2004 IEEE 6th Workshop on Multimedia Signal Processing

Inlier Modeling for Multimedia Data Analysis

Rozenn Dahyot, Niall Rea, Anil Kokaram
Dept. of Electronic and
Electrical Engineering
Trinity College
Dublin 2, Ireland
Email: dahyot@mee.tcd.ie,
oriabhan@tcd.ie, akokaram@tcd.ie

Nick Kingsbury
Dept. of Engineering
University of Cambridge
Trumpington Street
Cambridge CB2 1PZ, UK
Email: ngk@eng.cam.ac.uk

Abstract— **This paper presents a robust method to estimate the unknown standard deviation of a centred normal distribution from a mixture density. This method is applied to different signal processing problems. The first one concerns silence segmentation from audio data. The second application deals with colour class parameter extraction. In this later case, the mean is also estimated from the observations.**

I. Introduction

High-level semantic extraction methods are heavily dependent on the reliability of low-level processes that often require the designer to intituively set tuning parameters [1], [2], [3]. Many of these low-level processes can be re-expressed as a desire to identify the occurrence of particular events expressed as features. In that case the task is one of separating and extracting a desired class of features $\mathcal{C}$ (inliers) from the polluting class $\overline{\mathcal{C}}$ (outliers [4]). It is possible to make some relatively loose assumptions about the distribution of the class $\mathcal{C}$ of interest, to allow its statistics to be estimated from an observed mixture distribution and separated from the outliers [5].

The modelling of inliers and outlier data has already been studied in robust statistics [6], [5]. For instance, in [6], the M-estimators are designed to perform robust estimation of parameters over a mixture of observations where the inlier class is modelled using a centred normal distribution. The scale parameter σ [6] that controls the rejection of the data in the M-estimation is then corresponding to the standard deviation of the inliers [5]. This parameter is usually estimated using the median or MAD estimators [7]. In [5], distributions for both inlier and outlier classes are proposed in the context of an image matching application. The inliers follow a Laplacian distribution also depending on a unique parameter σ (standard deviation), whereas outlier distribution is modeled by non-parametric methods computed over the observations [5].

Following [6], we consider a class of interest (inlier) with a centred normal distribution. This paper proposes a new mechanism for estimating inlier statistics (the standard deviation σ) that is a generalisation of a method proposed for edge based segmentation [8]. It employs a non-parametric technique for identifying lobes in measured distributions and so is more robust that previous approaches. This new mechanism is applied to multimedia data analysis. Two low level tasks illustrate the efficiency of the new method: detection of silence in audio data and table or court segmentation in sport video broadcasts [1], [2], [3].

II. Principle

A. The class of Inliers

Considering two independent random variables, X_a and X_b, following the same centred normal law $\mathcal{P}_X(x) \sim \mathcal{N}(0,\sigma)$, the random variable $Y = \sqrt{X_a^2 + X_b^2}$ has a Rayleigh distribution [8], [9]:

$$\mathcal{P}_Y(y) = \frac{y}{\sigma^2} \cdot \exp\left[-\frac{y^2}{2\sigma^2}\right] \cdot \mathcal{U}(y) \quad (1)$$

One way to estimate the standard deviation σ is to compute the distribution $\mathcal{P}_Y(y)$ using observation samples $\{y_k\}$, by a histogram for instance, and compute the value $Y_{\max \mathcal{C}}$ that maximises this distribution. The parameter σ, or standard deviation of the variables X, is then easily computed by:

$$\sigma_{\mathrm{MS}} = Y_{\max \mathcal{C}} \quad (2)$$

The notation MS for the estimate σ_{MS} refers to the Mean Shift procedure that has been used to compute $Y_{\max \mathcal{C}}$ (cf. section II-D). For comparison purposes, we also compute the standard least squares estimate using the observations $\{x_k\}$:

$$\sigma_{LS} = \mathbb{E}[x^2] \quad (3)$$

And the robust Median Absolute Deviation [7]:

$$\sigma_{\mathrm{MAD}} = 1.4826 \cdot \mathrm{median}|x - \mathrm{median}(x)| \quad (4)$$

Figure 1 shows the results of a simulation for those three estimations of the standard deviation. All three are close to the true value.

B. The disturbing outliers

However in practice, the two observed random variables, X_a and X_b, follow a mixture of two laws [8]. The first one is the class of interest, noted $\mathcal{C}$, of normal distribution $\mathcal{N}(0,\sigma)$. The second one, noted $\overline{\mathcal{C}}$, gathers all the outliers of the class $\mathcal{C}$. As defined in [4], *an outlier is a data point that contains no information about the system* - the inlier class $\mathcal{C}$ - *to be estimated*. Figure 2 shows the distribution $\mathcal{P}_Y(y)$ simulated using mixed observations of the random variables X_a and X_b.

0-7803-8578-0/04/$20.00 ©2004 IEEE

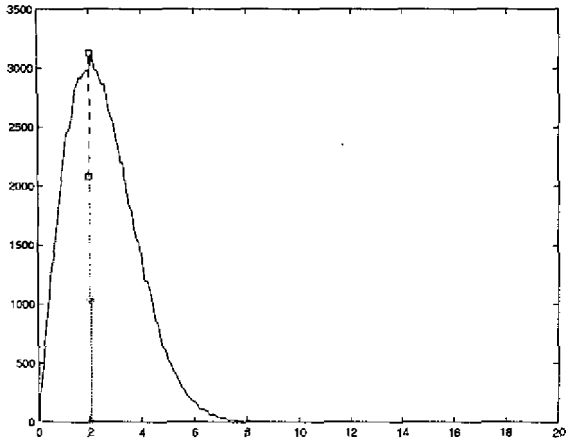


Fig. 1. Rayleigh Distribution $\mathcal{P}_Y(y)$ simulated with 100000 samples from $X_a \sim \mathcal{N}(0, \sigma = 2)$ and $X_b \sim \mathcal{N}(0, \sigma = 2)$: in *red solid* line $\sigma_{\mathrm{MS}} = 2.0413$, in *green dashdot* line $\sigma_{\mathrm{MAD}} = 1.9997$ and in *black dash* line $\sigma_{\mathrm{LS}} = 1.9996$.

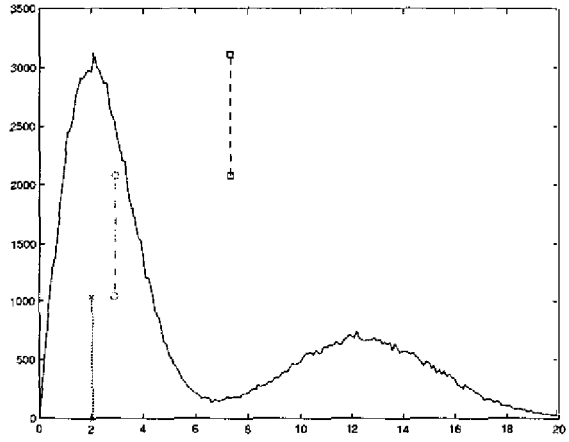


Fig. 2. Rayleigh Distribution with outliers: 100000 samples X_a, X_b have been independently generated from $\mathcal{N}(0, \sigma = 2)$ and mixed with 500000 samples from $\mathcal{N}(3, \sigma = 1)$ for X_a and from $\mathcal{N}(12, \sigma = 3)$ for X_b. Estimated standard deviation: in *red solid* line $\sigma_{\mathrm{MS}} = 2.0413$, in *green dashdot* line $\sigma_{\mathrm{MAD}} = 2.9209$ and in *black dash* line $\sigma_{\mathrm{LS}} = 7.3283$.

The estimation of the standard deviation using our method is closer to the true value than the other standard methods.

Depending on the proportion and the distribution of the outliers, the location of the relevant peak in $\mathcal{P}_Y(y)$ gets trickier. Assuming that the peak of interest for the estimation of σ is the closest to the value $y = 0$, we propose to use the mean shift procedure [10] to estimate the local maxima of interest $Y_{\max C}$ from $\mathcal{P}_Y(y)$.

C. Generalization

More generally, for a random variable Y computed from independent random variables $\forall i,\ X_i \sim \mathcal{N}(0, \sigma)$ such as:

$$Y = \sqrt{\sum_{i=1,\cdots,n} X_i^2}$$

has the χ function with n degrees of freedom as a probability density function [9]:

$$\mathcal{P}_Y(y) = \frac{2^{1-\frac{n}{2}}}{\Gamma\left(\frac{n}{2}\right)} \cdot \frac{y^{n-1}}{\sigma^n} \cdot \exp\left[-\frac{y^2}{2\sigma^2}\right] \cdot \mathcal{U}(y) \qquad (5)$$

The maximum of $\mathcal{P}_Y(y)$ is then linked to the unknown parameter σ such as:

$$\begin{aligned} Y_{\max C} &= \arg\max_y \mathcal{P}_Y(y) \\ &= \sqrt{n-1}\,\sigma \end{aligned} \qquad (6)$$

Once the maximum $Y_{\max C}$ is located, this relation provides an estimate of the unknown parameter σ.

D. Finding the maximum $Y_{\max C}$ using Mean Shift

The mean shift is a nonparametric estimator of the density gradient. By computing its zeros, the maxima of the distribution can then be located [10]. We collect a set of independent observations $\{y_k\}$ of the random variable Y. Considering the Epanechnikov kernel, the closest mode to the value $y = 0$ is computed using the following mean shift procedure:

$$\begin{array}{l} \text{Init } y = 0 \text{ (or } y = \min_k\{y_k\}) \\ \quad M_h(y) = \frac{1}{n_y} \sum_{y_k \in [y-h;y+h]} y_k - y \\ \quad y \leftarrow y + M_h(y) \\ \text{till convergence } Y_{\max C} = y \end{array} \qquad (7)$$

where n_y is the number of observation samples y_k is the interval $[y-h; y+h]$. The bandwidth parameter h, that controls the resolution of the mode selection, has been manually set in our applications in section III [10].

III. Applications

Application of this robust parameter estimation for edge segmentation in images has already been proposed in [8]. We consider here two other segmentation tasks. In section III-A, a silence detection method in audio streams is proposed. In section III-B, an automatic colour region segmentation is presented for sport video indexing purposes.

A. Silence detection in audio data.

In sport broadcast, the audio stream is composed of different source of sounds (audience, referees, etc.). One of particular importance, is the silence that appears in between those classes. We propose to use our method to segment the silence class C in audio data from the non-silence one $\overline{C}$. We assume that the hypotheses regarding the class C are met (cf. section II-A).

1) From stereo data: Considering a stereo audio signal $(s_r(k), s_l(k))$ (k, index the audio samples), we assume that the two data streams, s_r and s_l, are independent and provide the observations of our random variables X_a and X_b. The samples $\{y_k\}$ are computed by $y_k = \sqrt{s_r^2(k) + s_l^2(k)}$. Figure 3 shows the distribution $\mathcal{P}_Y(y)$ drawn from those observations. As the perception of silence is only possible on longer duration than only one audio sample (at a frequency $fs = 44100Hz$, a sample lasts for $0.03ms$), we propose another way to use our method in the next section.

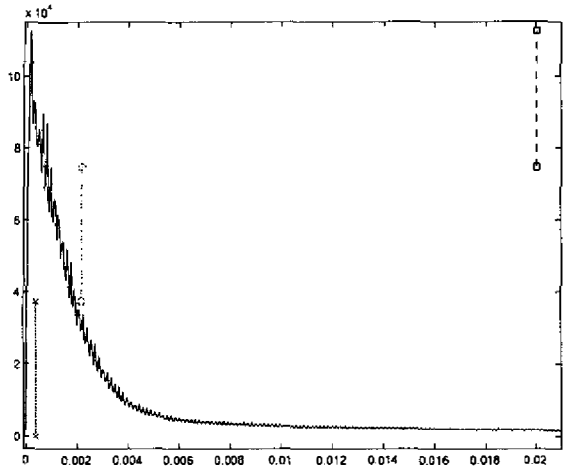


Fig. 3. Distribution $\mathcal{P}_Y(y)$ computed using stereo audio samples from a snooker broadcast. Estimated standard deviation: $\sigma_{\mathrm{MS}} = 7.7544e-004$, $\sigma_{\mathrm{MAD}} = 0.0022$ and $\sigma_{\mathrm{LS}} = 0.002$.

2) From mono stream audio data: When only one audio stream $s(k)$ is available, the observations y_k can be computed using successive samples as follows:

$$y_k = \sqrt{\sum_{i \in \Delta_k} s^2(i)} \tag{8}$$

The size of the temporal window Δ_k centred on the sample k defines the number n of data considered to compute samples of the random variable Y. It is also corresponding to the degree of freedom in the distribution of the inlier class (cf. equation 5).

3) Thresholding: One application of our method is the segmentation of the data set in between the two classes $\mathcal{C}$ and $\overline{\mathcal{C}}$. One simple way is to classify data such as:

$$\begin{cases} x_k \in \mathcal{C} \text{ if } |x_k| < 3\sigma \\ x_k \in \overline{\mathcal{C}} \text{ otherwise} \end{cases}$$

or

$$\begin{cases} y_k \in \mathcal{C} \text{ if } y_k < 3\sqrt{n-1}\,\sigma \\ y_k \in \overline{\mathcal{C}} \text{ otherwise} \end{cases}$$

The value 3σ insures that 99.7% of the class $\mathcal{C}$ are selected. Figure 4 shows an example of segmented audio signal of a tennis video computed using a temporal window $\Delta = 40ms$ [2].

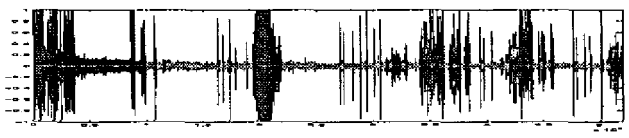
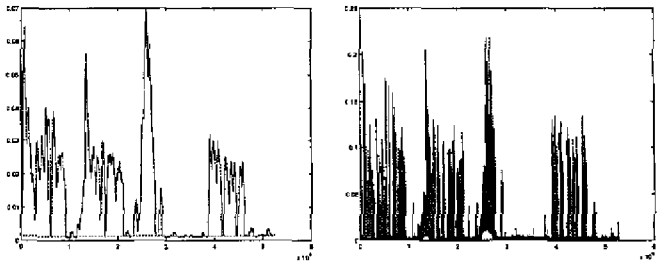


Fig. 4. Silence Detection in *Pierce* audio data [2]: in *green* the class $\mathcal{C}$ and in *blue* $\overline{\mathcal{C}}$.

For audio data, the variable Y, as defined in equation 8, correspond to the loudness or energy of the signal. It is a basic feature used for indexing sport videos [11]. Figure 5 shows the loudness information of a snooker broadcast computed at two different temporal resolutions Δ. The relative error of the estimation σ_{MS} at multiresolution (using equation 6) has been less than 0.01 in those experiments (Δ changing from $0.001s$ to $1s$). The accuracy of the estimation is then not sensitive to the choice of the temporal window Δ in computing Y.



$\Delta = 1s$ $\Delta = 0.001s$

Fig. 5. Loudness information Y (in *blue*) with the corresponding $Y_{\max \mathcal{C}}$ (in *green*) computed at different temporal resolution.

B. Colour Region segmentation in images

The application considered in this section concerns the court and snooker table detection in tennis and snooker videos [1], [2], [3]. The class $\mathcal{C}$ of interest is the homogeneus colour of those objects. Depending of the camera view (for indoor snooker videos) or varying lighting condition (in outdoor tennis videos), the statistics of this class of interest are temporally changing. We therefore propose to estimate those parameters for each image of the sequence instead of manually setting them [12]. From RGB images, we compute the variables:

$$\begin{cases} r = \frac{R}{R+G+B} \\ g = \frac{G}{R+G+B} \\ I = \frac{R+G+B}{3\times 255} \end{cases} \quad \text{and then} \quad \begin{cases} x_a = r - \mu_C^r \\ x_b = g - \mu_C^g \\ x_c = I - \mu_C^I \end{cases}$$

Following [1], the means μ_C^r, μ_C^g and μ_C^I are estimated by considering the maximum peak in the colour distribution $\mathcal{P}(r, g, I)$. This is performed in a coarse to fine way: first the maximum peak is located in the 3D colour histogram and then starting from those first estimates of the means, a Meanshift procedure is performed to refine the values μ_C^r, μ_C^g and μ_C^I.

The random variable Y is computed either $Y = \sqrt{X_a^2 + X_b^2}$ ($\mathcal{P}_Y(y|C)$ is a Rayleigh distribution) or $Y = \sqrt{X_a^2 + X_b^2 + X_c^2}$ ($\mathcal{P}_Y(y|C)$ is a Maxwell distribution). Figure 6 shows the distribution $\mathcal{P}_Y(y)$ computed using the visual data from an image of a snooker table (cf. figure 8 (d)).

The robustness of the method is assessed by performing the estimation of σ on each images of a snooker video shot as illustrated in figure 7. The stability in both the estimations of the means and the standard deviation insures the success of the segmentation of the regions of interest in the videos. Figure 8 shows the resulting segmentation of snooker table and a court in sport video images.

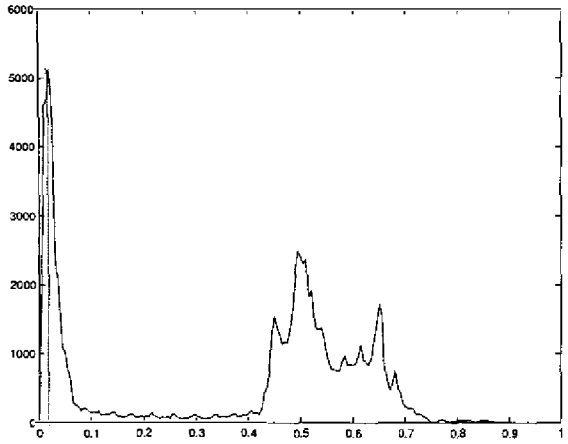


Fig. 6. Maxwell distribution computed with visual data (snooker broadcast image cf. figure 8 (d)). Estimated standard deviation $\sigma_{\mathrm{MS}} = 0.0174$.

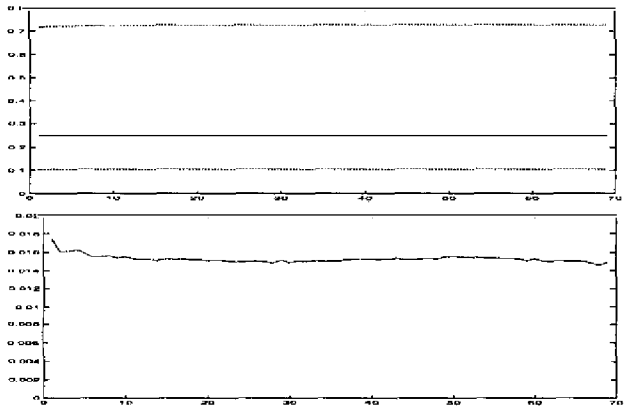
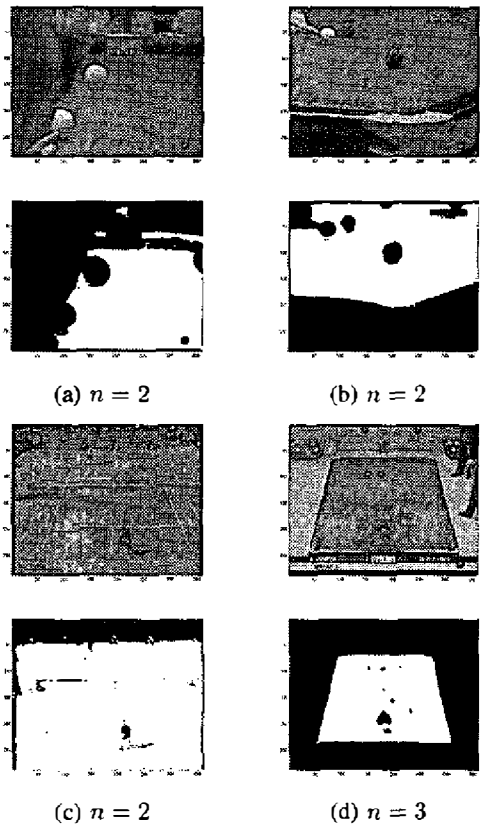


Fig. 7. Top: Means μ_C^r, μ_C^g and μ_C^I estimated over the shot. Bottom: the corresponding estimated standard deviation σ_{MS}. The variance of the σ_{MS} is less than 0.000002 over the sequence.



(a) $n = 2$ (b) $n = 2$ (c) $n = 2$ (d) $n = 3$

Fig. 8. Examples of unsupervised segmentation of court and snooker table (threshold at $3\sigma_{\mathrm{MS}}$).

IV. Conclusion

We proposed a method to robustly estimate the standard deviation of a class of data driven by a centred normal distribution from a observed mixture. This method has been successfully applied over two types of data, audio and visual, for segmentation purposes.

Acknowledgment

This work has been funded by the EU RTN MOUMIR (HP-99-108: www.moumir.org) and the EU NOE MUSCLE (FP6-5077-52).

References

[1] A. Ekin, A. M. Tekalp, and R. Mehrotra, "Automatic soccer video analysis and summarization," *IEEE Transaction on Image Processing*, vol. 12, no. 7, pp. 796–807, July 2003.

[2] R. Dahyot, A. C. Kokaram, N. Rea, and H. Denman, "Joint audio visual retrieval for tennis broadcasts," in *IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, Hong Kong, April 2003.

[3] N. Rea, R. Dahyot, and A. Kokaram, "Modeling high level structure in sports with motion driven hmms," in *IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP 2004)*, Montreal, Quebec, Canada, May 2004.

[4] T. P. McGarty, "Bayesian outlier rejection and state estimation," *IEEE transaction on Automatic Control*, pp. 682–687, October 1975.

[5] D. Hasler, L. Sbaiz, S. Süsstrunk, and M. Vetterli, "Outlier modeling in image matching," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 25, no. 3, March 2003.

[6] P. Huber, *Robust Statistics*. John Wiley and Sons, 1981.

[7] F. R. Hampel, E. M. Ronchetti, P. J. Rousseeuw, and W. A. Stahel, *Robust Statistics : The Approach Based on Influence Functions*. John Wiley and Sons, 1986.

[8] P. L. Rosin, "Edges: saliency measures and automatic thresholding," *Machine Vision and Applications*, vol. 9, pp. 139–159, 1997.

[9] A. Papoulis and S. U. Pillai, *Probability, Random Variables and Stochastic Processes*, fourth edition ed. Mc Graw Hill, 2002.

[10] D. Comaniciu and P. Meer, "Mean shift: A robust approach toward feature space analysis," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 24, no. 5, May 2002.

[11] R. Leonardi, P. Migliorati, and M. Prandini, "Semantic indexing of soccer audio-visual sequences: A multimodal approach based on controlled markov chains," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 14, no. 5, May 2004.

[12] H. Denman, N. Rea, and A. Kokaram, "Content-based analysis for video from snooker broadcasts," *in Special Issue on Video Retrieval and Summarization, Journal of Computer Vision and Image Understanding*, vol. 92, pp. 141–306, 2003.