In Proc. of IEEE Conf. on CVPR'2000, Vol.I, pp.133-138, Hilton Head Island, SC, 2000

Color Tracking by Transductive Learning

Ying Wu, Thomas S. Huang
Beckman Institute
University of Illinois at Urbana-Champaign
Urbana, IL 61801, U.S.A.
{yingwu, huang}@ifp.uiuc.edu

Abstract

One of the difficulties of color tracking is that color changes in different lighting conditions, and static color models would be inadequate to capture the non-stationary color distribution over time. Although some work has been done on adaptive color models, this problem still needs further investigation. Different from many other approaches, we formulate the non-stationary color tracking problem as a transductive learning *problem, in which the generalization of a trained color classifier is only defined on the pixels in a specific image, rather than the whole color space. This formulation offers a way to design and transduce color classifiers through non-stationary color distribution. Instead of assuming a color transition model, we assume that some unlabeled pixels in a new image frame can be "confidently" labeled by a "weak classifier" according to a preset confidence level. The proposed Discriminant-EM (D-EM) algorithm offers an effective way to transduce color classifiers as well as automatically select a good color space. Experiments show that D-EM successfully handles some problems in color tracking. As a component in our natural gesture interface, this algorithm gives tight bounding boxes of the hand or face regions in video sequences.*

1 Introduction

In current research of vision-based human computer interaction, the use of human body parts, such as hands and faces, motivates the research of tracking human body movements. Skin color offers an effective and efficient way to localize and track hand and face. The core of color tracking is color-based segmentation. According to the representation of color distribution in certain color spaces, current techniques of color tracking can be classified into two general approaches: non-parametric [12, 8, 7, 15] and parametric [14, 9, 16].

One of the non-parametric approaches is based on color histograms [12, 8, 7]. Since color space is quantized by the structure of the histogram, this technique shares the same problem with non-parametric density estimation, in which the level of quantization will affect the estimation. How to select a good quantization level of the color histogram is not trivial. Although non-uniform quantization would perform better than uniform quantization, it is much more complicated. Another non-parametric approach is proposed in [15] based on the self-organizing map (SOM), an unsupervised clustering algorithm to approximate color distribution. SOM can be viewed as a neural network-based vector quantization (VQ) algorithm. Although standard SOM algorithm also needs to specify the structure of SOM which acts the same role as the level of quantization, the algorithm proposed in [15] has the ability to find an appropriate structure by embedded growing, pruning and merging schemes. Generally, these non-parametric approaches work effectively when the quantization level is properly set and there are sufficient data.

Parametric approaches model the color density in parametric forms such as Gaussian distribution or Gaussian Mixture models [14, 9, 16]. Expectation-Maximization (EM) offers a way to fit probabilistic models to the observation data. The difficulty of *model order selection* could be handled by heuristics [9] or cross-validation.

However, when we try to apply these techniques to track human hand and face in some virtual environment (VE) applications, this problem is still made challenging by some special difficulties such as large variation in skin tone, unknown lighting conditions and dynamic scenes. In order to achieve user-independence, the tracking algorithm should be able to deal with the large variation in skin color for different people. One possible solution is to make a generic statistical model of skin color by collecting a huge training data set [7] so that the generic color model works for every user. However, collecting and labeling

such a huge database is not trivial.

Even though such a good generic color model can be obtained, we have to face another difficulty in color tracking: generic color models would be incapable to handle changing lighting conditions unless some invariants could be found. Many color tracking techniques assume controlled lighting. However, in many cases, the interested object may be shadowed by other objects or by the object itself so that the color looks very different. What is more, we cannot assume constant lighting sources, since the lighting directions, intensities and tones might change. In some VE applications, since the graphics rendered in the display keeps changing, the reflective lights would change the apparent color of objects. This color constancy problem is not trivial in color tracking.

Because of dynamic scenes and changing lighting conditions, the color distribution over time is non-stationary, since the statistics of color distribution will change with time. If a color classifier is trained under a specific condition, it may not work well in other scenarios.

There have been some researchers who have looked into the non-stationary color distribution problem in color tracking. Several methods have been proposed to approach this problem. A scheme of color model adaptation was addressed in [9], in which a Gaussian mixture model was used to represent color distribution, and a linear extrapolation was employed to adjust the parameters of the model by a set of labeled training data drawn from the new frame. However, since the new image is not segmented, this labeled data set is not reliable.

In [15], the scheme of *transduction of SOM* was proposed to update the weights and structure of the trained SOM to capture the new color distribution, according to a set of new training data, which consists of both labeled and unlabeled samples. Since the transduction of SOM combine unsupervised updating and supervised updating, a large number of labeled training data is not required.

In this paper, we try to investigate the problem of non-stationary color distribution in color tracking. We formulate this problem as a *transductive learning* problem, which offers an easier way to design color classifier in non-stationary color distribution. We fit this transductive learning problem into an EM framework. Combining both labeled and unlabeled data, the proposed Discriminant-EM (D-EM) algorithm can automatically select a good color space and relax the assumption of probabilistic structure of color model. Our algorithm has been applied to hand and face tracking. It gives tight bounding boxes of the hand or face regions in video sequences.

2 Color Features and Color Model

Each pixel is associated with a color feature vector. The issue of selecting color features must be addressed here. Different color spaces, such as HSI, RGB, normalized-RGB, have been used in current research. Many color histogram-based techniques use 2-D subspace of these 3-D color spaces, partly because much more storage and searching are needed in 3-D. For example, HSV space is reduced to HS subspace. However, hue and saturation become unstable when the intensity of a pixel is too large or too small, which means that the HS values are meaningless for dark or bright pixels. In some cases, simple intensity thresholding can segment objects well, but using HS would fail. Therefore, Reducing color space will lose some valuable color information.

Although these compact 3-D color spaces have substantial physical meanings, none of them is found to be able to give satisfactory color invariants through different lighting conditions. Considering that HSV color space is not a linear transformation of RGB space, we try to use a higher dimensional color space (6-D) by combining HSV and RGB spaces. Since this higher dimensional color space is redundant, it is not necessary to estimate probabilistic model parameters in such space. Instead, a linear subspace will be found by performing multiple discriminant analysis technique, which will be described in section 5. By this means, good color features for classification can be selected automatically.

Gaussian mixture models are employed here to model the color distribution. Let $\mathbf{x}$ be the color feature vector for each pixel. Its distribution in one image can be described as:

$$p(\mathbf{x}|\mathbf{\Theta}) = \sum_{j=1}^{C} p(\mathbf{x}|O_j;\theta_j)p(O_j) \tag{1}$$

where $\sum_{j=1}^{C} p(O_j) = 1$ and where $p(\mathbf{x}|O_j;\theta_j)$ is the conditional density for a pixel belonging to an object O_j in the image, and it has been parameterized by θ_j, and $\mathbf{\Theta} = \{\theta_j, j = 1, \ldots, C\}$. This conditional density can also be modeled by Gaussian mixtures:

$$p(\mathbf{x}|O_j;\theta_j) = \sum_{k=1}^{T} p(\mathbf{x}|c_k;\theta_{jk})p(c_k) \tag{2}$$

where $\sum_{k=1}^{T} p(c_k) = 1$ and where $p(\mathbf{x}|c_k;\theta_{jk})$ is the conditional density for a pixel belonging to a color

component c_k of the object O_j in the image. Each mixture component can be modeled by Gaussian distribution with mean μ_k and covariance matrix $\mathbf{\Sigma}_k$.

3 A Transductive Problem

It is a good practice to learn a generic color classifier by collecting a large labeled data set[7]. If some color invariants to lighting could be found, learning such a color classifier would suggest a direct and robust way to color tracking. However, when we consider the non-stationary color distribution over time, we do not generally expect to find such invariants. In fact, learning such a highly nonlinear color classifier may not be necessary.

The approach taken in [7] is an *inductive learning* approach, by which the color classifier learned should be able to classify any pixel in any image. Generally, this color classifier would be highly nonlinear, and a huge labeled training data set is required to achieve good generalization. However, the requirement of generalization could be relaxed to a subset of the data space. In color tracking, a color classifier M_t at time frame t could be only used to classify pixel $\mathbf{x}_j$ in the current specific image feature data set I_t so that this specific classifier M_t could be simpler. When there is a new image I_{t+1} at time $t+1$, this specific classifier M_t should be *transduced* to a new classifier M_{t+1} which works just for the new image I_{t+1} instead of I_t. The classification can be described as:

$$y_i = arg \max_{j=1,\ldots,C} p(y_j|\mathbf{x}_i, M_t, I_{t+1} : \forall \mathbf{x}_i \in I_{t+1}) \quad (3)$$

where y_i is the label of $\mathbf{x}_i$, and C is the number of classes. In this sense, we do not care the performance of the classifier M_{t+1} outside I_{t+1}. The *transductive learning* is to transduce the classifier M_t to M_{t+1} given I_{t+1}. Figure 1 shows the transduction of color classifiers.

This *transduction* may not always be feasible unless we know the joint distribution of I_t and I_{t+1}. Unfortunately, such joint probability is generally unknown since we may not have enough *a priori* knowledge about the transition in a color space over time. One approach is to assume a transition model, like the case in motion tracking by Kalman filter or Condensation[1], so that we can explicitly model $p(I_{t+1}|I_t)$. One of the difficulties of this approach is that a fixed transition model is unable to capture much dynamics. Although the issue of motion model switching by learning transition models has been addressed in [1], their scheme is not general. Another difficulty is that it may not be easy to identify parameters of the

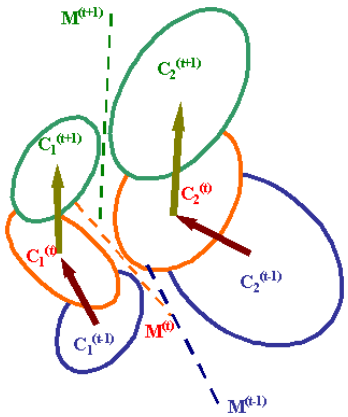


Figure 1: An illustration of transduction of classifiers.

transition models due to the insufficient labeled training data. The approach used in [9] assumes a linear transition model. However, the transition (updating) of color models is plagued since the newest image has not been segmented yet.

However, our assumption is different from the transition model assumption. We assume that the classifier M_t at time t can give "confident" labels to several samples in I_{t+1}, so that the data in I_{t+1} can be divided into two parts: labeled data set $\mathcal{L} = \{(\mathbf{x}_j, y_j), j = 1, \ldots, N\}$, and unlabeled set $\mathcal{U} = \{\mathbf{x}_j, j = 1, \ldots, M\}$, where N and M are the size of the labeled set and unlabeled set respectively, $\mathbf{x}_j$ is the color feature vector, and y_j is its label (such as skin or non-skin). Here, $\mathcal{L}$ and $\mathcal{U}$ are from the same distribution. Consequently, the transductive classification can be written as:

$$y_i = arg \max_{j=1,\ldots,C} p(y_j|\mathbf{x}_i, \mathcal{L}, \mathcal{U} : \forall \mathbf{x}_i \in \mathcal{U}) \quad (4)$$

In this formulation, the specific classifier M_t is transduced to another classifier M_{t+1} by combining a large unlabeled data set from I_{t+1}.

4 The EM framework

The Expectation-Maximization (EM) approach can be applied to this transductive learning problem, since the labels of unlabeled data can be treated as missing values.

The training data set $\mathcal{D}$ is a union of a set of labeled data set $\mathcal{L}$ and a set of unlabeled set $\mathcal{U}$. When we assume sample independency, the model parameters $\mathbf{\Theta}$ can be estimated by maximizing *a posteriori* probability $p(\mathbf{\Theta}|\mathcal{D})$. Equivalently, this can be done by maximizing $\lg(p(\mathbf{\Theta}|\mathcal{D}))$. Let $l(\mathbf{\Theta}|\mathcal{D}) =$

$\lg(p(\mathbf{\Theta})p(\mathcal{D}|\mathbf{\Theta}))$. When introducing a binary indicator $\mathbf{z}_i = (z_{i1}, \ldots, z_{iC})$, where $z_{ij} = 1$ iff $y_i = O_j$, and $z_{ij} = 0$ otherwise, we have:

$$l(\mathbf{\Theta}|\mathcal{D}, \mathcal{Z}) = \lg(p(\mathbf{\Theta})) + \sum_{\mathbf{x}_i \in \mathcal{D}} \sum_{j=1}^{C} z_{ij} \lg(p(O_j|\mathbf{\Theta})p(\mathbf{x}_i|O_j; \mathbf{\Theta}))$$

The EM algorithm estimates the parameters $\mathbf{\Theta}$ by an iterative hill climbing procedure, which alternatively calculates $E(\mathcal{Z})$, the expected values for all unlabeled data, and estimates the parameters $\mathbf{\Theta}$ given $E(\mathcal{Z})$. The EM algorithm generally reaches a local maximum of $l(\mathbf{\Theta}|\mathcal{D})$. It consists of two iterative steps:

- E-step: set $\hat{\mathcal{Z}}^{(k+1)} = E[\mathcal{Z}|\mathcal{D}; \hat{\Theta}^{(k)}]$
- M-step: set $\hat{\Theta}^{(k+1)} = arg\ max_\theta\, p(\Theta|\mathcal{D}; \hat{\mathcal{Z}}^{(k+1)})$

where $\hat{\mathcal{Z}}^{(k)}$ and $\hat{\Theta}^{(k)}$ denote the estimation for $\mathcal{Z}$ and Θ at the k-th iteration respectively.

When the size of the labeled set is small, EM basically performs an unsupervised learning, except that labeled data are used to identify the components. If the probabilistic structure, such as the number of components in mixture models, is known, EM could estimate true probabilistic model parameters. Otherwise, the performance could be very bad. Generally, when we do not have such prior knowledge about the data distribution, a Gaussian distribution is always assumed to represent a class. However, this assumption is often invalid in practice, which is partly the reason that unlabeled data hurt the classifier.

5 D-EM: A Transduction Algorithm

EM often fails when structure assumption does not hold. One approach to this problem is to try every possible structure and select the best one. However, it needs more computational resources. An alternative is to find a mapping such that the data are clustered in the mapped data space, in which the probabilistic structure could be simplified and captured by simpler Gaussian mixtures. MDA offers a possible way to relax the assumption of probabilistic structure, and the EM supplies MDA enough labeled data to identify most discriminating features for classification.

5.1 Multiple Discriminant Analysis

Multiple Discriminant Analysis (MDA) [3] is a natural generalization of Fisher's linear discrimination (LDA) in the case of multiple classes. MDA offers many advantages and has been successfully applied to many tasks such as face recognition. The basic idea behind MDA is to find a linear transformation $\mathbf{W}$ to map the original d_1 dimensional data space to a new d_2 space such that the ratio between the between-class scatter and within-class scatter is maximized in the new space.

MDA offers a means to catch major differences between classes and discount factors that are not related to classification. Some features most relevant to classification are automatically selected or combined by the linear mapping $\mathbf{W}$ in MDA, although these features may not have substantial physical meanings any more. Another advantage of MDA is that the data are clustered to some extent in the projected space, which makes it easier to select the structure of Gaussian mixture models. Details can be found in [3].

5.2 D-EM Algorithm

It is apparent that MDA is a supervised statistical method, which requires enough labeled samples to estimate some statistics such as mean and covariance. By combining DMA with the EM framework, our proposed method, Discriminant-EM algorithm (D-EM), is such a way to combine supervised and unsupervised paradigms. The basic idea of D-EM is to identify some "similar" samples in the unlabeled data set to enlarge the labeled data set so that supervised techniques are made possible in such an enlarged labeled set.

D-EM begins with a weak classifier learned from the initial labeled set. Certainly, we do not expect much from this weak classifier. However, for each unlabeled sample $\mathbf{x}_j \in \mathcal{U}$, the classification confidence $\mathbf{w}_j = \{w_{jk}, k = 1, \ldots, C\}$ can be given based on the probabilistic label $\mathbf{l}_j = \{l_{jk}, k = 1, \ldots, C\}$ assigned by this weak classifier.

$$l_{jk} = \frac{p(\mathbf{W}^T\mathbf{x}_j|O_k)p(O_k)}{\sum_{k=1}^{C} p(\mathbf{W}^T\mathbf{x}_j|O_k)p(O_k)} \tag{5}$$

$$w_{jk} = \lg(p(\mathbf{W}^T\mathbf{x}_j|O_k))\ k = 1, \ldots, C \tag{6}$$

Euqation(6) is just a heuristic to weight unlabeled data $\mathbf{x}_j \in \mathcal{U}$, although there may be many other choices.

After that, MDA is performed on the new weighted data set $\mathcal{D}' = \mathcal{L} \bigcup \{\mathbf{x}_j, \mathbf{l}_j, \mathbf{w}_j : \forall \mathbf{x}_j \in \mathcal{U}\}$, which is linearly projected to a new space of dimension $C-1$ but unchanging the labels and weights, $\hat{\mathcal{D}} = \{\mathbf{W}^T\mathbf{x}_j, y_j : \forall \mathbf{x}_j \in \mathcal{L}\} \bigcup \{\mathbf{W}^T\mathbf{x}_j, \mathbf{l}_j, \mathbf{w}_j : \forall \mathbf{x}_j \in \mathcal{U}\}$. Then parameters $\mathbf{\Theta}$ of the probabilistic models are estimated on $\hat{\mathcal{D}}$, so that the probabilistic labels are given by the Bayesian classifier according to Equation(5). The algorithm iterates over these three steps, "Expectation-Discrimination-Maximization". The algorithm can be

terminated by several methods such as presetting the iteration times, comparing a threshold and the difference of the parameters between consecutive two iterations, and using cross-validation.

It should be noted that the simplification of probabilistic structures is not guaranteed in MDA. If the components of data distribution are mixed up, it is very unlikely to find such a linear mapping. Our experiments show that D-EM works better than pure EM.

5.3 Tracking by D-EM

The application of D-EM to color tracking is straightforward. In our current implementation, in the transformed space, both classes (foreground and background) are represented by a Gaussian distribution with three parameters, the mean μ_i, the covariance $\mathbf{\Sigma}_i$ and *a priori* probability P_i.

We use three schemes to bootstrap the tracking. The first method is by manually collecting and labeling some pixels (100 samples) from both the interested object and background. An alternative is by putting the interested object in the middle of the image so that some data can be automatically collected. The third method is to detect the moving region by image differences in the first several frames. We assume that we are interested in the object with the largest motion.

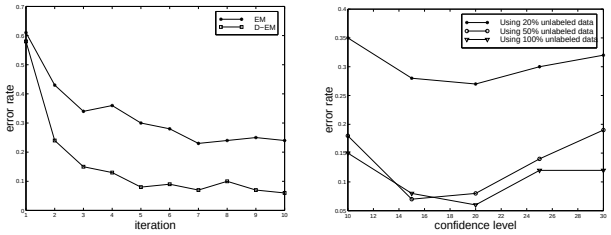
For each new image I_t, by setting a confidence level, the color classifier M_{t-1} at time $t-1$ divides I_t into two parts: labeled set $\mathcal{L}_t$ and unlabeled set $\mathcal{U}_t$. $\mathcal{L}_t$ is confidently labeled by M_{t-1}. The D-EM algorithm identifies some "similar" samples in $\mathcal{U}_t$ to the labeled samples in an unsupervised sense. Therefore, good discriminating color features can be automatically selected through the enlarged labeled data set. After a Bayesian classifier is designed in the new feature space, it is used to probabilistically label I_t. Through several iterations, the classifier M_{t-1} has been transduced to M_t by D-EM.

6 Experiments

6.1 Simulation

At current time t in tracking, since M_{t-1} may not be able to give a good segmentation on I_t, the image at time t is not labeled (segmented) so that the ground truth for the new data set is not available. However, to evaluate our algorithm, we assume the ground truth is known to calculate classification errors, although such errors are not available in real applications.

We use two "hand images" (resolution 100×75), where I_1 is a segmented image, and I_2 is the same as I_1 except that the color distribution of I_2 is transformed by shifting the R element of every pixel by 20 such that I_2 looks like adding a red filter. A color classifier is learned for I_1 with error rate less than 5%. In this simple situation, this color classifier would fail to correctly segment hand region from I_2, since the skin color in I_2 is much different. Actually, it has error rate of 35.2% on I_2.



(a) (b)

Figure 2: (a) shows the comparison between EM and D-EM. (b) shows the effect of number of labeled and unlabeled data in D-EM

Figure 2.a shows the comparison between EM and D-EM. In this experiment, both EM and D-EM converge after several iterations, but D-EM gives a lower classification error rate (6.9% vs. 24.5%). To investigate the effect of the unlabeled data used in D-EM, we feed the algorithm a different number of labeled and unlabeled samples. The number of labeled data is controlled by the confidence level. In this experiment, confidence level is the same as the size of the labeled set. In general, combining unlabeled data can largely reduce the classification error when labeled data are very few. When using 20% (1500) unlabeled data, the lowest error rate achieved is 27.3%. When using 50% (3750) unlabeled data, the lowest error rate drops to 6.9%. The transduced color classifier gives around 30% more accuracy. Figure 2.b shows the effect of different sizes of labeled and unlabeled data sets in D-EM.

6.2 Hand and Face Localization

This color tracking algorithm is applied to a gesture interface, in which hand gesture commands are localized and recognized to serve as inputs of a virtual environment application. These experiments ran at 15-20Hz on a single processor SGI O2 R10000 workstation.

Figure 3 and Figure 4 show two examples of hand and face localization in a typical lab environment. Both cases are difficult for static color models. In Figure 3, the skin color in different parts of hand are dif-

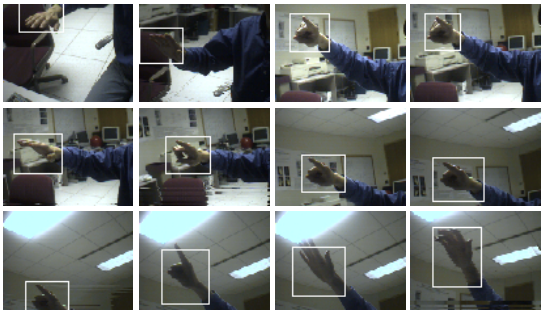


Figure 3: Hand Localization by D-EM

Figure 4: Face localization by D-EM

ferent. The camera moves from downwards to upwards and the lighting conditions on the hand are different. Hand becomes darker when it shades the light sources in several frames. In Figure 4, skin color changes a lot when the head moves back and forth, and turns around.

7 Conclusion

This paper presents a study of the problem of non-stationary color tracking. We formulate this problem as a *transductive learning* problem, which offers a way to design and transduce color classifiers in non-stationary color distribution through image sequences. Instead of assuming a transition model, we assume that some unlabeled pixels in a new image frame can be "confidently" labeled by a "weak classifier" according to a preset confidence level. Integrating discriminant analysis and the EM framework, the proposed Discriminant-EM (D-EM) algorithm offers a means to relax the assumption of probabilistic structures of data distribution and automatically select a good color space. As a component in a natural gesture interface, this algorithm gives tight bounding boxes of the hand or face region in video sequences.

One of the future research directions of this approach is to explore the nonlinear case of MDA. The convergence and stability analysis should be studied in the future work. Currently, the confidence level is an important parameter in the transduction to control the size of labeled set. It needs further investigation.

8 Acknowledgments

This work was supported in part by National Science Foundation Grant IRI-9634618 and Grant CDA-9624396. The authors would like to appreciate the anonymous reviewers for their comments.

References

[1] A.Blake, M.Isard, "Active Contours" Springer-Verlag, 1998.

[2] D. Comaniciu and P. Meer, "Robust Analysis of Feature Spaces: Color Image Segmentation" *Proc. of IEEE CVPR'97*, June 1997, 750-755.

[3] R.Duda and P.Hart, "Pattern Classification and Scene Analysis", New York:Wiley, 1973 (The 2nd Version with D.Stork unpublished)

[4] B.V.Funt and G.D.Finlayson, "Color Constant Color Indexing", *IEEE Trans. on Pattern Analysis and Machine Intelligence*, Vol.17, pp.522-529, 1995

[5] K. Imagawa, S. Lu and S. Igi, "Color-Based Hands Tracking System for Sign Language Recognition", *Proc. of IEEE FG'98*, pp.462-467, 1998.

[6] M. Isard and A. Blake "A mixed-state Condensation tracker with automatic model-switching" *Proc. 6th Int. Conf. Computer Vision*, 1998

[7] M.Jones and J.Rehg, "Statistical Color Models with Application to Skin Detection". *Compaq Cambridge Research Lab. Technical Report CRL 98/11*, 1998

[8] R. Kjeldsen and J. Kender, "Finding Skin in Color Images" *Proc. IEEE FG'96*, pp.312-317, 1996

[9] Y.Raja, S.McKenna and S.Gong, "Colour Model Selection and Adaptation in Dynamic Scenes", *Proc. ECCV'98*, 1998

[10] C.Rasmussen and G.Hager, "Joint Probabilistic Techniques for Tracking Objects Using Multiple Part Objects", *Proc. IEEE FG'98*, 1998.

[11] D.Saxe and R.Foulds, "Toward Robust Skin Identification in Video Images". *Proc. IEEE FG'96*, 1996

[12] M.J.Swain and D.H.Ballard, "Color Indexing", *Int. J. Computer Vision*, Vol.7, No.1, pp.11-32, 1991.

[13] V.Vapnik, "The Nature of Statistical Learning Theory", Springer-Verlag, 1995

[14] C. Wren, A. Azarbayejani, T. Darrel, and A. Pentland, "Pfinder: Real-Time Tracking of the Human Body", *In Photonics East, SPIE Proceedings* vol.2615, Bellingham, WA, 1995.

[15] Y. Wu and T. S. Huang "An Adaptive Self-Organizing Color Segmentation Algorithm with Application to Robust Real-Time Human Hand Localization", *Proc. of Asian Conference on Computer Vision*, Taiwan, 2000

[16] J.Yang, W.Lu and A.Waibel, "Skin-color Modeling and Adaptation", *Proc. of Asian Conference on Computer Vision*, pp.687-694, 1998