four-centroid environment (for more complicated illuminations across the corridor - installed indoor). We found that the Adaboost module takes $0.11ms$ per subwindow and the illumination compensation takes $0.31ms$ per centroid.

4. CONCLUSION AND FUTURE WORK

In this paper, we proposed a method to track faces across illumination changes and occlusion. The method is based on leveraging the strengths of Adaboost and the illumination model proposed by Kale and Jaynes. The method works reliably on challenging low resolution indoor and outdoor face videos which can occur in a surveillance type scenario. From our observations, we would like to stress that illumination compensation leads to better face recognition performance. Even though the approach is intended for monocular face tracking, it can handle slight pose changes and works well when the cameras are placed at vantage points where large and sustained pose changes do not occur e.g. in hallways. We are also currently extending the system proposed in this paper across a wider area within our building so that a handoff is made to an appropriate camera based on the track from the previous one.

5. REFERENCES

[1] P.Viola and M.Jones, "Robust real-time face detection," *IJCV*, 2004.

[2] S. Avidan, "Ensemble tracking," *Proc. of CVPR*, 2005.

[3] T.Chateau, V.G.Belille, F.Chausse, and J.T. Lapreste, "Real-time tracking with classifiers," *Proc. of ECCV*, 2006.

[4] M.Isard and A.Blake, "Condensation - conditional density propogation for visual tracking," *IJCV*, vol. 21, no. 1, pp. 695–709, 1998.

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

[6] A.D.Jepson, D.J.Fleet, and T.F.ElMaraghi, "Robust online appearance models for visual tracking," *IEEE Trans. on PAMI*, October 2003.

[7] P.N.Belhumeur and D.J.Kriegman, "What is the set of images of an object under all possible illumination conditions," *IJCV*, vol. 28, no. 3, pp. 1–16, 1998.

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

[9] A.Kale and C.Jaynes, "A joint illumination and shape model for visual tracking," *Proceedings of IEEE CVPR*, pp. 602–609, 2006.

[10] W.T.Freeman and J.B.Tenenbaum, "Learning bilinear models for two-factor problems in vision," *Proc. of CVPR*, 1997.

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

[12] B.Achermann and H.Bunke, "Combination of classifiers on the decision level for face recognition," Tech. Rep., Institut fur Informatik und angewandte, Mathematik,, Universitat Bern, 1996.

[13] D.Comaniciu, V.Ramesh, and P.Meer, "Real time tracking of non-rigid objects using mean shift," *Proc. of CVPR*, 2000.