TRACKING HUMAN FACES IN INFRARED VIDEO

Christopher K. Eveland *Diego A. Socolinsky* *Lawrence B. Wolff*

Equinox Corporation
9 West 57th Street
New York, NY 10019

Equinox Corporation
207 East Redwood Street
Baltimore, MD 21202

Equinox Corporation
9 West 57th Street
New York, NY 10019

{eveland,diego,wolff}@equinoxsensors.com

ABSTRACT

Detecting and tracking face regions in image sequences has applications to important problems such as face recognition, human-computer interaction, and video surveillance. Visible sensors have inherent limitations in solving this task, such as the need for sufficient and specific lighting conditions, as well as sensitivity to variations in skin color. Thermal infrared (IR) imaging sensors image emitted light, not reflected light, and therefore do not have these limitations, providing a 24-hour, 365-day capability while also being more robust to variations in the appearance of individuals.

In this paper, we present a system for tracking human heads that has three components. First, a method for modeling thermal emission from human skin that can be used for the purpose of segmenting and detecting faces and other exposed skin regions in IR imagery. Second, the segmentation model is applied to the CONDENSATION algorithm for tracking the head regions over time. This includes a new observation density that is motivated by the segmentation results. Finally, we examine how to use the tracking results to refine the segmentation estimate.

1. INTRODUCTION

While much work has been done on detecting [1, 2, 3] and tracking [4, 5] humans in image sequences, most of the effort has been with visible sensors. Here we examine the problem from the point of view of a thermal IR sensor. In particular we examine both mid-wave (MWIR) and long-wave (LWIR) sensors.

There are several advantages of using IR over traditional visible wavelength sensors. These advantages arise from the fact that most light in the mid-to-long wave IR is emitted rather than reflected. [6] This leads first to a 24-hour, 365-day capability since the scene will not be dependent on, and will be nearly invariant to, external lighting sources such as the sun or man-made lights.

In addition to the lighting invariance, another aspect that is of particular importance to detecting and tracking faces is the relative uniformity of emissivity values of skin among different members of the population. This contrasts to work done in the visible spectrum, where albedo can vary significantly from person to person. While work has been done on finding invariants related to skin color in the visible spectrum [7], this problem can be much more easily solved with a calibrated IR sensor. More information on calibrated IR can be found in Section 2.

Given calibrated IR as its input, we present in Section 3 a model of skin imagery that can be used to segment images into three classes: skin, covered skin (as by clothes or hair), and background, which is anything else. We work with indoor environments, but allow for warm items such as computers in the background, without significant distraction from clutter.

Such an ability to differentiate between skin and background pixels can obviously be of great use when trying to track face regions. Many tracking algorithms using visible spectrum sensors use background subtraction [2] to serve this function. [5] Such background subtraction typically requires a fixed camera mount, or limited camera motion. While it is possible to use motion to perform such segmentation through 3D reconstruction [8], the relatively simple pixel process that can be used in conjunction with thermal IR sensors has significant advantages in computational cost, robustness, and the ability to detect stationary targets.

With the segmentation, we proceed in Section 4 to perform tracking. We use the segmentation as a feature set to track with, which has several advantages of the more traditional approach of using edge features and contours. First, since the quality of the skin segmentation is sufficient for head detection, the tracker can self-initialize reliably. The second advantage of using the higher-level features is that the tracker is less prone to distractors. This is because most

THIS RESEARCH WAS SUPPORTED BY THE DARPA HUMAN IDENTIFICATION AT A DISTANCE (HID) PROGRAM, CONTRACT # DARPA/AFOSR F49620-01-C-0008.

Report Documentation Page

Form Approved
OMB No. 0704-0188

1. REPORT DATE	2. REPORT TYPE	3. DATES COVERED
2006		**00-00-2006 to 00-00-2006**

4. TITLE AND SUBTITLE	5a. CONTRACT NUMBER
Tracking Human Faces in Infrared Video	5b. GRANT NUMBER
	5c. PROGRAM ELEMENT NUMBER
6. AUTHOR(S)	5d. PROJECT NUMBER
	5e. TASK NUMBER
	5f. WORK UNIT NUMBER
7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES) **Equinox Corporation,9 West 57th Street,New York,NY,10019**	8. PERFORMING ORGANIZATION REPORT NUMBER
9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES)	10. SPONSOR/MONITOR'S ACRONYM(S)
	11. SPONSOR/MONITOR'S REPORT NUMBER(S)

12. DISTRIBUTION/AVAILABILITY STATEMENT
Approved for public release; distribution unlimited

13. SUPPLEMENTARY NOTES
The original document contains color images.

14. ABSTRACT
see report

15. SUBJECT TERMS

16. SECURITY CLASSIFICATION OF:			17. LIMITATION OF ABSTRACT	18. NUMBER OF PAGES	19a. NAME OF RESPONSIBLE PERSON
a. REPORT	b. ABSTRACT	c. THIS PAGE			
unclassified	**unclassified**	**unclassified**		**10**	

Fig. 1. Uncalibrated (left) and calibrated LWIR images. There is significant pixel-wise non-uniformity in responsivity which is removed by the calibration process.

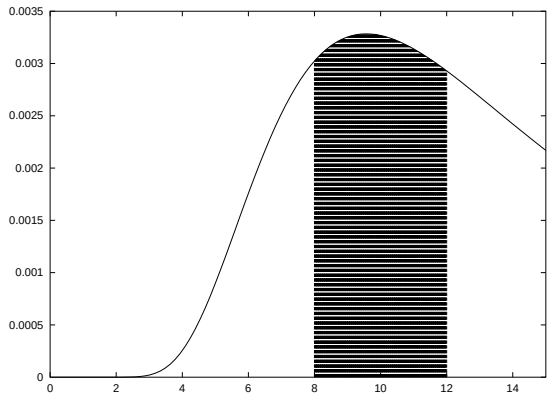

Fig. 2. The Planck curve for a black-body at 303K (roughly skin temperature), with the area to be integrated for an $8-12\mu$m sensor shaded.

of the background is classified as such. While some regions may be miss-classified, their extent tends to be minimal.

Just as the results of the segmentation can be useful for tracking, the results of tracking can aid in the segmentation. Since the segmentation algorithm is trained on a fixed population, and environmental conditions and variations in the population can cause variations in the observed intensities of skin, we refine the skin-model estimate using data collected from the tracker. This is described in detail in Section 5. Finally, we present results and conclusions in Sections 6 and 7.

2. FROM THERMAL VIDEO TO PHYSICAL MEASUREMENTS

In order to perform proper analysis, it is necessary that thermal IR imagery be radiometrically calibrated. Radiometric calibration achieves a direct relationship between the grayvalue response at a pixel and the absolute amount of thermal emission from the corresponding scene element. This relationship is called responsivity. Thermal emission is measured as flux in units of power such as W/cm^2. The grayvalue response of thermal IR pixels for LWIR and MWIR cameras is linear with respect to the amount of incident thermal radiation. The slope of this responsivity line is called the *gain* and the y-intercept is the *offset*. The gain and offset for each pixel on a thermal IR focal plane array is significantly variable across the array. That is, the linear relationship can be, and usually is, significantly different from pixel to pixel. This is illustrated in Figure 1 where both calibrated and uncalibrated images are shown of the same subject.

While radiometric calibration provides non-uniformity correction, it also provides the further advantage of data where environmental factors contribute to a much lesser degree. This is due to the relationship back to a physical parameter of the imaged object, its emissivity.

Since the responsivity of LWIR/MWIR sensors is very linear, the pixelwise linear relation between grayvalues and flux can be computed by a process of two-point calibration. Images of a black-body radiator covering the entire field of view are taken at two known temperatures, and thus the gains and offsets are computed using the radiant flux for a black-body at a given temperature.

Note that this is only possible if the emissivity curve of a black-body as a function of temperature is known. This is given by Planck's Law, which states that the flux emitted at the wavelength λ by a blackbody at a given temperature T in $W/(cm^2\mu m)$ is given by

$$W(\lambda, T) = \frac{2\pi hc^2}{\lambda^5 \left(e^{\frac{hc}{\lambda kT}} - 1\right)} \tag{1}$$

where h is Planck's constant, k is Boltzmann's constant, and c is the speed of light in a vacuum. To relate this to the flux observed by the sensor, the responsivity, $R(\lambda)$ of the sensor must be taken into account. This allows the flux observed by a specific sensor from a black-body at a given temperature to be determined:

$$W(T) = \int W(\lambda, T) R(\lambda) d\lambda \ . \tag{2}$$

For our sensors, the responsivity is very flat between 8 and 12 (3 to 5 respectively) microns, so we can simply integrate Equation 1 for λ between 8 and 12. The Planck curve and the integration process are illustrated in Figure 2.

One can achieve marginally higher precision by taking measurements at multiple temperatures and obtaining the gains and offsets by least squares regression. For the case

of indoor thermal images containing human faces, we take each of the two fixed temperatures to be below room temperature and above skin temperature, to obtain the highest quality calibration for all levels of IR emission.

It should be noted that calibration has a limited life span. If a LWIR/MWIR camera is radiometrically calibrated indoors, taking it outdoors where there is a significant ambient temperature difference will cause the gain and offset of linear responsivity of the focal plane array pixels to change. Therefore, radiometric calibration must be performed again. This effect is mostly due to the optics and focal plane array (FPA) heating up, and causing the sensor to 'see' more energy as a result. Also, suppose two sequences are collected with different LWIR/MWIR cameras but with the exact same model number, identical camera settings and under the exact same environmental conditions. Nonetheless, no two thermal IR focal plane arrays are ever identical and the gain and offset of corresponding pixels between these separate cameras will be different. Radiometric calibration standardizes all thermal IR data collections, whether they are taken under different environmental conditions, with different cameras, or at different times.

3. MODELING SKIN IN THERMAL IR

For the purposes of segmentation, we classify pixels in indoor scenes as belonging to one of three classes: exposed skin, covered skin (by clothing or hair), and background (everything else).

Our goal is a probabilistic model that can help segment these three regions. This takes the form:

$$P(c|r) \equiv P(c_i|r), \tag{3}$$

where $c \in C = \{c_S, c_C, c_B\} = \{c_1, c_2, c_3\}$ are the three classes of interest. Of course, in general we could classify into n categories $c_1, c_2, \ldots, c_n$. The images we are segmenting come from a calibrated image, R, and r is the radiance coming from the pixel in R under consideration.

We use a time-dependent model for the scene's flux probability density of the form

$$P^t = \sum_{i=1}^{n} \pi_i^t P_i^t, \qquad \sum_{i=1}^{n} \pi_i^t = 1, \tag{4}$$

where P_i^t is the class-conditional density for class c_i, and π_i are the class priors at time t. The initial class-conditional densities P_i^0 are obtained from training data, by hand segmenting the desired classes in a video sequence, and estimating the corresponding densities. Note that we assume no parametric form for P_i^t, for any $t \geq 0$.

While the densities P_i^0 can be estimated from training data, the class priors π_i^t cannot, since the relative abundance of each class in the training sequence may not be representative of that in the test data. However, given an image frame R^t at time $t > 0$, we can estimate the optimal priors in the maximum likelihood sense. At this point it should become clear why we choose to segment into three classes. Although we are interested only in skin and background for the purposes of detection and tracking, modeling the background as two components is useful. Since our prior model of background cannot take into account all possibilities, we break it into the two most prominent components, those being covered skin and room temperature office objects. Since we do not know the relative frequencies of these classes ahead of time, we adjust them along with the prior for the skin class at this point. This is done by maximizing the logarithmic likelihood of R^t as a function of the priors, given as

$$\log L = \sum_{r \in R^t} \log \left[\sum_{i=1}^{n} \pi_i^t P_i^t(r) \right] \tag{5}$$

We can compute the corresponding partial derivatives as

$$\frac{\partial \log L}{\partial \pi_j^t} = \sum_{r \in R} \frac{P_j^t(r)}{\sum_{i=1}^{n} \pi_i^t P_i^t(r)}. \tag{6}$$

Note that the maximization of Equation 5 should be performed for $\pi^t = (\pi_1^t, \pi_2^t, \ldots, \pi_n^t)$ in the convex constraint set

$$K = \{\mathbf{x} \in \mathbb{R}^n \mid \sum_{i=1}^{n} x_i = 1,\, x_i \geq 0\}, \tag{7}$$

where subscripts denote coordinate components, and $n = 3$ in our case. Geometrically, K is the face of the unit simplex in $\mathbb{R}^n$ which does not lie on a coordinate plane (see Figure 3).

In order to maximize Equation 5 numerically, we use the method of iterated projections [9]. Let $\Phi_K : \mathbb{R}^n \rightarrow K$ denote the projection map onto the set K (guaranteed to exist by the convexity of K). The method of iterated projections proceeds according to the following scheme

$$\pi^{t,k+1} = \Phi_K(\pi^{t,k} + \epsilon(\nabla \log L)(\pi^{t,k})), \tag{8}$$

with $\epsilon > 0$. This scheme converges to a local maximum $\pi^{t,\infty}$ of Equation 5. It remains to show the construction of Φ_K. For $\mathbf{x} \in \mathbb{R}^n$, the projection onto the hyperplane passing through the standard basis vectors e_i, $i = 1, \ldots, n$, is given by

$$(\Phi_K^1(\mathbf{x}))_i = x_i + \frac{1}{n}(1 - \sum_{j=1}^{n} x_j). \tag{9}$$

The projection from the hyperplane onto the constraint set K is given by

$$(\Phi_K^2(\mathbf{x}))_i = \begin{cases} 0 & \text{if } x_i < 0 \\ x_i + \frac{1-S^+}{C^+} & \text{otherwise} \end{cases} \tag{10}$$

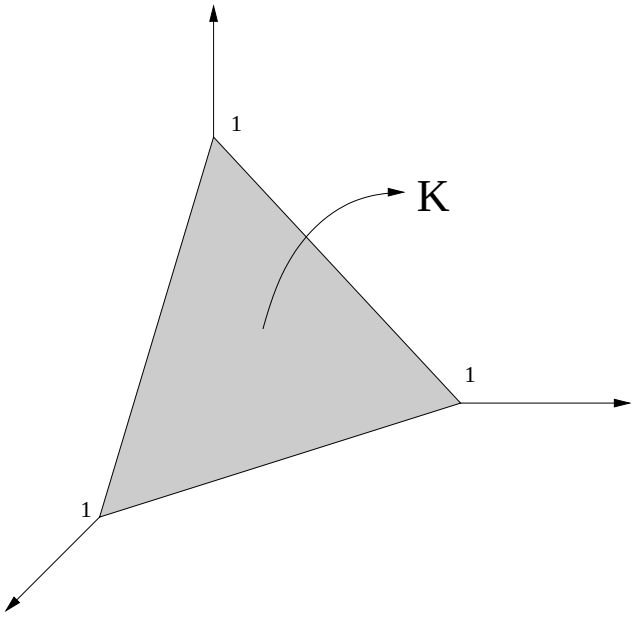

Fig. 3. Constraint set for the maximum likelihood estimation of mixture parameters.

where S^+ is the sum of the non-negative coordinates of $\mathbf{x}$, and C^+ is the number of strictly positive coordinates. Now the desired projection Φ_K can be written as $\Phi_K = \Phi_K^2 \circ \Phi_K^1$.

In order to obtain a good estimate of the global maximum likelihood, we use the scheme in Equation 8 within an iterative random restarts loop. Multiple runs of the constrained local optimization are performed starting at random perturbations of the best local optimum in order to locate a (hopefully) global optimum. As with any non-convex global problem, we have no guarantee of finding the global maximum, however, in experiments with synthetic data the method described above is able to find close-to-optimal solutions.

Once the optimal class priors have been determined, the mixture (Equation 4) constitutes our estimate of the probability density for the thermal flux in the current video frame. Recall that our intermediate goal is to obtain the posterior probabilities for each of our constituent classes (skin, covered skin and background). These can be computed from the information at hand via Bayes rule, which in this context takes the form

$$P^t(c_j|r) = \frac{\pi_j^t \, P_j^t(r)}{\sum_{i=1}^n \pi_i^t P_i^t(r)}. \qquad (11)$$

Equation 11 allows us to compute, for any given pixel in the current frame, the likelihood of belonging to each class, and thus constitutes a soft segmentation of the image. This process can be repeated for each incoming frame, to obtain updated estimates based on the latest data. While

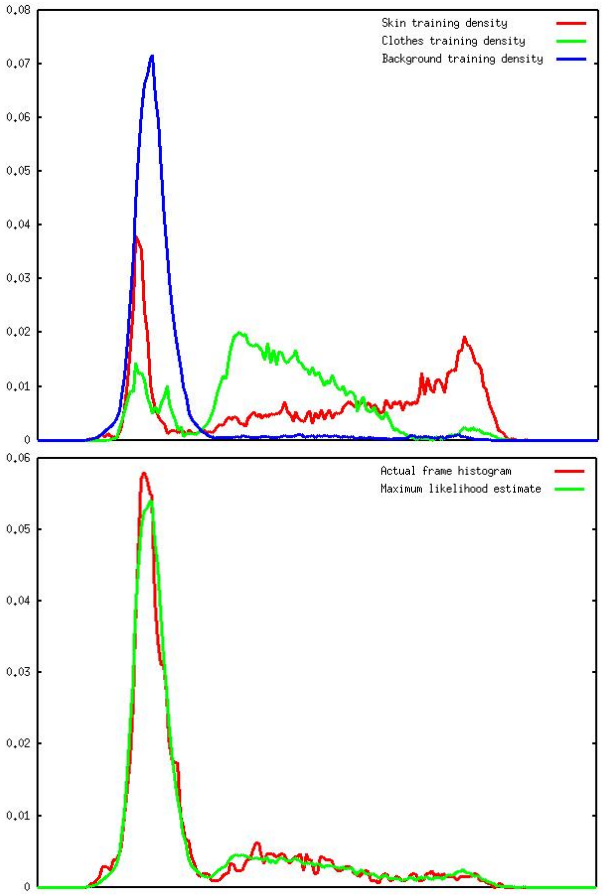

Fig. 5. Left: Training densities for skin, clothing and background. Right: Comparison of the histogram for the LWIR frame in Figure 4 with maximum likelihood estimate based on the training densities on the left.

we have adorned our class conditional densities P_i^t with a superscript denoting time, we have not yet explained how those densities vary as t changes. The time-adaptation of the class conditional densities is discussed in Section 5, after the tracking procedure has been introduced.

A result of applying the posterior density estimation described above to a LWIR/MWIR image is shown in Figure 4, where we can see probability images for exposed skin, covered skin and background. Some intuition into the maximum likelihood method may be gained from Figure 5, where we see the training densities (building blocks), together with a comparison of the actual frame histogram with the ML estimate. Note how the class priors are correctly estimated, yielding a very good semi-parametric estimate of the actual density.

We should mention that it is possible to incorporate the class priors into the variable representing the tracking state (see Section 4), but doing so increases the dimensionality of

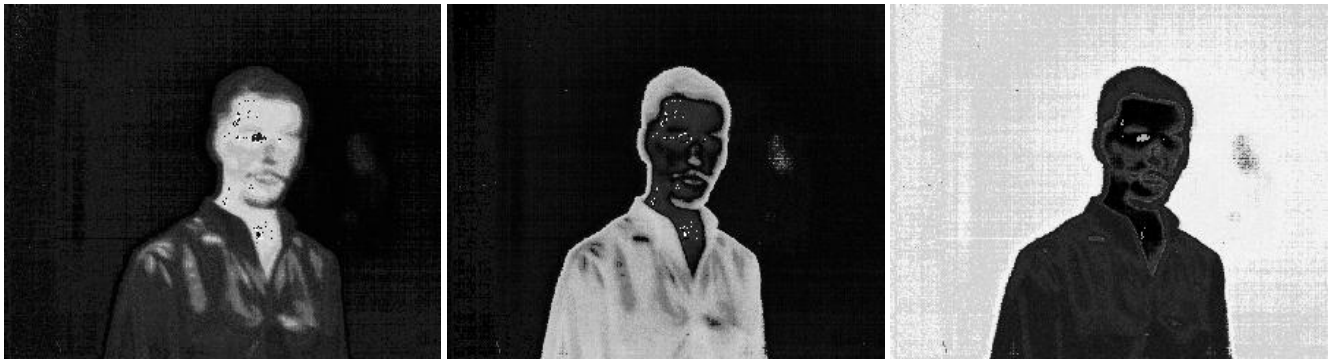

Fig. 4. Probabilities of skin, clothing or hair covered skin and background computed via Equation 11.

the state space. This degrades our ability to estimate probability densities, and sharply increases the computational cost of such estimation. We avoid this by separating the estimation of class priors from the tracking stage itself. Furthermore, the class priors are independent of the position of the tracked object within the scene, and of its motion parameters, so by decoupling their estimation from that of the tracking state variable we incur no accuracy cost.

4. TRACKING SKIN IN THERMAL IR

Let us review the basics of Bayesian tracking. Once a soft segmentation has been computed, we can use it to track faces in the scene. We model faces simply as arbitrarily oriented ellipses, with variable sizes and positions. Let $\mathcal{X}$ denote the state space of all such ellipses. The task of tracking consists of selecting an element of $\mathcal{X}$ for each video frame at time t. More generally, one normally wishes to estimate a probability density on the state space, encoding the likelihood that the tracked object is in a given configuration. Once this density has been computed, a number of estimators can be applied in order to recover the single state which we believe best corresponds to the object's parameters. In our case, we use the MAP estimator, which simply selects the state with highest likelihood.

Reasoning on the relative likelihood of different states in $\mathcal{X}$ is based on a series of independent observations up to time t, $\mathcal{Z}_t = \{\mathbf{Z}_1, \mathbf{Z}_2, \ldots, \mathbf{Z}_t\}$, corresponding to consecutive frames of video. These observations are related to states in $\mathcal{X}$ via an observation model $Q(\mathbf{Z}_t|\mathbf{x}_t)$. Furthermore, successive states are related via a dynamical model $\mathbf{x}_t = f(\mathbf{x}_{t-1})$, which expresses the physics that govern the system. For instance, it might relate velocity to position, or it might describe how uncertainty increases over time. Although this rule may not be known explicitly, one assumes that there is some model of it that is known, and is expressed as $Q(\mathbf{x}_t|\mathbf{x}_{t-1})$.

Armed with these models, and the observations $\mathcal{Z}_t$, it is possible to determine $Q_t(\mathbf{x}_t|\mathcal{Z}_t)$, the desired probability density on the state space $\mathcal{X}$, by an application of Bayes' rule combined with the dynamics equation to obtain:

$$Q_t(\mathbf{x}_t|\mathcal{Z}_t) = \frac{Q_t(\mathbf{Z}_t|\mathbf{x}_t)}{Q_t(\mathbf{Z}_t)} \times \int_{\mathbf{x}_{t-1}} Q_t(\mathbf{x}_t|\mathbf{x}_{t-1}) Q_{t-1}(\mathbf{x}_{t-1}|\mathcal{Z}_{t-1}) d\mathbf{x}_{t-1} \quad (12)$$

If the densities in Equation 12 are not Gaussian, and/or the dynamics model is non-linear, the numerical computation of Equation 12 is complex and time consuming, as there is no closed-form method for evaluating the left-hand-side. As is now standard in the tracking literature, we use a particle filter method for estimating the posterior densities, as in the CONDENSATION [10, 11] algorithm.

To evaluate $Q(\mathbf{Z}_t|\mathbf{x}_t)$, where $\mathbf{Z}_t$ is an observation (i.e. an image), we look at a series of fixed-length segments piercing the boundary of the ellipse. In order to minimize computational time, the segments lie on a fixed grid on the image plane, as shown in Figure 6. We wish to compute the likelihood that a conjectured model (an ellipse in this case) explains the observed data. That is the likelihood that a given ellipse coincides with the boundary of a face. If the segments piercing the ellipse's boundary are not too close to each other, this can be well approximated by the product of the likelihoods that each single segment intersects the boundary of a face. Thus it remains to compute that likelihood for a single segment.

This quantity is computed under the hypothesis that the conjectured ellipse does indeed correspond with a face boundary. Under that hypothesis, we can assume that radiances on the 'interior' half of the segment are drawn independently and identically distributed with respect to the class conditional density for skin, as estimated in Section 3. On the 'exterior' half of the segment, we can assume that the radiances have been drawn likewise, but with respect to the complementary density (the probability of not-skin). In order to cope with the uncertainty as to the appearance of pix-

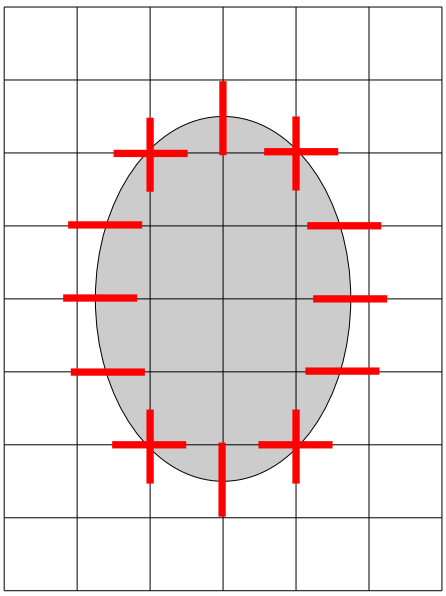

Fig. 6. Example of segments on a fixed grid used to compute the likelihood of the conjectured model given the data.

els immediately nearby the boundary and the non-elliptical nature of the human face, we may ignore radiances coming from pixels on the segment immediately adjacent to the boundary of the ellipse. Decomposing the i^{th} segment as a union of interior, exterior and middle, we write the corresponding likelihood as

$$SL_i = \left[\prod_{interior} P_s(r)\right]\left[\prod_{exterior} P_{ns}(r)\right], \qquad (13)$$

and therefore

$$Q(z|x) = \prod_{i=1}^{N} SL_i, \qquad (14)$$

where N is the number of segments piercing the ellipse $x \in \mathcal{E}$. In order to obtain a more robust estimate of the relative likelihoods for different hypothesized states, instead of taking the product of the individual pixel posterior likelihoods, we average the likelihoods for groups of three pixels along the piercing segment, and multiply those values as in Equation 13.

Since faces are positioned above the neck, there is normally no direct transition from skin to background or clothing at the bottom boundary of the face. Computing Equation 13 including segments piercing the lower portion of the ellipse would bias our likelihood estimate, since we do not expect to see skin/not-skin transitions in that area. Therefore, we ignore an elliptical arc spanning an angle of 45° about the bottom of the ellipse.

We use a simple dynamical model of the form

$$Q_t(\mathbf{x}_t|\mathbf{x}_{t-1}) = K(x_{t-1} - v_t), \qquad (15)$$

where K is a kernel composed of two Gaussian distributions, as pictured in Figure 7. One of the Gaussians has small variance, and is intended to thoroughly explore the area of state space where the maximum likelihood is expected to occur. The second, higher variance component is meant to explore the surrounding area in case the estimated position for the maximum likelihood state is incorrect. The v_t term in Equation 15 is an estimate of the velocity at which the tracked object is moving in the plane, and thus all but two of its five coordinates are zero. We obtain an estimate of v_t using a recursive filter on instantaneous estimates $\hat{v}_t$, as follows

$$\hat{v}_t = E[Q_{t-1}(\mathbf{x}_{t-1}|\mathcal{Z}_{t-1})] - E[Q_{t-1}(\mathbf{x}_{t-2}|\mathcal{Z}_{t-2})] \qquad (16)$$

$$v_t = \beta\,\hat{v}_t + (1-\beta)\,\hat{v}_{t-1}, \qquad (17)$$

where β is the learning rate, which we set at 0.25. These simple dynamics give us a rather robust estimate of velocity, under the assumption of generally uniform motion. The estimate degenerates somewhat when the tracked object speeds up or changes direction, but it recovers quickly, as the high variance component of Equation 15 searches for the object in an extended neighborhood of its expected location.

Much like for the class priors estimated in Section 3, it is possible (maybe even customary [10, 11, 12, 13]) to include the velocity parameters into the state variable for the tracked object. However, once again, this increases the dimensionality of the state space, thus rendering density estimation less accurate and more computationally expensive. For that reason, we separate the velocity estimation from that of the other state variables. Our velocity estimate is not meant to be very precise, just a rough guess to aid the Bayesian tracker by exploring more thoroughly the space about the expected location of the target. Experimentally, we see that the estimate is accurate enough to achieve this goal.

5. TRACKING-DRIVEN ADAPTATION

Section 3 introduced the ideas involved in estimating the posterior flux densities for each material class, based on hand-labeled training data, maximum likelihood estimation of class priors and Bayes Rule. However, it did not deal with the issue of adaptation of the class conditional densities, moving them away from the training data to better fit the observed radiances. That is the subject of the current section.

Let us suppose for a moment that at a given time step, we know the tracker has successfully located the skin region. Under this assumption, we have access to a sample of radiances from the skin class (those pixels in the interior of the region) which is more representative of the appearance of skin in this particular video sequence, and for this particular individual, than the training data used to estimate the skin density. Therefore we should be able to update our skin

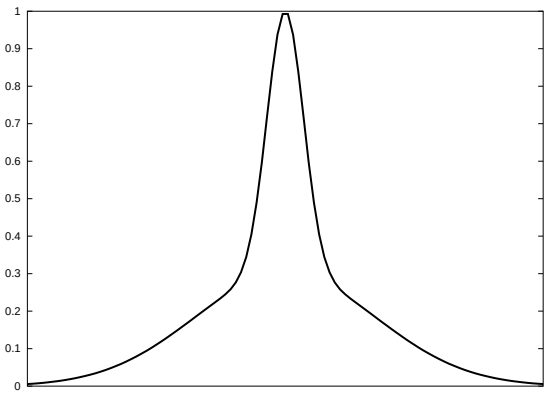

Fig. 7. Mixture of Gaussians kernel used in the dynamics model.

density estimate to better match the current conditions. Of course, in reality we do not know whether the tracker has correctly located the skin area, and thus cannot directly update our estimates. We now propose a confidence criterion which will allow us to carry through a similar procedure without resorting to unobtainable ground truth information.

The result of our CONDENSATION tracker developed in Section 4 is a probability density function Q (or an estimate thereof) on a configuration space representing the possible states of the tracked target in the scene. This density represents the likelihood that the tracked object is in a certain configuration given the current video frame and the history of the object's states. In Section 4 we used the MAP estimator to arrive at the position of the object based on this density. We propose to use the uncertainty in Q as an indication of how likely the tracker is to have succeeded in locating the target. The uncertainty in Q can be measured by its entropy:

$$H(Q) = \int_{\mathcal{E}} Q \log Q, \tag{18}$$

where the logarithm is understood to be zero wherever Q vanishes. It can be shown that for continuous spaces, the distribution with lowest entropy is the delta distribution, and the highest entropy for a fixed variance is attained by the normal distribution with that variance. For discrete probability spaces, the lowest entropy is achieved by a distribution whose mass is concentrated at a single point, and the highest by the uniform distribution.

Estimating the uncertainty of a high-dimensional non-parametric density is not a computationally straightforward task. This is a direct consequence of Bellman's curse of dimensionality. [14] While the entropy as defined in Equation 18 gives an exact measure of the uncertainty of the underlying distribution, we can construct a related measure which can be estimated much more readily. Let the underlying space of the probability density Q be $\mathbb{R}^n$, and for $1 \leq i \leq n$, denote by Q_i the marginal density of Q with respect to the i^{th} coordinate:

$$Q_i(x_i) = \int Q(x_1, \ldots, x_n) dx_1 \ldots \hat{dx_i} \ldots dx_n, \tag{19}$$

where $\hat{dx_i}$ denotes ommission of the i^{th} coordinate differential. The chain rule for entropies states that

$$H(Q) = \frac{1}{n} \sum_{i=1}^{n} H(Q_i) + \frac{1}{2} \sum_{i,j=1}^{n} H(Q_i|Q_j), \tag{20}$$

where the second term is the sum of the relative entropies between the respective marginals. On the other hand, we also have that

$$H(Q) = \sum_{i=1}^{n} H(Q_i) - \frac{1}{2} \sum_{i,j=1}^{n} I(Q_i, Q_j), \tag{21}$$

where the second term is the sum of the mutual informations between the respective marginals. It follows from Equations 20 and 21 that

$$\frac{1}{n} \sum_{i=1}^{n} H(Q_i) \leq H(Q) \leq \sum_{i=1}^{n} H(Q_i), \tag{22}$$

with equality attained on the left when all marginals of Q are equal, and on the right when Q is a product of its marginal densities.

As a consequence of the previous discussion, for a fixed dimension, the average entropy of the marginals of Q, denoted $\bar{H}(Q)$, is uniformly equivalent to the entropy of Q itself, and thus is equally useful to us as a measure of the uncertainty in Q. In contrast with $H(Q)$, however, $\bar{H}(Q)$ can be easily estimated even for non-parametric densities, since it relies only on one-dimensional computations.

We can learn the relative entropy values for 'good' and 'bad' tracking by observing the tracker's behavior on training sequences. From this we obtain upper and lower bounds for the mean marginal entropy of the posterior state-space density Q_t, denoted $\bar{H}_{\min}$ and $\bar{H}_{\max}$. Using these learned bounds, we define the adaptation rate by

$$\alpha_t = \min(\max(\frac{\bar{H}(Q_t) - H_{\min}}{\bar{H}_{\max} - \bar{H}_{\min}}, 0), 1). \tag{23}$$

Using the adaptation rate, we modify our class-conditional density for skin, P_1^{t+1} as follows

$$P_1^{t+1} = \alpha_t \, P_1^t + (1 - \alpha_t) \, \xi_t, \tag{24}$$

where ξ_t is a non-parametric estimate of the greyvalue distribution strictly inside the MAP tracking state.

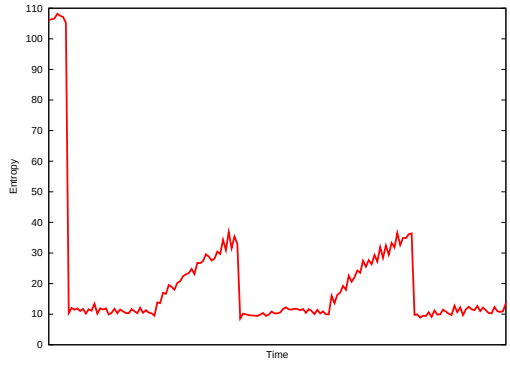

Fig. 8. Entropy as a function of time for an intentionally distracted CONDENSATION tracker. Note the rise and fall in entropy as the tracker loses and re-acquires the target.

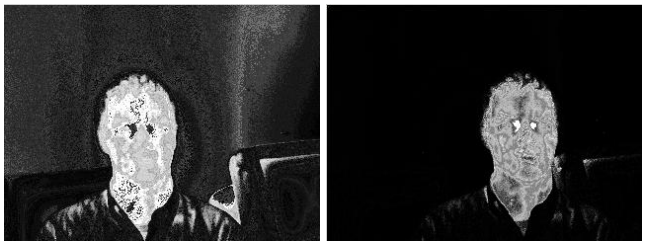

Fig. 9. Estimates of the probability of skin before and after adaptation.

Figure 9 shows the effect of adaptively learning the probability density for skin based on entropy-weighted tracking results. These images show the (posterior) probability-of-skin images without and with adaptation. Note how some background objects such as the warm chassis of a computer monitor look somewhat like skin on the image on the left. Additionally the pixels on the wall behind the subject have a small, but non-negligible probability of skin. With tracking-driven adaptation we obtain the image on the right, where essentially all distractors have been ruled out, and the only pixels with significant probability of being skin are those on the subject's face.

6. EXPERIMENTAL RESULTS

We performed tracking experiments on several LWIR and MWIR indoor image sequences. Training and testing data was acquired and calibrated according to the procedures outlined in Section 2. Using a typical hand-segmented frame from the training sequence, we create initial histogram estimates for the class-conditional densities of skin, clothing or hair-covered skin, and background. These, together with a test sequence, are the initial inputs to our tracker. Typical results for the soft segmentation portion of the algorithm can be seen in Figure 4.

Detection of the face and tracker initialization are done automatically by seeding the first image frame with a uniform distribution of particles representing ellipses at different positions, sizes, and orientations. After the first frame, the estimated densities evolve according to the Bayesian tracking procedure in Section 4. Figure 10 shows the initial estimate of the face location, obtained without operator intervention, and four typical frames from an 80 frame MWIR sequence. Note that the initial estimate is rather good, but it greatly improves as the tracking process evolves.

We collected ground-truth positions of the subject's head for the sequence in Figure 10, by hand-placing the ellipse best fitting the face in each frame. Figure 11 shows a comparison of ground-truth $x-$ and y-coordinates of the center of the ellipse for each of the 80 frames in the sequence. The mean absolute (L^1) error for the x-coordinate estimate is 1.6 pixels, with a standard deviation of 1.37 pixels, while for the y-coordinate we obtain a mean error of 2.9 pixels, with a standard deviation of 2.0 pixels. We should note that estimating the vertical position of the best bounding ellipse is a harder task for both the human operator and the automated tracker, since the hairline and chin are not as clear in the images as the sides of the face.

We can validate the dynamical model in Equation 15 by comparing the velocity estimated using the recursive filter (Equation 17) with that obtained by differencing the ground-truth position data. This comparison is shown in Figure 11, in which we see that after a brief period of uncertainty, the estimate is a good approximation to the observed velocities. Recall that this estimate is obtained without adding velocity variables to the state space, and is therefore not the same as that which we might obtain by differencing the position estimates. In fact, our velocity estimate is predictive, and it does not lag one frame behind the position estimate.

Lastly, we should mention the computational cost involved in tracking using this method. Up to the present time, we have made no effort to optimize the tracker for real-time performance. As a result, it currently takes approximately 5 seconds to process each frame, on dual PIII 1Ghz computer with 512Mb of memory. This falls clearly short of the real-time standard, but we believe that judicious optimization can yield processing times of over 5Hz on currently existing off-the-shelf hardware.

7. CONCLUSIONS

We introduced a methodology for tracking human faces in calibrated thermal infrared imagery. The use of calibrated imagery allows us to use training data acquired before the tracking stage, with the assurance that its thermal flux values will be comparable to those measured during tracking.

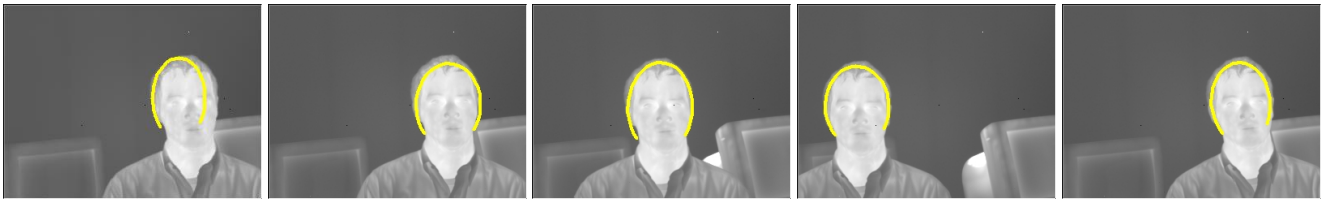

Fig. 10. Initial estimate of face location (left), and tracking results for frames 7, 18, 38 and 78 of an 80 frame MWIR sequence.

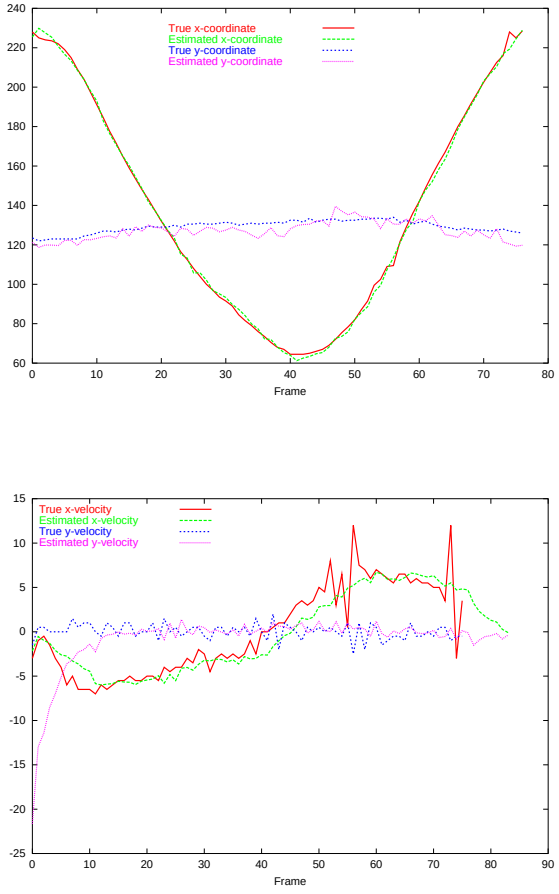

Fig. 11. Top: Estimates and ground-truth positions for MWIR sequence in Figure 10. Bottom: Estimates and ground-truth velocities for MWIR sequence in Figure 10

This calibration step is a critical difference between tracking in the visible and thermal domains. When working with visible imagery, one has to worry about the effects of lighting and color constancy (or lack thereof) on the acquired data. In sharp contrast with that situation, calibrated thermal imagery gives us lighting invariant data, very suitable for robust tracking.

Our method consists of several portions of independent interest. Firstly we outline a scheme for modeling and segmenting skin, covered skin and background in thermal imagery based on training data plus a maximum likelihood estimation step. We model each frame's histogram as a mixture of non-parametric densities, and estimate the class priors in order to best approximate the actual flux density.

Tracking proceeds along the lines of the CONDENSATION algorithm, with a likelihood model based on the posterior densities for each of the three segmented classes. In section 6 we show results of applying the tracker to standard indoor MWIR scenes. Furthermore, we compare our results to hand-measured ground truth positions and velocities. This comparison shows that our tracker is very accurate, usually within 2 pixels of the hand-labeled face centroid.

We go beyond simple tracking, by allowing the tracking results to feed back into the segmentation stage. This is accomplished using the average marginal entropy of the posterior state density as a measure of tracking accuracy. Based on this measure, we can adapt our class-conditional densities, thus yielding a better segmentation (and therefore tracking) that can be obtained with pre-computed training data alone.

A number of extensions of this work should be considered, including handling multiple subjects and the resulting occlusion problem. The most interesting area for future refinement of this method lies within the feedback loop between the tracking and segmentation stages. While we provide an effective means of adapting the class-conditional densities based on an entropy weighting criterion, there is certainly room for further experimentation in this area. As we mentioned above, real-time implementation of the algorithm remains a challenge, but we believe it possible on off-the-shelf hardware, after careful optimization. The promising preliminary results in this paper, together with the illu-

mination invariance properties afforded by thermal infrared imagery, make this an attractive avenue of investigation.

8. REFERENCES

[1] B. Menser and F. Muller, "Face detection in color images using principal components analysis," in *Proceedings Seventh International Conference on Image Processing and its Applications*, July 1999, vol. 2, pp. 620–624.

[2] C. Stauffer and W.E.L. Grimson, "Adaptive background mixture models for real-time tracking," in *CVPR*, June 1998, vol. 2.

[3] C. K. Eveland, K. Konolige, and R. Bolles, "Background modeling for segmentation of video-rate stereo sequences," in *CVPR*, 1998.

[4] S. Birchfield, "Elliptical Head Tracking Using Intensity Gradients and Color Histograms," in *Proceedings CVPR 1998*, June 1998.

[5] M. Isard and J. MacCormick, "Bramble: A bayesian multiple-blob tracker," in *Proc. 8th Int. Conf. Computer Vision*, 2001.

[6] L. Wolff, D. Socolinsky, and C. Eveland, "Quantitatve measurement of illumination invariance for face recognition using thermal infrared imagery," in *Proceedings CVPR Workshop on Computer Vision Beyond the Visible Spectrum*, December 2001.

[7] J.C. Terrillon, M. David, and S. Akamatsu, "Automatic Detection of Human Faces in Natural Scene Images by Use of a Skin Color Model and of Invariant Moments," in *Proceedings Third International Conference on Face and Gesture Recognition*, 1998, pp. 112–117.

[8] Randal C. Nelson, "Qualitative detection of motion by a moving observer," *Internation Journal of Computer Vision*, vol. 7, no. 1, pp. 33–46, November 1991.

[9] J.-L. Lions R. Glowinski and R. Trémolières, *Numerical Ananlysis of Variational Inequalities*, North-Holland, Amsterdam, 1981.

[10] M. Isard and A. Blake, "Contour Tracking by Stochastic Propagation of Conditional Density," in *Proceedings ECCV*, 1996.

[11] M. Isard and A Blake, "Condensation - conditional density propagation for visual tracking," *International Journal of Computer Vision*, vol. 29, no. 1, pp. 5–28, 1998.

[12] A. Blake, M. Isard, and D. Reynard, "Learning to track the visual motion of contours," *Artificial Intelligence*, vol. 78, pp. 101–134, 1995.

[13] J. MacCormick, *Probabilistic models and stochastic algorithms for visual tracking*, Ph.D. thesis, Oxford University, 2000.

[14] V. Vapnik, *Statistical Learning Theory*, John Wiley and Sons, 1998.