An EM-like algorithm for color-histogram-based object tracking

Zoran Zivkovic Ben Kröse

Intelligent and Autonomous Systems Group
University of Amsterdam
The Netherlands
email:{zivkovic,krose}@science.uva.nl

Abstract

The iterative procedure called 'mean-shift' is a simple robust method for finding the position of a local mode (local maximum) of a kernel-based estimate of a density function. A new robust algorithm is given here that presents a natural extension of the 'mean-shift' procedure. The new algorithm simultaneously estimates the position of the local mode and the covariance matrix that describes the approximate shape of the local mode. We apply the new method to develop a new 5-degrees of freedom (DOF) color histogram based non-rigid object tracking algorithm.

1. Introduction

Visual data is often complex and there are usually many data points that are not well explained by the applied models. In order to deal with the outliers robust estimation techniques are very important for solving vision problems [8]. Vision problems are often very specific and the methods from robust statistics [7] need to be modified in such a way that they are made appropriate for vision problems. A robust method that is often used for solving vision problems [3, 2, 1] is the 'mean-shift' procedure [9]. Data samples are used to get a kernel based approximation of the probability density function [17]. The mean-shift algorithm is a procedure to search for a local mode of the empirical density function. The position of the local mode is known to be very tolerant to outliers.

Efficient color-histogram-based tracking presented in [3] is based on the mean-shift procedure. Color histogram is a very robust representation of the object appearance [16]. In [3] the shape of the tracked non-rigid object is represented by an ellipse. A similarity function is defined between the color histogram of the object and the color histogram of a candidate ellipsoidal region from a new image from an image sequence. The mean-shift procedure is used to find the region in the new image that is the most similar to the object. See section 5 and [3] for more details. The problem of adapting the ellipse that approximates the shape of the object when the shape and the size of the tracked object change remained unsolved. Some local shape descriptors were used in [5]. In [3] after each tracking step the ellipse is adapted by checking a +10% larger and a -10% smaller ellipse and choosing the best one. In [14] an extensive search is performed within a range of scales of the ellipse.

In this paper we present an extension of the mean-shift algorithm. Instead of only estimating the position of a local mode the new algorithm simultaneously estimates the covariance matrix that describes the shape of the local mode. This is illustrated in figure 1. Further, we show how the algorithm can be applied to color-histogram-based object tracking in a similar way as in [3]. We propose a 5-DOF color-histogram-based tracking method that estimates the position of the tracked object but also simultaneously estimates the ellipse that approximates the shape of the object. The new algorithm solves the mentioned problem of adapting the ellipse in an efficient way.

The paper is organized as follows. In section 2 we introduce mean-shift as a robust estimation technique. In section 3 we present how the mean-shift can be viewed as an EM-like algorithm. In section 4 we extend the EM-like algorithm to estimate also the local scale. In section 5 we apply the new algorithm to color histogram based tracking. Some experiments are given in section 6 and in section 7 we report some conclusions.

2. Extreme outlier model

We will denote a data set of N independent samples by $\mathcal{X} = \{\vec{x}_1, ..., \vec{x}_N\}$. Let us assume that the probability density function $p(\vec{x})$, for example a Gaussian $p(\vec{x}) = \mathcal{N}(\vec{x}; \vec{\theta}, V)$, is a good generative model for our data. Maximum Likelihood (ML) estimates for the mean vector $\vec{\theta}$ and the covariance matrix V are the values that maximize the likelihood function $\prod_{i=1}^{N} p(\vec{x}_i)$. Often in practice we are confronted with a data set which is polluted by some outliers. Uniform distribution $1/A$ (where A is the area of the domain of $\vec{x}$) can be used to model the outliers. If e presents the probabil-

1063-6919/04 $20.00 © 2004 IEEE

ity that a data sample is an outlier, we can write a common generative model, that takes into account the outliers, as:

$$p'(\vec{x}_i) = e/A + (1-e)p(\vec{x}_i). \tag{1}$$

The likelihood of the data is now $\prod_{i=1}^N p'(\vec{x}_i)$ and Taylor expansion of the likelihood in $(1-e)$ is given by:

$$(e/A)^N + (e/A)^{N-1}(1-e)\sum_{i=1}^N p(\vec{x}_i) + O((1-e)^2). \tag{2}$$

In an extreme case where there are a lot of outliers, e is close to 1 and only the first two terms matter [12]. The first term is constant and the ML estimates are obtained by maximizing $\sum_{i=1}^N p(\vec{x}_i)$. For Gaussian p, the objective function to be maximized can be written as:

$$f(\vec{\theta}, V) = \sum_{i=1}^N \mathcal{N}(\vec{x}_i; \vec{\theta}, V). \tag{3}$$

Given a fixed V and if we add $1/N$ in front, (3) resembles an empirical density estimate using Gaussian kernels, where V can be regarded as the bandwidth factor [17]. The mean-shift can be used to get a robust estimate of $\vec{\theta}$ - the mode of this empirical density function. In section 4 of this paper we show how to get also robust estimates for V using this extreme outlier model. Vision problems often involve analyzing only a local part of an image and disregarding the data from the rest of the image regardless of how large the image is. The extreme outlier model is obviously appropriate for such problems.

The robust statistics procedure called 'iteratively reweighted least squares' (IRLS) [6] is very similar to the mean-shift procedure. In fact we can see the mean-shift as a version of IRLS for the extreme outlier model. In a similar way, the new procedure we present here is a special version of the robust scale estimators [11]. We mentioned that V in (3) can be regarded as the bandwidth factor in kernel-based density estimation. However, the objective in bandwidth estimation [17] is quite different.

3. Mean-shift as an EM-like algorithm

If each data point has also a weight factor ω_i, a more general version of (3) is given by:

$$f(\vec{\theta}, V) = \sum_{i=1}^N \omega_i \mathcal{N}(\vec{x}_i; \vec{\theta}, V). \tag{4}$$

We would like to find the parameters $\vec{\theta}$ and V for which the maximum value of (4) is achieved. This can be done iteratively using EM-like iterations [4, 13]. From the Jensen's inequality we get:

$$\log f(\vec{\theta},V) \geq G(\vec{\theta},V,q_1,...,q_N) = \sum_{i=1}^N \log\left(\frac{\omega_i \mathcal{N}(\vec{x}_i;\vec{\theta},V)}{q_i}\right)^{q_i} \tag{5}$$

where q_i-s are arbitrary constants that meet the following requirements:

$$\sum_{i=1}^N q_i = 1 \text{ and } q_i \geq 0. \tag{6}$$

Let us assume that the current estimate values of the parameters are denoted by $\vec{\theta}^{(k)}$ and $V^{(k)}$. The E and M steps described below are repeated then until convergence:

1. E step: find q_i-s to maximize G while keeping $\vec{\theta}^{(k)}$ and $V^{(k)}$ fixed. It is easy to show that the maximum (equality sign in (5)) is achieved for:

$$q_i = \frac{\omega_i \mathcal{N}(\vec{x}_i; \vec{\theta}^{(k)}, V^{(k)})}{\sum_{i=1}^N \omega_i \mathcal{N}(\vec{x}_i; \vec{\theta}^{(k)}, V^{(k)})}. \tag{7}$$

2. M step: maximize G from (5) with respect to $\vec{\theta}$ and V while keeping q_i-s constant. The q_i-s are now fixed we need to minimize only a part of G that depends on the parameters:

$$g(\vec{\theta}, V) = \sum_{i=1}^N q_i \log \mathcal{N}(\vec{x}_i; \vec{\theta}, V). \tag{8}$$

From $\frac{\partial}{\partial\vec{\theta}} g(\vec{\theta}, V) = 0$ we get:

$$\vec{\theta}^{(k+1)} = \sum_{i=1}^N q_i \vec{x}_i = \frac{\sum_{i=1}^N \vec{x}_i \omega_i \mathcal{N}(\vec{x}_i; \vec{\theta}^{(k)}, V^{(k)})}{\sum_{i=1}^N \omega_i \mathcal{N}(\vec{x}_i; \vec{\theta}^{(k)}, V^{(k)})} \tag{9}$$

Note that this update equation for the position estimate is equivalent to the 'mean shift' update equation for the Gaussian kernels. For other kernel types this might be different. This new EM-like view of the problem will lead to update equations for V as described next.

4. Scale selection

If $p^*(\vec{x})$ is the true distribution of the data, the expected value of (3) is:

$$E\left[f(\vec{\theta},V)\right] = \int_{\vec{x}} p^*(\vec{x})\mathcal{N}(\vec{x};\vec{\theta},V). \tag{10}$$

This can be seen as a smoothed version of the original p^* and the maximum with respect to V does not have some desirable properties [15]. For example, if p^* is locally a Gaussian $\mathcal{N}(\vec{x};\vec{\theta}^*,V^*)$, the expected value (10) is a smoothed Gaussian $\mathcal{N}(\vec{x};\vec{\theta}^*,V^*{+}V)$. The expected maximum is for $\vec{\theta} = \vec{\theta}^*$, but unfortunately for the trivial value $V = 0$ since the value at the local mode is decreasing with larger V. We normalize the result by multiplying density estimate (4) by $|V|^{\gamma/2}$ and we get what we can call a 'γ-normalized' function:

$$f_\gamma(\vec{\theta}, V) = |V|^{\gamma/2} f(\vec{\theta}, V). \tag{11}$$

Under the same assumption that the local mode is approximately a Gaussian, the value at the mode will now be proportional to $|V|^{\gamma/2}/|V^*{+}V|^{1/2}$. The maximum with respect to V is at:

$$\frac{\partial}{\partial V} \frac{|V|^{\gamma/2}}{|V^*{+}V|^{1/2}} = 0 \tag{12}$$

Since $\frac{\partial}{\partial V}|V| = |V| \left[2V^{-1} - \text{diag}(V^{-1})\right]$ we get:

$$\gamma|V|^\gamma \left[2V^{-1} - \text{diag}(V^{-1})\right] |V^*{+}V|$$
$$-|V|^\gamma |V^*{+}V| \left[2(V^*{+}V)^{-1} - \text{diag}((V^*{+}V)^{-1})\right] = 0. \tag{13}$$

From here we get $\gamma V^{-1} = (V^*{+}V)^{-1}$ and $V = \frac{\gamma}{1-\gamma}V^*$. Obviously only for $\gamma\epsilon(0,1)$ we get a positive value. For $\gamma = 1/2$ it follows that expected maximum is for $V = V^*$. The solution using the γ-normalized function is not biased and this is a desirable property of an estimation algorithm.

The extreme outlier model in the limit case can be explained also as a model where only one observation is not an outlier [12]. Then it is understandable that V can not be estimated reliably using this model. The γ-normalization can be seen as introducing a certain informative prior for V to regularize the solution and get non-biased estimates. Another interesting connection is with some image filtering algorithms. For example, in [10] γ-normalized image convolution was studied for selecting the scale of the filtering operator. If we have a 2D case and we replace p^* in (10) with an image, the connection with the image convolution is evident.

The EM-like iterative algorithm from the previous section can be applied to the γ-normalized function. The only difference is in the M-step. Instead of (8) we have now:

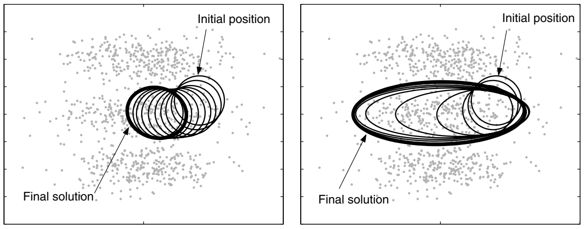
$$g(\vec{\theta}, V) = \sum_{i=1}^{N} q_i \log |V|^{\gamma/2} \mathcal{N}(\vec{x}; \vec{\theta}, V). \tag{14}$$

The position update equation (9) stays the same. From $\frac{\partial}{\partial V} g(\vec{\theta}, V) = 0$ it is easy to show that the update equation for V in the M-step is given by:

$$\vec{V}^{k+1} = \beta \sum_{i=1}^{N} q_i (\vec{x}_i - \vec{\theta}^{(k)})(\vec{x}_i - \vec{\theta}^{(k)})^T, \tag{15}$$

where $\beta = 1/(1-\gamma)$.

In figure 1 an example is shown to illustrate the performance of the new algorithm. The simulated data consists of 600 samples generated using a mixture of three Gaussian distributions. The three modes are clearly visible in figure 1. The iterations (the 2-sigma contours of the estimated Gaussian) of the mean-shift procedure are plotted in figure 1a. In figure 1b we show the iterations of the new EM-like algorithm with $\gamma = 1/2$ ($\beta = 2$). We can observe how the new algorithm simultaneously estimates both the position of the local mode and the covariance matrix that describes the shape of the mode. Note that $\beta = 2$ is appropriate if the underlying distribution is Gaussian. If some other distribution is approximated by a Gaussian some other value for β might be needed in order to avoid biased solution. Similar parameter and similar discussion is also given in the standard robust statistics methods [11, 7]. The difference is that the results we present here are for the extreme outlier model.



a) mean-shift iterations b) EM-shift iterations

Figure 1: Performance of the two algorithms on simulated 2D data.

5. Color histogram tracking

We assume that the shape of a non-rigid object is approximated by an ellipsoidal region in an image. Initially the object is selected manually or detected using some other algorithm, background subtraction for example. Let $\vec{x}_i$ denote a pixel location and $\vec{\theta}_0$ the initial location of the center of the object in the image. The second order moment can be used to approximate the shape of the object:

$$V_0 = \sum_{\text{all the pixels that belong to the object}} (\vec{x}_i - \vec{\theta}_0)(\vec{x}_i - \vec{\theta}_0)^T. \tag{16}$$

Further, the color histogram is used to model the object appearance. Let the histogram have M bins and let the function $b(\vec{x}_i) : R^2 \rightarrow 1, ..., M$ be the function that assigns a color value of the pixel at location $\vec{x}_i$ to its bin. The color histogram model of the object consists then of the M values of the M bins of the histogram $\vec{o} = [o_1, ..., o_M]^T$. The value of the m-th bin is calculated by:

$$o_m = \sum_{i=1}^{N_{V_0}} \mathcal{N}(\vec{x}_i; \vec{\theta}_0, V_0) \delta\left[b(\vec{x}_i) - m\right], \tag{17}$$

where δ is the Kronecker delta function. We use the Gaussian kernel $\mathcal{N}$ to rely more on the pixels in the middle of the object and to assign smaller weights to the less reliable pixels at the borders of the objects. We use only the N_{V_0} pixels from a finite neighborhood of the kernel and the pixels further than 2.5-sigma are disregarded.

5.1. Similarity measure

Let us assume that we have a new image from an image sequence and the object we are tracking is present in the image. The goal of a tracking algorithm is to find the object in the new image. Let an ellipsoidal region in the new image be defined by its position $\vec{\theta}$ and its shape described by the covariance matrix V. The color-histogram that describes the appearance of the region is $\vec{r}(\vec{\theta}, V)$ and the value of the m-th bin is calculated by:

$$r_m(\vec{\theta}, V) = \sum_{i=1}^{N_V} \mathcal{N}(\vec{x}_i; \vec{\theta}, V)\delta\left[b(\vec{x}_i) - m\right]. \qquad (18)$$

The similarity of the region to the object is defined by the similarity of their histograms. As in as in [3] we use Bhattacharyya coefficient as a measure of similarity between two histograms:

$$\rho\left[\vec{r}(\vec{\theta}, V), \vec{o}\right] = \sum_{m=1}^{M} \sqrt{r_m(\vec{\theta}, V)}\sqrt{o_m}. \qquad (19)$$

The first order Taylor approximation around the current estimate $\vec{r}(\vec{\theta^{(k)}}, V^{(k)})$ is given by:

$$\rho\left[\vec{r}(\vec{\theta}, V), \vec{o}\right] \approx c_1 + c_2 \sum_{i=1}^{N_V} \omega_i \mathcal{N}(\vec{x}_i; \vec{\theta}, V), \qquad (20)$$

where c_1 and c_2 are some constant factors and

$$\omega_i = \sum_{m=1}^{M} \sqrt{\frac{o_m}{r_m(\vec{\theta}^{(k)}, V^{(k)})}}\delta\left[b(\vec{x}_i) - m\right]. \qquad (21)$$

Since the last term in (20) has the same form as (4) we can use the new EM-like algorithm to search for the local maximum of the similarity function (20). The weights are recalculated before each iteration using (21) and then the update is done using (7),(9) and (15). Some practical issues are presented next.

5.2. Practical algorithm

For the sake of clarity we present here the whole algorithm:

Input: the object model $\vec{o}$, its initial ($k = 0$) location $\vec{\theta}^{(k)}$ and shape defined by $V^{(k)}$.

1. Compute the values of the color histogram of the current region defined by $\vec{\theta}^{(k)}$ and $V^{(k)}$ from the current frame using (18).

2. Calculate weights using (21).

3. Calculate q_i-s using (7).

4. Calculate new position estimate $\vec{\theta}^{(k+1)}$ using (9).

5. Calculate new variance estimate $V^{(k+1)}$ using (15).

6. If no new pixels are included using the new elliptical region defined by the new estimates $\vec{\theta}^{(k+1)}$ and $V^{(k+1)}$ stop, otherwise set $k \leftarrow k + 1$ and go to 1.

The procedure is repeated for each frame. In the simplest version the position and shape of the ellipsoidal region from the previous frame are used as the initial values for the new frame.

The maximum of (20) is well defined with respect to V for $\beta = 1$. However since we disregard the samples further than 2.5-sigma and it is easy to show that we should use $\beta \approx 1.1$. The correct value for the β depends on the noise that is present in the image sequence. Small errors in choice of β leads to slightly biased solution but since the ellipse is just an approximation of the shape this is acceptable.

Furthermore, because of the approximation that the weights ω_i are constant during one iteration the convergence proof does not hold. An additional line search should be performed to make sure that we increase the value of (19) as it was mentioned in [3]. However the approximation is usually good in the small neighborhood and this is not needed. This was also noted for the mean-shift algorithm presented in [3].

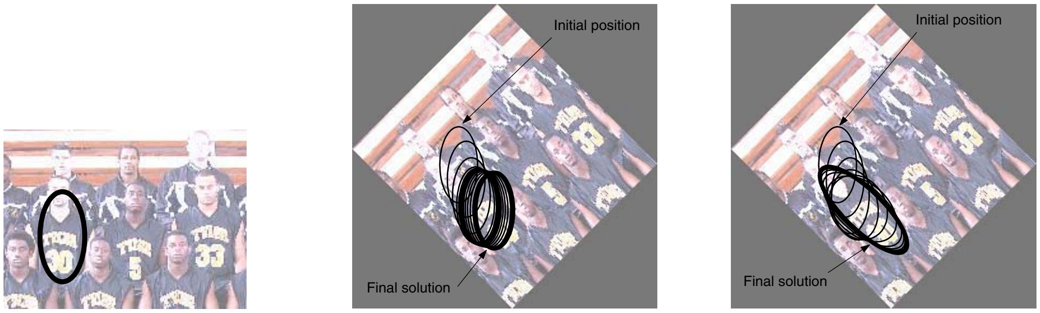
Finally, since we estimate more parameters (5 instead of only the position (2 parameters) as in [3], the algorithm is inherently less stable. This should be mentioned as a possible disadvantage of the new algorithm. However, we did not notice any problems in practice.

6. Experiments

The new 5-DOF color-histogram-based tracking was applied to a number of sequences and some results are reported in this section. The position and shape of the tracked objects is represented by the dashed ellipse.

First in figure 2 we illustrate the performance of the algorithm. A player is selected as indicated by the elliptical in figure 2a. For better presentation we increased the brightness of the images we present here. The original images was darker. The image is scaled 1.5 times in the vertical direction and then rotated for 45 degrees as presented in figure 2b. We use the initial shape of the region and we manually select a position in the new rotated and scaled image. The

1063-6919/04 $20.00 © 2004 IEEE



a) the selected region b) 'mean-shift' search c) EM-shift search

Figure 2: The mean shift and the new EM-like algorithm

iterations and the final of the mean-shift procedure are presented in figure 2b. In figure 2c we present the iterations and the final solution of our algorithm. Both the new shape and the position are accurately estimated. The new elliptical region contains the same content as the content of the initial region.

The 'hall' sequence (figure 3)is a long low-quality video from a surveillance camera. We used only H and S from HSV color space to be more robust to the light effects. The objects was represented using a 8×8 histogram in the HS space. Since the objects were walking people we did not expect the orientation of the ellipse that approximates the shape of the objects to be other than vertical. Therefore we constrained V to be diagonal. Two frames from that represent a typical situation from the video are presented in figure 3a. The object moves towards the camera and the size of the object changes considerably. Standard mean-shift tracking from [3] fails to adapt to these size changes. This is similar to the sequence that was used in [14]. Our algorithm has no problems with adapting as the much slower extensive search method from [14].

The 'PETS1' is a sequence from the standard data set from www.visualsurveillance.org. The covariance matrix is now not constrained to be diagonal since the vehicles are also changing orientation. We used RGB space and $8\times8\times8$ histogram. Two frames from the sequence are shown in 3b.

The 'hand' sequence is used to demonstrate the full 5-DOF color-histogram-based tracking. To be robust to light conditions we used again 8×8 histogram in the HS space. The hand is tracked. The sequence has 250 frames and the position and the shape of the hand are changing rapidly. In figure 3c we can see that the new algorithm can track the hand and also adapt to the shape of the object. Hand tracking was used for example in [18]. However the algorithm they used is not very robust and can be used only for single colored objects.

Finally, in figure 4 we present the number of iterations of the algorithm for the 'hand' sequence. The average number of iterations per frame was approximately 6. This is slightly more then 4 that was reported for the mean-shift based iterations in [3]. The computational complexity of one iteration of the new algorithm is slightly higher than the computational complexity of the mean-shift. On average our algorithm is around 2 times slower but still fast enough for real-time performance. In our current implementation the algorithm works comfortably in real-time on a 1GHz PC.

7. Conclusions

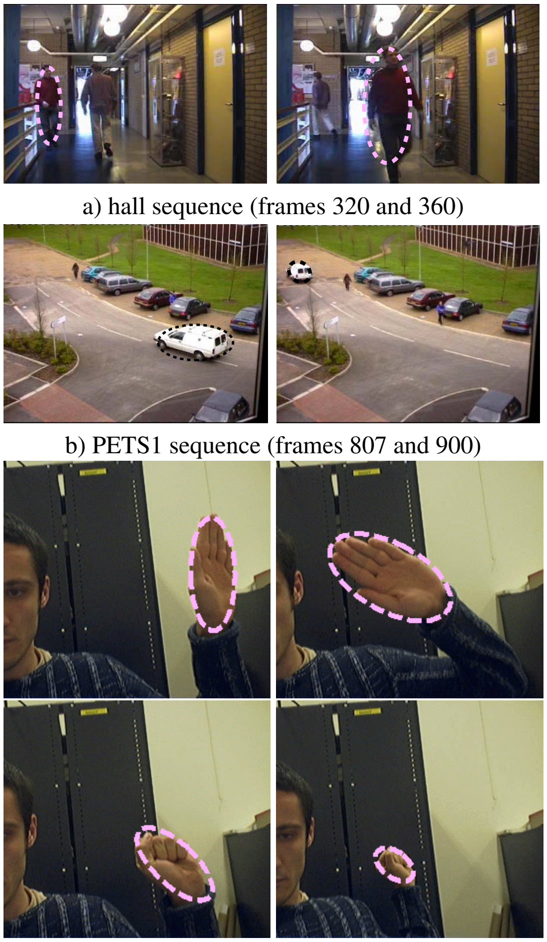
We presented a new 5-DOF color-histogram-based nonrigid object tracking. We demonstrated that he new algorithm can robustly track the objects in different situations. The algorithm can also adapt to changes in shape and scale of the object. The algorithm works in real-time and the computational cost is only slightly higher than for the previously proposed algorithms that had problems with shape and scale changes. The new color-histogram-based object tracking procedure is based on a natural extension of the mean-shift algorithm that can be useful also for many other vision problems. This is a topic of our further research.

Acknowledgments

The work described in this paper was conducted within the EU Integrated Project COGNIRON (”The Cognitive Companion”) and was funded by the European Commission Division FP6-IST Future and Emerging Technologies under Contract FP6-002020.

References

[1] H. Chen and P. Meer. Robust computer vision through kernel density estimation. *A. Heyden et al. (Eds.):ECCV 2002, LNCS2350*, pages 236–250, 2002.



a) hall sequence (frames 320 and 360)

b) PETS1 sequence (frames 807 and 900)

c) hand sequence (frames 0,100,200 and 250)

Figure 3: Some tracking results

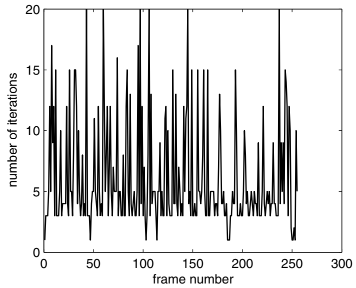


Figure 4: Number of iterations per frame for the hand sequence

[2] D.Comaniciu and P. Meer. Mean shift: A robust approach toward feature space analysis. *IEEE Trans. Pattern Anal. Machine Intell.*, 24(5), 2002.

[3] D.Comaniciu, V.Ramesh, and P.Meer. Real-time tracking of non-rigid objects using mean shift. *Proc. IEEE Conf. on Computer Vision and Pattern Recognition*, 2:142–149, 2000.

[4] A.P. Dempster, N. Laird, and D.B.Rubin. Maximum likelihood from incomplete data via the em algorithm. *Journal of the Royal Statistical Society, Series B (Methodological)*, 1(39):1–38, 1977.

[5] G.R.Bradski. Computer vision face tracking as a component of a perceptual user interface. *Proc.IEEE Workshop on Applications of Computer vision*, pages 214–219, 1998.

[6] P. W. Holland and R. E. Welsch. Robust regression using iteratively reweighted least-squares. *Commun. Statist.-Theor. Meth.*, (6):813–827, 1977.

[7] P. Huber. *Robust Statistics*. Wiley, 1981.

[8] Special issue. Robust statistical techniques in image understanding. *Computer Vision and Image Understanding*, 2000.

[9] K.Fukunaga and L.D.Hostetler. The estimation of the gradient of a density function, with applications in pattern recognition. *IEEE Trans. Information Theory*, 21:32–40, 2002.

[10] T. Lindeberg. Feature detection with automatic scale selection. *International Journal of Computer Vision*, 30(2):77–116, 1998.

[11] R. Maronna. Robust M-estimators of multivariate location and scatter. *The Annals of Statistics*, 4(1):51–67, 1976.

[12] T. Minka. The ‘summation hack’ as an outlier model. *Tutorial note*, 2003.

[13] R. M. Neal and G. E. Hinton. A new view of the EM algorithm that justifies incremental, sparse and other variants. *In, M. I. Jordan editor, Learning in Graphical Models*, pages 355–368, 1998.

[14] R.Collins. Mean-shift blob tracking through scale space. *Proc. IEEE Conf. on Computer Vision and Pattern Recognition*, 2003.

[15] B.W. Silverman. *Density Estimation for Statistics and Data Analysis*. Monographs on Statistics and Applied Probability, Chapman and Hall, 1986.

[16] M. Swain and D. Ballard. Color indexing. *Intl. J. of Computer Vision*, 7(1):11–32, 1991.

[17] M. P. Wand and M. C. Jones. *Kernel Smoothing*. Chapman and Hall, 1995.

[18] C. Wren, A. Azarbayejani, T. Darrell, and A. Pentland. Pfinder: Real-time tracking of the human body. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 19(7):780–785, 1997.