MEAN-SHIFT BACKGROUND IMAGE MODELLING

M. Piccardi and T. Jan

Computer Vision Research Group
University of Technology, Sydney, Australia
E-mail: massimo@it.uts.edu.au

ABSTRACT

Background modelling is widely used in computer vision for the detection of foreground objects in a frame sequence. The more accurate the background model, the more correct is the detection of the foreground objects. In this paper, we present an approach to background modelling based on a mean-shift procedure. The mean shift vector convergence properties enable the system to achieve reliable background modelling. In addition, histogram-based computation and the new concept of local basins of attraction allow us to meet the stringent real-time requirements of video processing.

1. INTRODUCTION AND RELATED WORKS

Background subtraction is commonly used in computer vision to detect foreground objects in videos, typically from surveillance and traffic monitoring applications. Background subtraction namely consists of an image subtraction between the current frame and an image of the scene's static background. Such a subtraction results in an image containing only the "foreground" objects, which are the objects of interest for the computer vision application.

Several techniques have been proposed in the literature for background subtraction [1], [2], [3], [4], [5], [6]. Since the static background and its appearance change over time due to both geometry and illumination changes, it is mandatory that the background image is kept updated over time. Moreover, for practical reasons, such an update must be unsupervised. Hence, most of the approaches from the literature propose estimation of the background image directly from the frame sequence; this estimate is commonly referred to as the *background model*. At each pixel location, the background model may be represented simply by one numerical value. However, in many practical cases, it needs to be described by a distribution (or a mixture of distributions, as in [2]) to reflect the changing nature of the scene's background. Therefore, the problem becomes that of estimating a probability density function (PDF) at each pixel location, quantifying the probability that a certain pixel value is background.

A natural solution to approximate such a probability density function is that of using the histogram of pixel values from the last N frames. If N is chosen as an adequately large number, the histogram is more likely to reflect background values rather than foreground, at least in its dominant modes. However, it is well known that the histogram is often not an adequate estimate of the true PDF due to its discrete nature, and that smoothing is needed for improvement. Amongst the many techniques proposed for histogram smoothing, Kernel Density Estimation (KDE) is probably one of the most popular. KDE has also been used for background modelling in important surveillance applications [3]. However, KDE application is extremely sensitive to the kernel bandwidth [7], leading in many practical cases to a poor estimate of the true PDF. A technique with lower dependency on the bandwidth parameter would be able to overcome this problem.

Recently, a novel technique based on the mean shift vector was successfully proposed for PDF estimation in applications of image segmentation and object tracking [8], [9], [10], [7]. The mean shift technique is an iterative gradient-ascent method with nice convergence properties allow it to detect the modes of a multi-variate distribution and their covariance matrix. The mean shift technique needs only know the range of the bandwidth typical of a certain application, and such a requirement can often be easily satisfied [9]. However, due to its iterative nature, the computational cost is generally high and can become prohibitive in real-time applications such as video surveillance where a PDF must be computed and updated at each pixel location. A recent paper has proposed to initialize the background model with a mean-shift procedure and use mode propagation to update the model [6]. In this way, the computational load remains reasonably limited. In this paper, we present instead a novel approach to background modelling where both initialization and update are carried out by means of a mean shift procedure. Specific optimizations are proposed so as to greatly reduce the computational cost. This makes our procedure able to meet real-time constraints, while, at the same time, the experimental results provided in this paper prove the accuracy of the approach.

2. MEAN-SHIFT VECTOR BASED MODE ESTIMATION

Given a set of data point x_i, the mean shift vector in a simple one-dimensional case can be expressed as:

$$m(x) = \frac{\sum_{i=1}^{n} x_i g((x - x_i/h)^2)}{\sum_{i=1}^{n} g((x - x_i/h)^2)} - x \quad (1)$$

where x is an arbitrary point in the data space (it can even be one of the x_i points), h is a positive value called the analysis bandwidth and $g(u)$ is a special function with bounded support; $g(u)$ is defined as the first derivative of another bounded-support function, $k(u)$, called the kernel profile [9]. Typical kernel profiles are the Epanechnicov kernel:

$$k_E(x) = \{ \begin{array}{ll} 1 - x & 0 \leq x \leq 1 \\ 0 & x > 1 \end{array} \quad (2)$$

and the truncated normal kernel:

$$k_N(x) = \{ \begin{array}{ll} \exp(-\frac{1}{2}x) & 0 \leq x \leq 1 \\ 0 & x > 1 \end{array} \quad (3)$$

When $m(x)$ is applied to the original point, x, it results in a new position, x^1; this process can be repeated and an iterative procedure defined in this way:

$$x^{l+1} = m(x^l) + x^l \quad (4)$$

For a kernel with a convex and monotonically decreasing profile, convergence of x^l for $l = 1, 2, \ldots$ can be proven. The iterative mean shift procedure is, in essence, a gradient ascent method where the step size is initially large and decreases towards convergence. This eliminates the need for a step size selection procedure and can be regarded as a major advantage with respect to traditional gradient-based methods [9].

The mean shift procedure of Eq. 4 can also be used to detect the main modes of the data set, i.e. a set of kernel functions that can be used to approximate the true distribution of the data. In order to do that, some tessellation of the data space must be chosen first and convergence studied for all points. All points belonging to a same mode will converge to a single stationary point, which is the mode center, or mean, μ. Mode detection can be regarded also as a form of data clustering. In addition, mode detection requires also the estimate of the mode variance, σ. Given all the x_i, $i = 1, \ldots, t_u$ points converging to a same u mode and assuming Gaussian kernels, this is a fitting problem allowing the least-squares solution [10]:

$$\sigma_u^2 = h^2 [\frac{\sum_{i=1}^{t_u} m(x_i)(\mu_u - x_i)}{\sum_{i=1}^{t_u} m(x_i)^2} - 1] \quad (5)$$

where μ_u and σ_u are the u mode's mean and standard deviation and h is the analysis bandwidth. It can also be proven that if the underlying distribution is normal, the mode detection is not influenced by the analysis bandwidth. However, since in many practical cases the distributions deviate from normality, h has an influence on the mode detection procedure. It has been suggested that, if the range of variation of h is known *a priori* (a requirement met in many applications), its optimal value can be chosen based on a stability procedure [10].

3. MEAN-SHIFT VECTOR BASED BACKGROUND MODELLING

The mean shift procedure can be used for modelling the background distribution at a given pixel location by using a set of recent background values, x_i, as the data set. However, the computational costs can become prohibitive, since this procedure must be repeated for all pixel locations in the frame and updated at a rate that is able to respect the background dynamic. In the following, we propose a procedure able to drastically limit such a computational load based on the following steps:

- Histogram-based mean shift computation: we compute the histogram from the background values, x_i. Then, we reformulate Eq. 1 in terms of the non-empty histogram bins and use it in the mean shift computation in place of the data samples;
- Computation of the local basins of attraction: we introduce the concept of local basins of attractions, which are ranges $[a, b]$, where the mean shift vector value is pre-computed. We then convert the explicit computation of the mean shift vector into a direct substitution procedure;
- Reconstruction of the approximated PDF from the modes: the mean shift procedure provides a mean to detect the modes of the data set. However, once the modes have been found, the contribution of each mode to the approximated PDF must be estimated. We provide a framework where the PDF is given by a selection of weighted modes.

3.1. Histogram-based mean shift computation

Since pixels from standard camera frames can assume only integer values, the histogram of a set of $x_i, i = 1, \ldots, n$ pixel values can be easily defined by using unit integer bins. Based on the histogram, the mean shift vector can be redefined as:

$$m(x) = \frac{\sum_{i=1}^{m} i y_i g((x - i/h)^2)}{\sum_{i=1}^{m} y_i g((x - i/h)^2)} - x \quad (6)$$

Example of local basin of attraction:

Hyp.: histogram with values in range [1,16]: {0, 0, 0, 0, 5, 6, 9, 4, 6, 9, 12, 16, 15, 7, 4, 0}

Analysis bandwidth $h = 3$, initial position $x^0 = 10.45$; trajectory:

$x^0 = 10.45$; $x^1 = 11.2097$; $x^2 = 11.7077$; $x^3 = 11.7077$

$\rightarrow$ 11.7077 is the stationary point for x^0, and so is for all points in range (10,11).

Fig. 1. Example of a local basin of attraction

where m is the number of histogram bins, i is the histogram bin and y_i is the i-bin's value. If $m \ll n$, the computation of Eq. 6 is significantly faster than that of Eq. 1. In our implementation, we map the non-empty histogram bins only on a list data structure and use it for the mean shift computation. Since background values tend to be local in time, m is often largely less than n and a corresponding speed up is achieved. In addition, the list form for the histogram is also highly efficient for histogram update. The data set is updated at a pre-defined rate so as to add a new sample and remove the oldest one in a first-in/first-out manner. The histogram is updated by correspondingly modifying the related bins or creating and removing bins as needed.

3.2. Computation of the local basins of attraction

Given the bounded nature of the kernel profile, the set of histogram bins contributing to $m(x)$ in Eq. 6 can be expressed as:

$$\{i|i : x - h \leq i \leq x + h, i \in \{1, \ldots, m\}\} \qquad (7)$$

Given the finite number and integer nature of the i values, $[l, u]$ intervals (open or closed) get defined where all x points share the same set of i values. For instance, for h = 2.2 all points in the open interval $(8.2, 8.6)$ share the same set {7, 8, 9, 10} of i values. Let us now add the assumption of adopting the Epanechnicov kernel: in this case, $g(\cdot)$ is the uniform kernel and, hence, all x points in $[l, u]$ turn out to have *exactly the same mean-shifted position*. Consequently, they also will all then follow exactly the same trajectory in the data space up to a same stationary point. This allows us to avoid the explicit computation of the mean shift vector for all those points $\{x|x \in [l, u]\}$ for which the relation $[l, u] \rightarrow m(x)$ has already been computed at least once. We named the $[l, u]$ intervals the *local basins of attraction* (LBAs) since all the belonging points share a same mean-shifted position (whereas by basin of attraction we just mean the domain of data points converging to a same mode). Fig. 1 shows an example of such a local basin of attraction.

LBA	x1	x2	x3	x4
(8,9)	8.8478			
$[9, 9]$	9.6613	10.0536*	11.2097*	**11.7077***
(9,10)	10.0536	11.2097*	**11.7077***	
$[10, 10]$	10.6761	11.2097*	**11.7077***	
(10,11)	11.2097	**11.7077***		
$[11, 11]$	11.4928	**11.7077***		
(11,12)	**11.7077**			

Table 1. Example of table (portion) with local basins of attraction and their trajectories (stationary points are in bold; *:iterations skipped thanks to direct substitution)

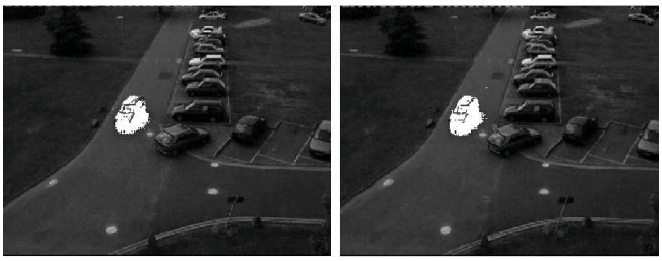
Given an assigned analysis bandwidth, k, we proceed by creating a table containing all the LBAs and their associated trajectories. We start filling the table by computing the trajectories for the local basins of attraction in positions close to the histogram's maxima. Table 1 shows an example under the same hypotheses as Fig. 1. The trajectory for interval (11,12) is computed first in this case, converging to stationary point 11.7077. When trajectory for interval (10,11) is computed, at the first iteration we obtain 11.2097 and, since this value belongs to the already computed interval (11,12), we immediately terminate its trajectory to the same stationary point without any further explicit mean shift computations. We proceed similarly for the other basins of attraction. Although the actual computational load is obviously data dependent, in this way we are often able to compute the stationary point in just 1-2 iterations. This allows us fast clustering of the original data points into their modes.

3.3. Reconstruction of the approximated PDF from the modes

Once the modes have been found with the mean shift procedure, they must be recombined into the approximated PDF, each u-mode being a Gaussian distribution of the type:

$$G_u(x) = \frac{1}{\sqrt{2\pi\sigma_u^2}} exp(-\frac{(x - \mu_u)^2}{2\sigma_u^2}) \qquad (8)$$

Recombination can be provided in different ways, for instance in terms of Sum of Modes or as Maximum Likelihood. The computational load is very limited thanks to the low number of modes (usually 1 to 3 in practical cases) and far less than that of a KDE based approach with one kernel per data point (typically in the order of 50). If multiple-channel images are used, probability independence between channels can be assumed and the overall PDF computed as the product of the individual-channel PDFs. In any case, in the reconstruction of the PDF mode weighting has to be considered. To this aim, we experimented an approach where the approximated PDF is given by the selection of the highest weighted mode:



for analysis bandwidth h=8 **for analysis bandwidth h=4**

Fig. 2. Example of foreground segmentation based on the mean-shift background model

$$PDF(x) = max_u(f_u(x)) \tag{9}$$

with each weighted mode $f_u(x)$ given by $f_u(x) = k_u G_u(\frac{x}{k_u})$ so as to provide similarity scaling both on the $x-$ and $y-$ axis. The k_u parameter was estimated as t_u/σ_u, where t_u is the number of points converging to the u-mode. Since regular tessellation is used for mode detection, this proves a good estimate of the 'size' of the mode normalized to its standard deviation, σ_u. Such an approach provided more accurate results in the experiments than a simple selection of modes. However, it adds to the overall computational costs in a real-time implementation.

Fig. 2 shows an example of foreground object detection achieved by the proposed background modelling on frames from a PETS 2001 sequence. Foreground detection at each pixel location was performed as:

$$PDF_{col}(x) = \prod_{v=r,g,b} PDF_v(x) > t_h \tag{10}$$

with $t_h = 1e^{-32}$. Fig. 2.a shows results obtained with analysis bandwidth h = 8, while Fig 2.b with h = 4. Results prove the accuracy of the background model and give evidence of the low sensitivity of the mean shift analysis to the analysis bandwidth parameter, making application easy in many practical cases.

4. CONCLUSIONS

In this paper, we have proposed a new approach for background modelling based on a mean shift vector procedure. The mean shift vector enjoys interesting properties providing us with accurate modelling of the background distribution. However, a standard implementation of the mean shift is not possible due to the excessive computational load. To this aim, we have introduced the new concept of local basins of attraction, allowing us to drastically limit the computational load deriving from mode detection. Once the modes are detected, the actual computation of the background PDF requires very limited computation, far less than that of a KDE-based approach with a kernel per data point. Finally, experimental results in this work show that the mean-shift background model is able to provide accurate foreground object detection.

5. REFERENCES

[1] C. Wren, A. Azarbayejani, T. Darrell, and A. P. Pentland, "Pfinder: real-time tracking of the human body," *IEEE Trans. on Pattern Analysis and Machine Intelligence*, vol. 19, no. 7, pp. 780–785, July 1997.

[2] C. Stauffer and W. E. L. Grimson, "Adaptive background mixture models for real-time tracking," in *Proc. IEEE CVPR*, 1999, pp. 246–252.

[3] A. Elgammal, D. Harwood, and L. S. Davis, "Non-parametric model for background subtraction," in *Proc. ECCV*, 2000, pp. 751–767.

[4] I. Haritaoglu, D. Harwood, and L. S. Davis, "W4: real-time surveillance of people and their activities," *IEEE Trans. on Pattern Analysis and Machine Intelligence*, vol. 22, no. 8, pp. 809–830, Aug. 2000.

[5] R. Cucchiara, C. Grana, M. Piccardi, and A. Prati, "Detecting moving objects, ghosts, and shadows in video streams," *IEEE Trans. on Pattern Analysis and Machine Intelligence*, vol. 25, no. 10, pp. 1337–1342, Oct. 2003.

[6] B. Han, D. Comaniciu, and L. Davis, "Sequential kernel density approximation through mode propagation: applications to background modeling," in *Proc. ACCV - Asian Conf. on Computer Vision*, Feb. 2004.

[7] P. Meer, *Robust techniques in computer vision*, Emerging Topics in Computer Vision. Prentice Hall PTR, available March 2004.

[8] Y. Cheng, "Mean shift, mode seeking, and clustering," *IEEE Trans. on Pattern Analysis and Machine Intelligence*, vol. 17, no. 8, pp. 790–799, Aug. 1995.

[9] D. Comaniciu and P. Meer, "Mean shift: a robust approach toward feature space analysis," *IEEE Trans. on Pattern Analysis and Machine Intelligence*, vol. 24, no. 5, pp. 603–619, May 2002.

[10] D. Comaniciu, "An algorithm for data-driven bandwidth selection," *IEEE Trans. on Pattern Analysis and Machine Intelligence*, vol. 25, no. 2, pp. 281–288, Feb. 2003.