Australian Journal of Basic and Applied Sciences, 3(4): 4017-4025, 2009
ISSN 1991-8178

Kernel-Based Hand Tracking

[1]Aras Dargazany, [2]Ali Solimani

[1]Department of ECE, Shahrood University of Technology
[2]Department of ECE, Shahrood University of Technology

Abstract: In this work, a new method is proposed for hand tracking based on a density approximation and optimization method. Considering tracking as a classification problem, we train an approxiator to recognize hands from its background. This procedure is done by extracting feature vector of every pixel in the first frame and then building an approximator to construct a virtual optimized surface of pixels for similarity of the frames which belong to the hand of those frames related to the movie. Received a new video frame, approximator is employed to test the pixels and build a surface. In this method, the features we use is color RGB corresponding to the feature space. Conducting simulations, it is demonstrated that hand tracking based on this method result in acceptable and efficient performance. The experimental results agree with the theoretical results.

Key word: Hand Tracking, Kernel Density, Approximator.

INTRODUCTION

Hand tracking is known as a crucial and basic ingredient computer vision and involves in many research fields of object tracking in recent years. Generally, tracking is the task of finding the object states (including: position, scale, velocity, feature selecting and many other important parameters) obtained from a sequential series of images. Humans can simply recognize and track an object immediately even in the presence of high clutter, occlusion, and non-linear variations in the background as well as in the shape, direction or even the size of the target object. However, hand tracking can be a difficult and challenging task for a machine. If we consider tracking as a classification problem, the choice of a good classifier contributes us to distinguish hands from its constant background. Recently, many solutions are proposed to deal with hands motion. Kernel Density was selected as feature weights to be used in building a surface of similarity. At the next step, in order to obtain the weighting coefficients of the entire image, each image (incoming frame) is processed. A specific hand was considered as the desired object to be tracked in our experiments as reported. Firstly, color frames were converted into gray level images, and then a kernel function is employed; furthermore weights of pixels were obtained for each frame. However, in this work, we propose a novel approach to manipulate similarity coefficients. For each color band, we process each image to obtain these specific coefficients. In addition, the approximator (Birchfield, S., 1998; Black, M. and D. Fleet, 2000) is applied by employing the first frame coefficients and compared with the other sequential frames coefficients. The proposed method offers several advantages. For instance, it can be very resistant against difficulties such as partial occlusion, blurring caused by camera shaking, deformation of object position and any other sorts of translation. This is due to employing color information as feature vectors in the proposed technique. Moreover, it can recognize hands with large aspect of changes. In addition, the proposed method can also be used for both hands. The proposed method can successfully perform real-time tracking algorithm in which we demonstrate that it is capable of tracking hands over long periods of time.

Hand Representation as a Target:

In order to be characterizing the hands as our desirable targets, at first a feature space should be chosen. The reference Hand model as target model is represented by the target probability density function (pdf) shown as q in the chosen feature space. For example, the reference feature space modeled in this work can be selected to be the color pdf of the hands. The hands in the target model can be considered the center of the spatial location 0. In the next frame, a hand as the target candidate is defined at location y, and is distinguished by the pdf $p(y)$.

Corresponding Author: Aras Dargazany, Department of ECE, Shahrood University of Technology
E-mail: aras.dargazany@gmail.com

Each of these pdfs are estimated from the data. In order to satisfy the lower cost of computation which is imposed by using the discrete densities-bin histograms should be initialized. Therefore, we have the two models indicating the location of hand in the first frame, Target Model $\hat{p}(y)$, and in the following frames as well, known as candidate model $\hat{q}(y)$

$$\sum_{u=1}^{m} \hat{p}_u = 1, \sum_{u=1}^{m} \hat{q}_u = 1 \tag{1}$$

The histogram as mentioned above is not the best nonparametric density estimate (Scott, D.W., 1992), but it seems enough for our own purposes. The other discrete density estimates can be also used. We indicate the similarity function between Target and candidate models. The parameter $\hat{\rho}(y)$ plays an important role for likelihood and the local maxima, found in the image identifying the existence of objects in the subsequent frames possessing representations resembling to the hand in the target model which is defined in the first frame.

$$\hat{\rho}(y) = \rho\{\hat{p}(y), \hat{q}\} \tag{2}$$

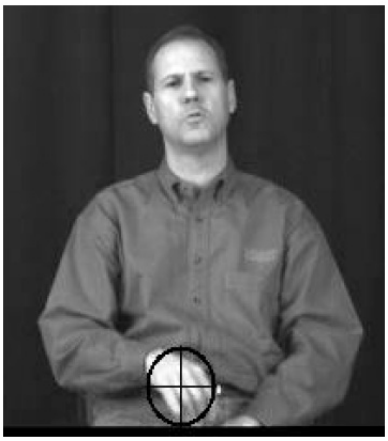

Fig. 1: The coordinates of the feature space of hand in in thethe target model in the first frame

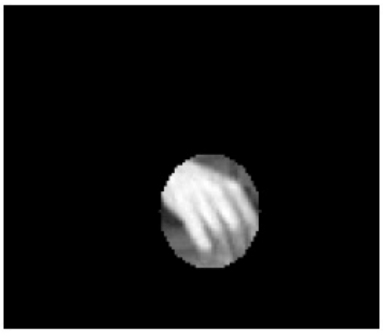

Fig. 2: The complete feature space of the hand target model in the first frame.

If some spectral information is required for specifying the hand in target model, the similarity function $\hat{\rho}(y)$ can be varied largely in the neighboring locations upon the lattice of the image in the frames and the spatial information in feature space is lost. Just for finding the maximum of these functions, the gradient-based optimizing methods seem very difficult to deploy and just an coasty exhaustive search can be utilized.

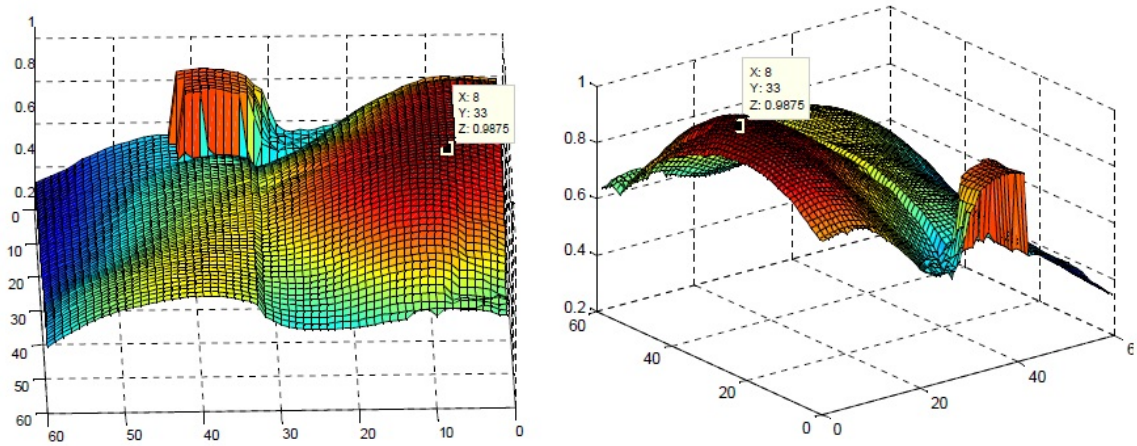

Fig. 3: The surface of similarity function and the coordinates of the maximum in the similarity of the hand in the target model and candidate in the first and second frame consecutively.

We make regulations for the similarity function by applying the mask (Kernel function) upon the hands with an isotropic-monotonic kernel in the feature space of the spatial domain. While the weights of the kernel, which carries the spatial information in continuous form, are being utilized for identifying the feature space representations, the similarity function changes into a smooth function of y.

Hand in the Target Model:

A hand as the target model is represented by an restricted region, here an ellipsoidal region used, in the frame as an image. For eliminating the impact of different target dimensions, all targets, hands, should be at first normalized to a unit circle in order to have a normalized region of feature space. This goal would be obtained by an independent rescaling of the row and column dimensions: h(x,y).Then Let us have the normalized pixel locations of the hands (or hand) in the region known as the target model at the first frame. The center of the region is right on 0 assumed in the first frame, target model. An isotropic kernel, with a convexity and monotony specifications, has a Decreasing-slope kernel profile identified as *k(x),* which gives smaller weights to the pixels of the image which are farther from the center of ellipse. Through these assigned weights, the robustness of the density estimation will be increased, since the marginal Pixels of the normalized region are the least important, influenced by occlusions, clutter or interferences of the background. The function b is assigned to the pixels at the location of x and the index of their bins in the quantized feature Space of the normalized region. The probability indicating the feature is u: 1….m in the hand region of the first frame, target Model, is then computed as

$$q_u = \sum_{i=1}^{u} \hat{k}(x)\delta(b(x)-u) \tag{3}$$

Where there is the delta function indicating the histogram normalizing. There would be also some normalization constants which come from imposing the condition that the summation of q as target probability would be summed into one, because the summation as shown above for target, hand, is one.

Hands in Target Candidates Region:

The normalized pixel locations of the subsequent frames, target candidate, have a center on y in the current next frame. This normalization procedure should be continued right as well as was applied on the frame of the target model. The same kernel function, with a definite bandwidth, the feature probability in the next frames, target candidate, is

$$p_u(y) = \sum_{i=1}^{u} \hat{k}(x)\delta(b(x)-u) \tag{4}$$

There is also constants called normalization constant. This constant is not dependent upon y the center location, the reason is that the location of the pixels are constituted in a certain lattice of pixels and then the shift of center y would be the lattice nodes as well. So that this constant can be computed and assigned by a kernel function and also different values of bandwidth in this function. This bandwidth as given before will

show the hand. Scale in different frames if calculated well enough as the target candidate. Do not forget that the pixels number should be taken into consideration during the tracking procedure.

Similarity:

As mentioned before, the similarity function (2) derives from the characteristics of the kernel function exactly when hand in the target model and candidate, first frame as reference and next ones, are shown (Aherne, F., N. Thacker 1998; Avidan, S., 2001). Also we can say that different kernel gives a different similarity function and efficient gradient-based optimizing methods can be utilized in order to find its extremum, especially here maximum. Existence of the continuous kernel yields a numerical procedure right between the different locations on the lattice of the image. The used target representations do not confine the way that the similarity between two frames is measured and various functions can be constructed (Puzicha, J., Y. Rubner, 1999) for Measures of different similarity of histogram for in an experimental evaluation.

Metric Coefficient of Similarity Function:

In order to find the similarity, we should define a distance between the different location of hand in both target model and candidate. For accommodating the comparisons of several hands as targets, this measure of distance must be based on a metric configuration. In this way, we can specify the difference and similarity between the two various discrete distributions as:

$$D(y) = (1 - \hat{\rho}(y)) \tag{5}$$

On which we consider the similarity function

$$\rho = \sum_{u=1}^{m} (p(y).q)\,\hat{}\,1/3 \tag{6}$$

Each pattern estimate of these coefficients among p and q (Kailath, T., 1967).These coefficient achieved is sort of a specific measure (Lin, J., 1991) in which there is a numerical-geometrical translation. In fact the angle is converted into the m-dimensional vectors. p and q are kinds of continuous distributions that are precisely considered through showing them on the surface. We are also able to paraphrase (6) in the form of correlation among the vectors. The coefficient like its concern with the classifying measure of data, and quality of the pattern measurement, and straightforward forms of different kinds of probability distributions are represented (Djouadi, A., O. Snorrason, 1990; Kailath, T., 1967).These estimates were also already used in field of computer vision. These bounds have also been used in (Konishi, S., A. Yuille, 1999) in order to make the determination of the effective edge detection. The Kullback-Leibler divergence among the probability distribution and the product of their peripherals were used in (Viola, P. and W. Wells, 1997).

Location of Hand as TARGET:

Locating the hand as the target in the current frame, then the distance (5) must be reduced in the form of a function y. This process begins at the location of the hand target in the previous frame as discussed before (the target model for the hand) and seeks into the adjacent pixels as well. Because the similarity function results in smooth distance, this trend goes on gradient-based data in which it is defined through an average estimation procedure through the vectors of the mean shift (Comaniciu, D. and P. Meer, 2002). It deals with the optimizations methods ,formulated according to the gradient and the Hessian (Press, W., S. Teukolsky, 1992).As feature extraction ,Color data is selected among others as the untranslatable feature of the hands as the target feature, and the same tendency may involve also texture and edges as well if necessary, other mixture of the features are possible too. Of the frame 1st, supposed that the data is available first in the localization of the hand in the initial frame for the hands to be tracked in the target models (Black, M. and D. Fleet, 2000) and then there should be sorts of representation for hand as an object to be capable of being updated and tracked in the subsequent frame as in candidate model in spite of various and diverse changes and variation in the colors as feature of the hands (McKenna, S., Y. Raja 1999).

Optimization of Metric Distance:

In this work, in order to decrease the distance, as discussed above, which is computed according to similarity function and is equal to that similarity coefficient minimizing. This procedure is just looking for the

hand as our own new target in the location of the current frame which actually is initialized by the previous position y0 in previous hand tracked as the target in the previous frame. Therefore, the probabilities of the hand in the new frame, the target candidate, at the position y0 should be computed at the beginning. Through the approximation of the equation around the parameters, we can have a new formula as well for the coefficient (6) which is achieved by some variation in their values:

$$\rho = \frac{1}{3}\sum_{u=1}^{m}(p(y).q)\,\hat{}\,1/3 + \frac{2}{3}\sum_{u=1}^{m}(p(y)(3\sqrt{\frac{q}{p}})\,\hat{}\,2 \tag{7}$$

The satisfaction of this approximation is good enough exactly at the time that the hand in the new frames, the target candidate, is not varying suddenly and largely in comparison with their previous location, because usually it is probable to be found nearby in the subsequent frames as discussed. Situation in which small limits is established for all probabilities, may not employ some of the features. In this way the weight of similarity patterns of two frames are recalculated as below:

$$w_i\sum_{u=1}^{m}(3\sqrt{\frac{q}{p}})\,\hat{}\,2\delta(b(x_i)-u) \tag{8}$$

Conclusively in order to find the minimum of the difference discussed in the similarity function (6), the similarity function formula as mentioned should be at the maximum of its level (7).But some of the parameters as seen are not based on the location of the hand as target in the new frame. Pay attention on the density as calculated according to the weights calculated above. The kernel at the new location in the present frame, along with the samples which are assigned weight according to weight (8). In this way, we have to find the optimum point, the maximum, which is the mode of this density as is illustrated in the surface and exists in the locality of the adjacent samples and can be fast searched for a local maximum, the nearest one in fact, so that would be tracked and persuited via an average calculation of the weighted sample throughout the surface, a process called mean shift (Comaniciu, D. and P. Meer, 2002). Through this trend, the weighting procedure of the samples should be update iteratively and also the kernel profile is shifted out of the current position into the new one based on the below equation:

$$y = \sum_{i=1}^{n} x_i w_i / \sum_{i=1}^{n} w_i \tag{9}$$

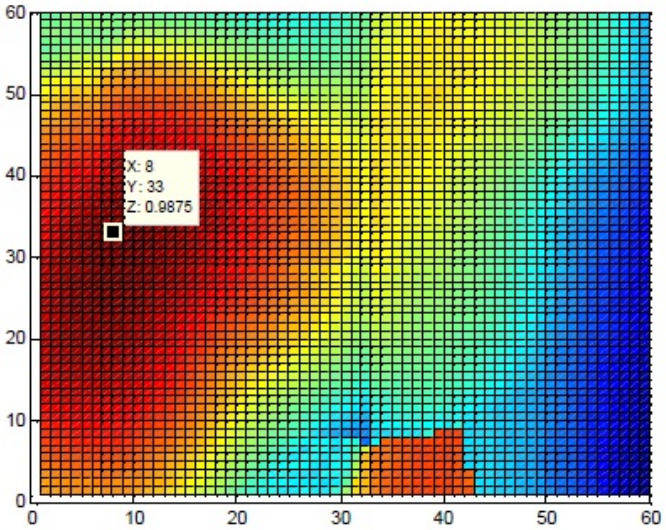

Fig. 4: The density of the similarity surface optimized and the maximum found through the density approximation

Overview of the Kernels:

Unlike other methods, in which no weight assigning methods are used, these methods are based on weights giving, called kernel, of different function and restricted area. This characteristic makes them a favorable choice to provide us with the frequency as well as temporal information for a given signal as reported in (Bar-Shalom, Y. and T. Fortmann, 1988). They can absolutely be implemented using multi-resolution techniques as reported in (Bascle, B. and R. Deriche, 1995). The advantage of such an approach is that some features which might not be detected at one resolution may be found at some other resolutions. The main contribution of the work is to introduce a new framework of efficient tracking the hands. We can show that through spatially masking the target with an isotropic function ,called kernel, and by using spatially-smooth similarity function , the target localization problem, means the hand locations, is then declined to a simple search into the basin of attraction obtained by the kernel function. In fact, we can regularize the similarity function, discussed later, by putting mask on the hands with an isotropic function, Kernel, in the spatial domain. When the kernel weights, carrying continuous spatial information, are used in constructing the feature space representations, $r\hat{}(y)$ becomes a smooth function in y. In the practical algorithm, we only make the iterations by computing the weights of the space features obtained by kernel function, deriving the new location in the next Step, and testing the size of the kernel shift would be in the next Step. Kernels with Epanechnikov profile (Comaniciu, D. and P. Meer, 2002)

$$k(x) = \{c(1 - x(x <= 1) \tag{10}$$

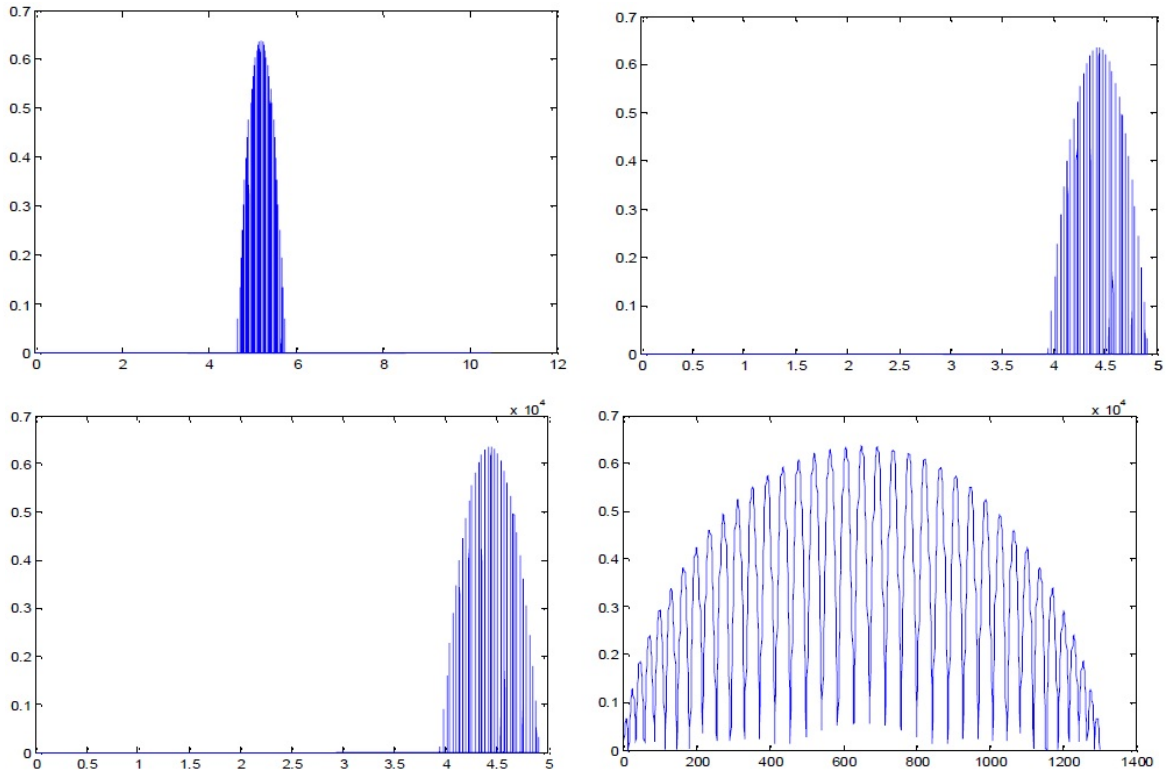

Fig. 1: The Kernel function weighting in 4 different frames 1,2,3,4

Scale Determination:

Based on the method explained previously (sec.4. 1) , in the first frame of the hand as the target model, the hand position in the current frame in which the distance as calculated through the similarity function should be minimized (5) is probably around the nearby samples of the recent position. Nevertheless, the hand scale varies usually and occasionally, so that (4) the bandwidth as mentioned previously in the kernel function should be scaled too, Because of the scale changeability due to the bandwidth parameter and covariance (5). Given by previous bandwidth in the previous frame, we are able then to estimate the optimized bandwidth in the

current frame by our method into the hand tracking by localizing it as the target candidate in constant number of times. The most typical outcome achieved, hopt, giving the largest Bhattacharyya coefficient, which is then preserved among other values. In order to prevent the sensitivity of the scale, value of the bandwidths concerning to current frame should be checked and measured based on the variance and covariance of the samples of the surface of optimization. Actually, the series of new bandwidths should be composed of important data about the moving and motion of the hands as the target in scaling, because it is better to be utilized effectively.

Experiments and Results:

This method as the kernel-density approximator hand tracker was employed over many frames of our sigh language data base movies. But it can also be generalized into object tracking as well. We were making presentation of some of the practical results in the application of sign language. The last experiment, color feature, the RGB, was considered as feature space and then it was converted into the histogram of 3*16 bins. This procedure was applied (sec 4.2). The Epanechnikov kernel was used for the histogram and the density approximator iterations was entirely dependent on the updating weighted averages as discussed above.

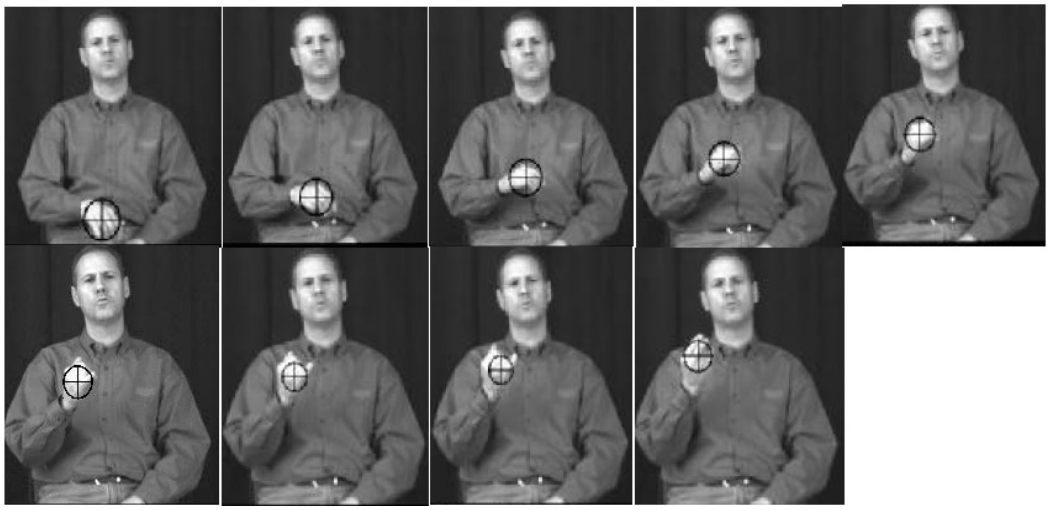

Fig. 3: An example for hand tracking in which we show 9 frames from a 50-frame long sequence. Frame 2, 4, 6, 8, 11, 13, 14, 16 and 18 respectively

Conclusions:

In this work, we presented an algorithm used specifically for hand tracking based on kernel-density approximation estimator. The proposed process uses only color RGB as feature obtained from frames image. The hand tracking is supposed to be an object tracking problem just by labeling the object in the first frame as the hand. The next of our best, approximator was based on using the weight vectors achieved in subsequent frames. Finally, approximator of the density was taken advantage for other frames to recognize hand from its, somehow, constant background. We also examined our method, to track different kind of hands as the targets in complicated backgrounds. The process was done in the first frame. The result in the other frames except the first one is what we decide to do in the future work and we also expect that this, may improve the performance of the tracking algorithm entirely. Furthermore, we can use this method as an estimator for moving model of the target, and then we can find the precise region of the hand as our desirable target using other methods as well.

REFERENCE

Aggarwal, J. and Q. Cai, 1999. “Human Motion Analysis: A Review,” Computer Vision and Image Understanding, 73: 428-440.

Aherne, F., N. Thacker and P. Rockett, 1998. “The Bhattacharyya Metric as an Absolute Similarity Measure for Frequency Coded Data,” Kybernetika, 34(4): 363-368.

Arulampalam, S., S. Maskell, N. Gordon and T. Clapp, 2002. “A Tutorial on Particle Filters for On-Line Non-Linear/Non-Gaussian Bayesian Tracking,” IEEE Trans. Signal Processing, 50(2): 174-189.

Avidan, S., 2001. "Support Vector Tracking," Proc. IEEE Conf. Computer Vision and Pattern Recognition, I: 184-191.

Bar-Shalom, Y. and T. Fortmann, 1988. Tracking and Data Association. Academic Press. 1988.

Bascle, B. and R. Deriche, 1995. "Region Tracking through Image Sequences," Proc. Fifth Int'l Conf. Computer Vision, pp: 302-307.

Birchfield, S., 1998. "Elliptical Head Tracking Using Intensity Gradients and Color Histograms," Proc. IEEE Conf. ComputerVision and Pattern Recognition, pp: 232-237.

Black, M. and D. Fleet, 2000. "Probabilistic Detection and Tracking of Motion Boundaries," Int'l J. Computer Vision, 38(3): 231-245.

Boykov, Y. and D. Huttenlocher, 2000. "Adaptive Bayesian Recognition in Tracking Rigid Objects," Proc. IEEE Conf.Computer Vision and Pattern Recognition, pp: 697-704.

Bradski, G.R., 1998. "Computer Vision Face Tracking as a Component of a Perceptual User Interface," Proc. IEEE Workshop Applications of Computer Vision, pp: 214-219.

Bue, A.D., D. Comaniciu, V. Ramesh and C. Regazzoni, 2002. "Smart Cameras with Real-Time Video Object Generation,"Proc. IEEE Int'l Conf. Image Processing, III: 429-432.

Caenen, G., V. Ferrari, A. Zalesny, L. VanGool, 2002. "Analyzing the Layout of Composite Textures," Proc. Texture 2002. Workshop, pp: 15-19.

Cham, T. and J. Rehg, 1999. "A Multiple Hypothesis Approach to Figure Tracking," Proc. IEEE Conf. Computer Vision andPattern Recognition, II: 239-219.

Chen, H. and T. Liu, 2001. "Trust-Region Methods for Real-Time tracking," Proc. Eigth Int'l Conf. Computer Vision, II: 717-722.

Chen, Y., Y. Rui and T. Huang, 2001. "JPDAF-Based HMM for Real- Time Contour Tracking," Proc. IEEE Conf. Computer Vision and Pattern Recognition, I: 543-550.

Collins, R., A. Lipton, H. Fujiyoshi and T. Kanade, 2001. "Algorithms for Cooperative Multisensor Surveillance," Proc. IEEE, 89(10): 1456-1477.

Comaniciu, D. and P. Meer, 2002. "Mean Shift: A Robust Approach Toward Feature Space Analysis," IEEE Trans. Pattern Analysis and Machine Intelligence, 24(5): 603-619.

Comaniciu, D., V. Ramesh and P. Meer, 2000. "Real-Time Tracking of Non-Rigid Objects Using Mean Shift," Proc. IEEE Conf. Computer Vision and Pattern Recognition, II: 142-149.

Cover, T. and J. Thomas, 1991. Elements of Information Theory. New York: John Wiley & Sons.

Cox, I. and S. Hingorani, 1996. "An Efficient Implementation of Reid's Multiple Hypothesis Tracking Algorithm and Its Evaluation for the Purpose of Visual Tracking," IEEE Trans. Pattern Analysis and Machine Intelligence, 18(2): 138-150.

DeCarlo, D. and D. Metaxas, 2000. "Optical Flow Constraints on Deformable Models with Applications to Face Tracking,"Int'l J. Computer Vision, 38(2): 99-127.

Djouadi, A., O. Snorrason and F. Garber, 1990. "The Quality of Training- Sample Estimates of the Bhattacharyya Coefficient," IEEE Trans. Pattern Analysis and Machine Intelligence, 12: 92-97.

Doucet, A., S. Godsill and C. Andrieu, 2000. "On Sequential Monte Carlo Sampling Methods for Bayesian Filtering," Statistics and Computing, 10(3): 197-208.

Elgammal, A., D. Harwood and L. Davis, 2000. "Non-Parametric Model for Background Subtraction," Proc. European Conf. Computer Vision, II: 751-767.

Ennesser, F. and G. Medioni, 1995. "Finding Waldo, or Focus of Attention Using Local Color Information," IEEE Trans. Pattern Analysis and Machine Intelligence, 17(8): 805-809.

Ferrari, V., T. Tuytelaars and L.V. Gool, 2001. "Real-Time Affine Region Tracking and Coplanar Grouping," Proc. IEEE Conf. Computer Vision and Pattern Recognition, II: 226-233.

Garcia, J., J. Valdivia and X. Vidal, 2001. "Information Theoretic Measure for Visual Target Distinctness," IEEE Trans. Pattern Analysis and Machine Intelligence, 23(4): 362-383.

Konishi, S., A. Yuille, J. Coughlan and S. Zhu, 1999. "Fundamental Bounds on Edge Detection: An Information Theoretic Evaluation of Different Edge Cues," Proc. IEEE Conf. Computer Vision and Pattern Recognition, pp: 573-579.

Kailath, T., 1967. "The Divergence and Bhattacharyya Distance Measures in Signal Selection," IEEE Trans. Comm.Technology, 15: 52-60.

Lin, J., 1991. "Divergence Measures Based on the Shannon Entropy," IEEE Trans. Information Theory, 37: 145-151.

McKenna, S., Y. Raja and S. Gong, 1999. "Tracking Colour Objects Using Adaptive Mixture Models," Image and Vision Computing J., 17: 223-229.

Press, W., S. Teukolsky, W. Vetterling and B. Flannery, 1992. Numerical Recipes in C, second ed. Cambridge Univ. Press.

Puzicha, J., Y. Rubner, C. Tomasi and J. Buhmann, 1999. "Empirical Evaluation of Dissimilarity Measures for Color and Texture," Proc. Seventh Int'l Conf. Computer Vision, pp: 1165-1173.

Scott, D.W., 1992. Multivariate Density Estimation. Wiley.

Viola, P. and W. Wells, 1997. "Alignment by Maximization of Mutual Information," Int'l J. Computer Vision, 24(2): 137-154.