Unsupervised Camera Motion Estimation and Moving Object Detection in Videos

Rozenn Dahyot
School of Computer Science and Statistics
Trinity College Dublin, Ireland
https://www.cs.tcd.ie/Rozenn.Dahyot/
Rozenn.Dahyot@cs.tcd.ie

Abstract

In this article, we consider the robust estimation of a location parameter using M-estimators. We propose here to couple this estimation with the robust scale estimate proposed in [Dahyot and Wilson, 2006]. The resulting procedure is then completely unsupervised. It is applied to camera motion estimation and moving object detection in videos. Experimental results on different video materials show the adaptability and the accuracy of this new robust approach.

Keywords: M-estimation, camera motion, moving object detection, robust estimation, video analysis

1 Introduction

Many problems in computer vision involve the separation of a set of data into two classes, one of interest in the context of the application and the remaining one. For instance, edge detection in images requires the thresholding of the gradient magnitude to discard noisy flat areas from the edges. The challenge is then to automatically select the appropriate threshold [Rosin, 1997].

Regression problems also involve the simultaneous estimation of the variance or standard deviation of the residuals/errors. The presence of a large number of outliers makes difficult the estimation of the parameters of interest. Performance of robust estimators is highly dependent on the setting of a threshold or scale parameter, to separate the good data (inliers) that fit the model, from the gross errors (outliers) [Chen and Meer, 2003]. The scale parameter, needed in M-estimation and linked to the scale parameter of the inliers residuals, is often set a priori or estimated by the Median Absolute Deviation. In [Dahyot and Wilson, 2006], a robust non-parametric estimation for the scale parameter has been proposed and then combined with a robust RANSAC [Fischler and Bolles, 1981] for object recognition. This paper proposes to combine the robust scale parameter estimation with a M-estimation of the camera motion parameter in videos. The whole scheme is unsupervised. The estimated scale parameter is also used to detect moving objects in the sequences.

2 Robust scale estimation

2.1 Observations

The observations consist in a set of independent samples $\{x_i\}$ of a random variable X. Its probability density function can be written as a mixture:

$$\mathcal{P}_X(x|\sigma,\theta) = \mathcal{P}_X(x|\sigma,\theta,\mathcal{C}) \cdot \mathcal{P}_X(\mathcal{C}) + \mathcal{P}_X(x|\overline{\mathcal{C}}) \cdot \mathcal{P}_X(x|\overline{\mathcal{C}}) \quad (1)$$

with the *pdf* $\mathcal{P}_X(x|\sigma,\theta,\mathcal{C})$ corresponding to a particular class $\mathcal{C}$ of interest (inliers) that depends onto one scale parameter σ and possibly also on a location parameter θ. The other *pdf* $\mathcal{P}_X(x|\overline{\mathcal{C}})$ in the mixture is generated by possible outliers occurring in the observations (class $\overline{\mathcal{C}}$) and the parameters of interest σ and θ do not depend on those outlying observations. $\mathcal{P}_X(\mathcal{C})$ is the proportion of inliers and $\mathcal{P}_X(\overline{\mathcal{C}})$ is the proportion of outliers.

In this work, we assume the distribution of the inliers to be a Generalized centred Gaussian [Aiazzi et al., 1999]:

$$\begin{array}{l} \mathcal{P}_X(x|\mathcal{C},\sigma,\theta) = \frac{1}{2\Gamma(\alpha)\cdot\alpha\cdot\beta^\alpha}\exp\left[\frac{|x(\theta)|^{1/\alpha}}{\beta}\right] \\ \text{with } \beta = \sigma^{1/\alpha}\cdot\left[\frac{\Gamma(\alpha)}{\Gamma(3\alpha)}\right]^{1/(2\alpha)} \end{array} \tag{2}$$

Setting the shape parameter $\alpha = 1$ (Laplacian law) and $\alpha = 1/2$ (Gaussian law) in equation (2), are two popular hypotheses [Hasler et al., 2003, Dahyot et al., 2004]. We assume α is known and focus on the estimation of the scale σ and θ.

2.2 Robust scale estimation knowing the location parameter θ

We now assume that we have $\mathbf{n}$ independent variable X_n of the same nature of the X previously defined. Samples for each X_n can be easily obtained for instance by splitting the original sample set of samples $\{x_i\}$ into $\mathbf{n}$ sets. Depending on the applications, the $\mathbf{n}$ random variables can also be naturally defined (see section 3). We define the variables:

$$\left\{\begin{array}{ll} Z & = \sum_{n=1}^{\mathbf{n}} |X_n|^{1/\alpha} \\ Y & = Z^\alpha \end{array}\right. \tag{3}$$

Inliers of Z and Y (in class $\mathcal{C}$) are the samples z_j or y_j computed with $\mathbf{n}$ independent samples of X such that $\forall i, x_i \in \mathcal{C}$. For $\mathbf{n} = 1$ and $Z = |X|^{1/\alpha}$, the *pdf* $\mathcal{P}_Z(z|\theta,\sigma,\mathcal{C})$ corresponds to the gamma distribution:

$$\mathcal{P}_Z(z|\theta,\sigma,\mathcal{C}) = \mathcal{G}_{Z|(\alpha,\beta)}(z) = \frac{z^{\alpha-1}}{\Gamma(\alpha)\cdot\beta^\alpha}\exp\left[-\frac{z}{\beta}\right], \quad z \geq 0 \tag{4}$$

When $\mathbf{n} > 1$, the *pdf* $\mathcal{P}_Z(z|\mathcal{C},\sigma)$ is the gamma function $\mathcal{G}_{Z|(\mathbf{n}\alpha,\beta)}(z)$, and the *pdf* of Y can easily be inferred [Dahyot and Wilson, 2006]. The maximum of the distributions $\mathcal{P}_Z(z|\mathcal{C},\sigma)$ and $\mathcal{P}_Y(y|\mathcal{C},\sigma)$ can be then computed:

$$\left\{\begin{array}{l} Z_{\max\mathcal{C}} = \beta\cdot(\mathbf{n}\alpha - 1),\ \mathbf{n}\alpha > 1 \\ Y_{\max\mathcal{C}} = [(\mathbf{n}-1)\ \alpha\ \beta]^\alpha,\ \mathbf{n} > 1 \end{array}\right. \tag{5}$$

Those maxima depend on the parameter σ by definition of β (cf. eq. (2)). From equation (5), the scale σ can be computed by:

$$\left\{\begin{array}{ll} \sigma_Z = \left(\frac{Z_{\max\mathcal{C}}}{\mathbf{n}\alpha-1}\right)^\alpha\left[\frac{\Gamma(3\alpha)}{\Gamma(\alpha)}\right]^{1/2}, & \mathbf{n}\alpha > 1 \\ \sigma_Y = \frac{Y_{\max\mathcal{C}}}{(\mathbf{n}-1)^\alpha\cdot\alpha^\alpha}\left[\frac{\Gamma(3\alpha)}{\Gamma(\alpha)}\right]^{1/2}, & \mathbf{n} > 1 \end{array}\right. \tag{6}$$

The maximum of the distributions of Y and Z has first to be located. Depending on the proportion and the values of the outliers, the localisation of the maximum needed in the estimation gets more difficult. We assume that the relevant maximum for the estimation is the closest peak to zero in the distributions $\mathcal{P}_Y(y|\sigma,\theta)$ and $\mathcal{P}_Z(z|\sigma,\theta)$.

2.3 Computation of the scale estimate in practice

The scale estimates are computed using the meanshift procedure on the set of samples of variables Y or Z, starting from the minimum sample value (or starting from zero). More details are presented in [Dahyot and Wilson, 2006]. However, in some signal processing applications, the digitised signal is discrete with known quantized levels in a finite domain. For instance, pixel values in video data are integers in $[0; 255]$. Most of all, the variable Z (or Y) has its values in a one-dimensional space. Therefore, as an alternative to the kernel representation of the distribution and the Mean Shift algorithm to perform the estimation, standard histograms can be easily used and their derivatives easily computed using filters. This is another practical and faster way to perform the estimation when dealing with a digitised signal. Both Y and Z perform similarly a robust scale estimation (see [Dahyot and Wilson, 2006]).

3 Applications

Section 3.1 presents an experiment for unsupervised moving object detection in static sequences. The variable Z is used for the scale estimation, there is no location parameter θ (the camera motion is null) and the shape parameter is chosen $\alpha = 1$. Section 3.1 extends those preliminary results to camera motion estimation and moving object detection. The variable Y is used for the scale estimation, the location parameter θ corresponds to a 6-dimensional camera motion vector and the shape parameter is chosen $\alpha = 1/2$.

3.1 Application to moving object detection in static camera sequences

3.1.1 Using colour video data $\mathbf{n} = 3$

We are considering two different colour images $I_t = (R_t, G_t, B_t)$ and $I_{t'} = (R_{t'}, G_{t'}, B_{t'})$ from a video sequence. The samples of the random variables X_n for $n \in \{1, 2, 3\}$ are computed as the inter-frame difference on each colour band for each position i of the pixel:

$$\left\{ \begin{array}{l} X_1 : \; x_i^{(1)} = R_{t'}(i) - R_t(i) \\ X_2 : \; x_i^{(2)} = G_{t'}(i) - G_t(i) \\ X_3 : \; x_i^{(3)} = B_{t'}(i) - B_t(i) \\ Z : \; z_i = |x_i^{(1)}| + |x_i^{(2)}| + |x_i^{(3)}| \end{array} \right. \tag{7}$$

The distribution of Z has been drawn using a histogram over the samples $\{z_i\}$ in figure 1(a). The estimated distribution of the inliers $\mathcal{P}_Z(z|\mathcal{C}, \sigma_Z)$ is also superimposed (with a rescaling factor to match the maxima).

3.1.2 Using grey level video data $\mathbf{n} = 2$

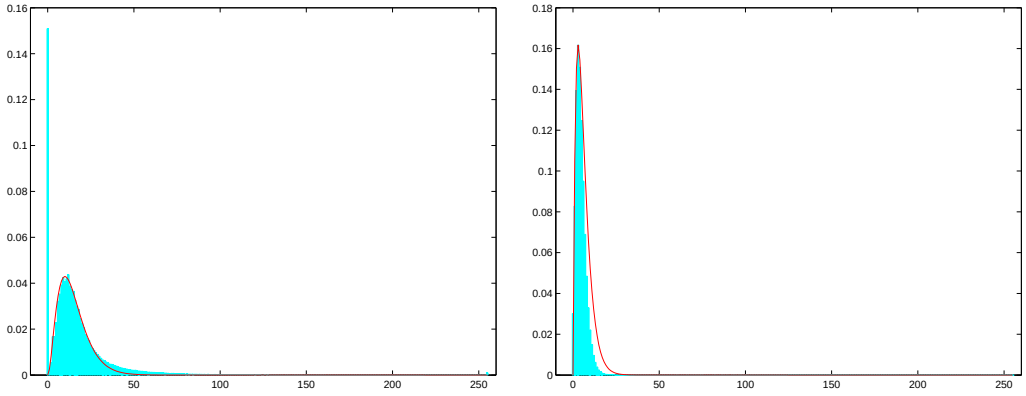
When the sequence is grey level, the samples of the variable Z can be computed using the backward and forward inter-frame differences:

$$\left\{ \begin{array}{l} X_1 : \; x_i^{(1)} = I_{t+1}(i) - I_t(i) \\ X_2 : \; x_i^{(2)} = I_t(i) - I_{t-1}(i) \\ Z : \; z_i = |x_i^{(1)}| + |x_i^{(2)}| \end{array} \right. \tag{8}$$

Figure 1 (b) shows the distribution of Z in this case.

3.1.3 Results

When comparing images from a video, the interframe differences contain outliers due to camera and object motion. However in most applications, it can be assumed that a sensible proportion of pixels are matching. This proposed scheme for standard deviation estimation can be



(a) $\mathbf{n} = 3$ (b) $\mathbf{n} = 2$

Figure 1: Distribution $\mathcal{P}_Z(z)$ (blue bars) with the fitted distribution of the inliers $\mathcal{P}_Z(z|\mathcal{C}, \sigma_Z)$ on real video data for inter-frame difference analysis.

used on the inter-frame differences in order to separate or locate the outliers in the observations. We are considering in this example a video recorded from a static camera. Only moving objects in the scene generate outliers. By setting a threshold using the estimated standard deviation, the moving objects are located using the decision rule $z_i > \mathbf{n}\ 3\ \sigma$ for each pixel position i.

Figure 2 shows an example of the moving object detection process. The inter-frame difference is computed between the median frame of the video as a model of the background model of the scene, and the frame at time $t = 100$. Segmentation of the movement is not perfect as it is only based on the pixel statistics. Some pixels from the moving objects have similar values as the background ones at the same location, therefore their differences are classified as inliers (un-moving regions). However, it is a simple method to roughly locate moving regions which, in this example, are objects of interest such as pedestrians and cars for a traffic surveillance application. A result on a grey level sequence is shown in figure 3.

3.2 Application to unsupervised camera motion estimation and moving object detection in moving camera sequences

We consider in this section the problem of camera motion estimation from video data. Camera motion estimation has many applications such as video restoration and content analysis [Bouthémy et al., 1999, Kokaram et al., 2003, Coldefy et al., 2004]. The motion parameters are 6-dimensional to take into account zoom, rotation and translation. The residuals $\{x_i\}$, corresponding to the displaced frame difference (*dfd*), are linearly depending on the camera motion parameters θ. Figure 5 shows two successive images in a sequence and their difference, respectively before and after camera motion compensation. Those two observations correspond to the residuals observed at the first and the last iteration of our robust estimation. Images of videos used for testing are shown in fig. 4.

3.2.1 Iterative estimation of the scale and location parameters

For each iteration, we estimate the scale parameter σ_Y on the set of residuals and then perform the estimation of the motion parameters until convergence:

$$
\begin{array}{l}
\text{Initialisation } \theta^{(0)} \\
\text{Repeat :} \\
\left|\begin{array}{l}
\sigma_\rho^{(m)} \leftarrow \text{Scale estimation on } \{y_i\} \\
\qquad\quad \text{computed from } \{x_i(\theta^{(m)})\} \\
\\
\theta^{(m+1)} = \arg\min \left\{ \sum_i \rho\left(\frac{x_i(\theta^{(m)})}{\sigma_\rho^{(m)}} \right) \right\}
\end{array}\right. \\
\text{Until convergence } \hat{\theta}\ \hat{\sigma}_\rho
\end{array}
\tag{9}
$$

Frame at $t = 100$

Median image of the sequence as *background*

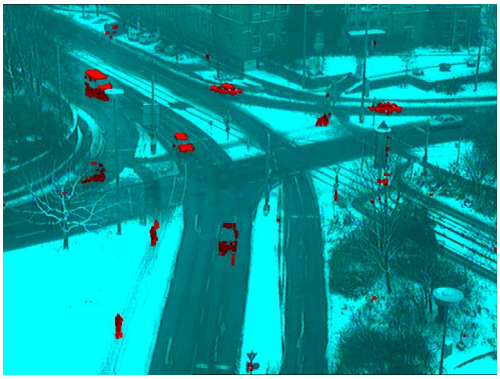
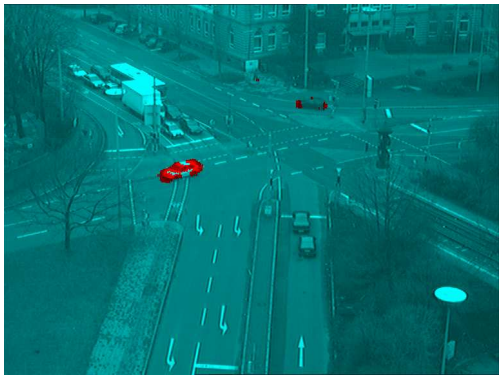


Figure 2: Colour sequence. Detection of moving objects (in red) based on the statistics of the difference of the colour pixels with the median image of the sequence (sequence *dtneu_winter*).

$t = 199$

$t = 200$

$t = 201$



Moving objects at $t = 200$

Figure 3: Grey-level sequence. Detection of moving objects in between successive frames, based on the statistics of the difference of grey level pixels(sequence *dt_passat03*).

(a) (b)

Figure 4: Videos used in the test. (a) is a simple video where the camera motion is known, and the dark rectangle is the only (outlier) object that moves differently from the background. (b) is a shot from an old film where several artifacts occur (blotch, flicker, sporadic and severe vertical displacement, etc).

The location parameter estimation is performed using M-estimation with a robust function ρ [Dahyot and Kokaram, 2004]. The initial guess $\theta^{(0)}$ is estimated using non-robust least squares.

Scale parameter estimation. Using the set of residuals $\{x_i\}$ samples of the variable X, we need first to define variable Y and to compute its samples. The residuals are a mixture of inliers and outliers as being defined in this article. The proportion α of inliers is unknown and the outliers correspond to unmatched pixels due, for instance, to moving objects (different movement to the camera motion), occlusions or artifacts (specially in old films). The outliers form localized areas in the *dfd* (cf. figure 5). Using this property, we draw samples of the random variable Y such as $y_i = \sqrt{x_i^2 + x_{i+1}^2}$ where x_i and x_{i+1} are two neighbouring residuals in the *dfd*. This strategy allows to preserve a similar proportion of inliers in the observations $\{y_i\}$. Using our estimation scheme with histograms for a faster computation, the scale parameter is set to $\sigma_\rho = 3 \times \sigma_Y$. This choice insures that 99% of the inliers are kept for the estimation of the camera motion parameters.

Accuracy of the estimates. Using video (a) (cf. figure 4) where the ground truth is known, the scale parameter and the motion parameters are estimated while adding gaussian noise of variances 10 and 100. Compared to the ground truth, the motion parameters are estimated with a Mean Square Error below 0.00007 on zoom-rotation parameters and 0.05 pixels on the translation parameters. The estimated scale parameter of the class of inliers is also stable over the sequence: the standard error of the estimate σ_Y of one grey level compared to the ground truth.

Scale Adaptability. On the video (b) (cf. fig. 4), our unsupervised robust camera motion estimation is also performed. No ground truth is known, however the estimated parameters are coherent with the one with a manually tuned estimation. The estimated scale parameter of the inliers in the *dfd* remains constant over 400 frames of the sequence (b). 10 frames show a slight over-estimation of the scale. Those artifacts correspond to sudden changes in the intensity values (flicker) that increase the *dfd* [Kokaram et al., 2003]. The automatic SD estimation allows to account for those changes in the data stream. The algorithm proposed in (9) has also always converged in our experiments undertaken on different videos.

3.2.2 Moving object detection

Figure 5 shows an example with two successive images from a video of cricket. The images of the residuals is also shown before and after motion compensation. The estimated scale σ_Y allows to take a decision in between pixels being outlier residuals (in black) and pixels being inlier residuals (in white). Those binary maps allows the detection of independent moving

objects in the sequence, and carry relevant information, for instance, for video understanding [Coldefy et al., 2004]. Our method provides an automatic thresholding method that does not require to manually set a threshold over the weights as in [Coldefy et al., 2004].

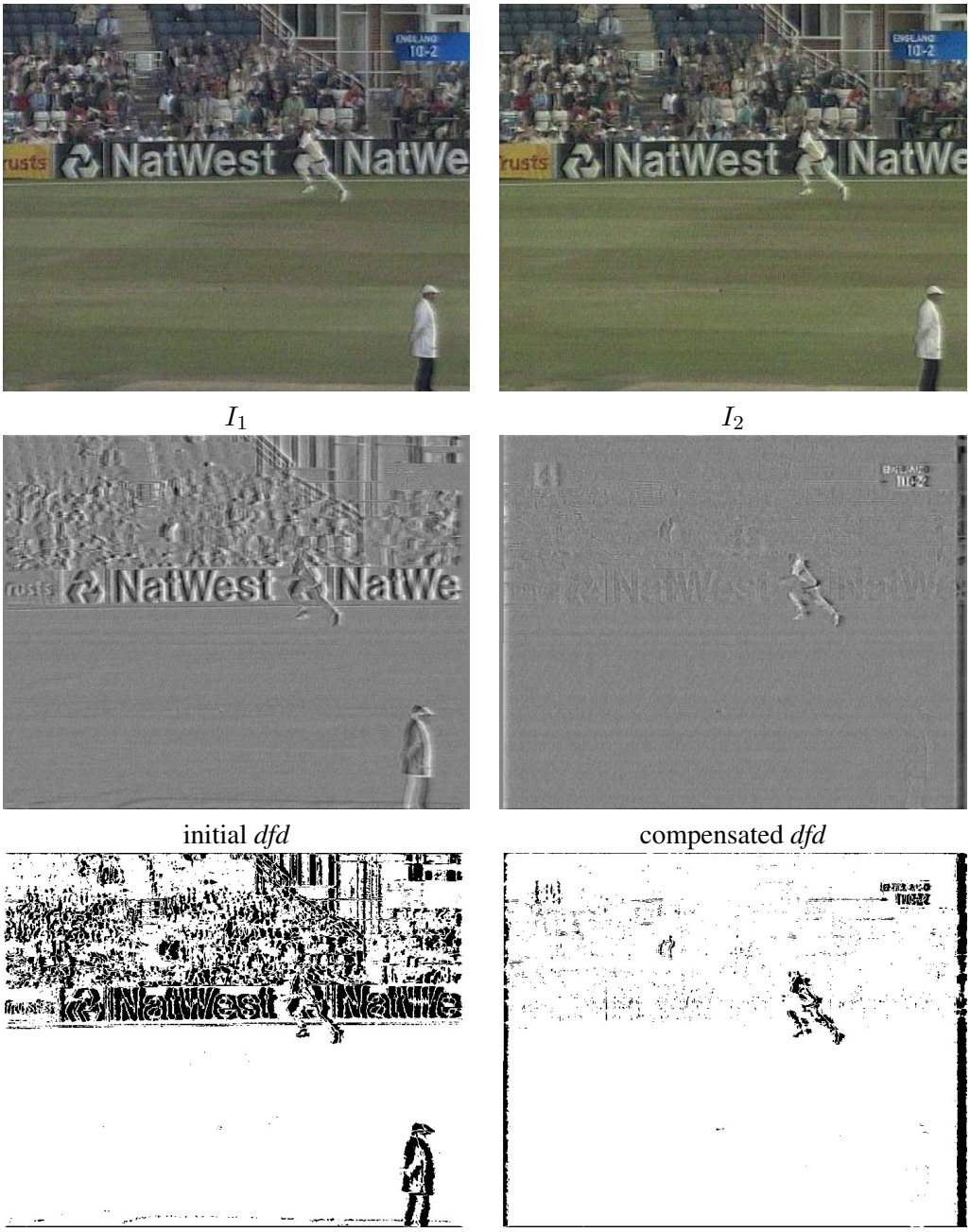


Figure 5: Application to camera motion estimation. Top: two successive images from a sport video. Middle: corresponding image of residuals when the motion is not compensated (left), and when it is compensated (right). Bottom: maps of the outliers (residuals above $3 \times \sigma^Y$).

4 Future work

In this article, we proposed an unsupervised method for camera motion estimation and moving object detection in video. The whole objects are not detected but only its parts that generate outliers in the *dfd*. Future work will aim at improving the segmentation of the moving objects by using mathematical morphology and/or hysteresis thresholding to take into account spatial correlation between neighbouring pixels.

Acknowledgments

This work has been funded by the European Network Of Excellence on *Multimedia Understanding through Semantics, Computation and Learning*, MUSCLE FP6-5077-52 (www.muscle-noe.org). We thank the institute KOGS/IAKS of Universität Karlsruhe for the video sequences (http://i21www.ira.uka.de/image_sequences/).

References

[Aiazzi et al., 1999] Aiazzi, B., Alparone, L., and Baronti, S. (1999). Estimation based on entropy matching for generalized gaussian pdf modeling. *IEEE Signal Processing Letters*, 6(6):138–140.

[Bouthémy et al., 1999] Bouthémy, P., Gelgon, M., and Ganansia, F. (1999). A unified approach to shot change detection and camera motion characterization. *IEEE Transactions on Circuits and Systems for Video Technology*, 9:1030–1044.

[Chen and Meer, 2003] Chen, H. and Meer, P. (2003). Robust regression with projection based m-estimators. In *International Conference on Computer Vision*, pages 878–885, Nice, France.

[Coldefy et al., 2004] Coldefy, F., Bouthemy, P., Betser, M., and Gravier, G. (2004). Tennis video abstraction from audio and visual cues. In *Proceedings IEEE Workshop on Multimedia Signal Processing, MMSP'2004*, pages 163–166, Siena, Italy.

[Dahyot and Kokaram, 2004] Dahyot, R. and Kokaram, A. (2004). Comparison of two algorithms for robust m-estimation of global motion parameters. In *proceedings of Irish Machine Vision and Image Processing Conference*, pages 224–231, Dublin, Ireland.

[Dahyot et al., 2004] Dahyot, R., Rea, N., Kokaram, A., and Kingsbury, N. (2004). Inlier modeling for multimedia data analysis. In *IEEE International Workshop on MultiMedia Signal Processing*, pages 482–485, Siena Italy.

[Dahyot and Wilson, 2006] Dahyot, R. and Wilson, S. (2006). Robust scale estimation for the generalized gaussian probability density function. *Advances in Methodology and Statistics (Metodološki zvezki)*, 3.

[Fischler and Bolles, 1981] Fischler, M. A. and Bolles, R. C. (1981). Random sample consensus: a paradigm for model fitting with applications to image analysis and automated cartography. *Commun. ACM*, 24(6):381–395.

[Hasler et al., 2003] Hasler, D., Sbaiz, L., Süsstrunk, S., and Vetterli, M. (2003). Outlier modeling in image matching. *IEEE Transactions on Pattern Analysis Machine Intelligence*, 25(3):301–315.

[Kokaram et al., 2003] Kokaram, A. C., Dahyot, R., Pitie, F., and Denman, H. (2003). Simultaneous luminance and position stabilization for film and video. In *Visual Communications and Image Processing*, pages 688–699, San Jose, California USA.

[Rosin, 1997] Rosin, P. L. (1997). Edges: saliency measures and automatic thresholding. *Machine Vision and Applications*, 9:139–159.