Background Subtraction with Dirichlet Process Mixture Models

Supplementary Material

Tom S. F. Haines and Tao Xiang

1 Colour Model

A simple method for filtering out shadows is to separate the luminance and chromaticity, and then ignore the luminance, as demonstrated by Elgammal et al. [1]. This tends to ignore too much information; instead a novel step is taken too reduce the importance of luminance. Firstly luminance is moved to a separate channel. Due to the DE assuming independence between components this is advantageous, as luminance variation tends to be higher than chromatic variation. To do this a parametrised colour model is designed. First the r, g, b colour space is rotated so luminance is on its own axis

$$\begin{pmatrix} l \\ m \\ n \end{pmatrix} = \begin{pmatrix} \sqrt{3} & \sqrt{3} & \sqrt{3} \\ 0 & \sqrt{2} & -\sqrt{2} \\ -2\sqrt{6} & \sqrt{6} & \sqrt{6} \end{pmatrix} \begin{pmatrix} r \\ g \\ b \end{pmatrix}, \tag{1}$$

then chromaticity is extracted

$$l' = 0.7176\, l, \qquad \begin{pmatrix} m' \\ n' \end{pmatrix} = \frac{0.7176}{\max(l, f)} \begin{pmatrix} m \\ n \end{pmatrix}, \tag{2}$$

where 0.7176 is the constant required to maintain a unit colour space volume[1] - details of its derivation are below. To obtain chromaticity the division should be by l rather than $\max(l, f)$, but this results in a divide by zero. Assuming the existence of noise when measuring r,g and b the division by l means the variance of m' and n' is proportional to $\frac{1}{l^2}$. Consequentially we have two competing goals - to estimate chromaticity and to limit the variance of this estimate. The use of $\max(l, f)$ introduces f, a threshold on luminance below which capping variance takes priority over chromaticity estimation - we fix this at 0.01. This colour space is then parametrised by r, which scales the luminance to reduce its importance against chromaticity

$$[l, m, n]_r = [r^{\frac{2}{3}} l', r^{-\frac{1}{3}} m', r^{-\frac{1}{3}} n']. \tag{3}$$

The volume of the colour space has again been held at 1. Robustness to shadows is obtained by setting r to a low value, as this reduces the importance of brightness changes. For most of the experiments presented r has been set to 0.75.

1. The post processor assumes a uniform distribution over colour, and hence needs to know the volume.

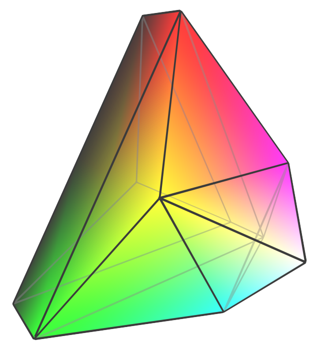

Fig. 1. Visualisation of the colour model - see text for details.

Normalisation: The colour model has its colour space visualised in Figure 1. A uniform prior is needed over the space, motivating a scaling of the space to have unit volume - this is the source of the constant 0.7176 in Equation 2. The sides of this surface are slightly curved; additionally the noise floor parameter for the colour model adjusts the volume. However, as neither of these adjust the volume by much it can be approximately modelled as a triangular prism from which three wedges are subtracted, and to which a triangular pyramid is added. The triangular prism has a depth of $\frac{2\sqrt{3}}{3}$ and its triangular cross section has a base of $\sqrt{6}$ and a height of $\frac{3\sqrt{2}}{2}$, giving it a volume of

$$\frac{2\sqrt{3}}{3} \frac{1}{2} \sqrt{6} \frac{3\sqrt{2}}{2} = \frac{\sqrt{36}}{2} \tag{4}$$

Each triangular wedge is a quarter the size of a cuboid (Application of halving for a triangle, twice for two dimensions), with edges of length $\frac{\sqrt{6}}{2}$, $\sqrt{\frac{45}{8}}$ and $\frac{\sqrt{3}}{3}$, giving a volume of

$$\frac{\sqrt{6}}{2} \sqrt{\frac{45}{8}} \frac{\sqrt{3}}{3} \frac{1}{4} = \frac{\sqrt{\frac{405}{4}}}{24} \tag{5}$$

Finally, the pyramid has an equilateral triangular base, where each side is of length $\frac{\sqrt{6}}{2}$, and a height of $\frac{\sqrt{3}}{3}$, so its volume is

$$\frac{3}{12}\frac{\sqrt{3}}{3}\left(\frac{\sqrt{6}}{2}\right)^2 \cot\left(\frac{\pi}{3}\right) = \frac{1}{8} \tag{6}$$

Combining these the final volume is given by

$$\frac{\sqrt{36}}{2} - \frac{3\sqrt{\frac{405}{4}}}{24} + \frac{1}{8} \approx 1.867 \tag{7}$$

To normalise we multiply by the inverse, which is given by ≈ 0.7176.

2 Approximation

A certain amount of optimisation is required to speed up the proposed model update algorithm. Specifically, to obtain real time performance an approximation is made when evaluating the student-t distribution

$$\mathcal{T}(x|v,\mu,\sigma^2) = \frac{\Gamma(\frac{v+1}{2})}{\Gamma(\frac{v}{2})\sqrt{v\pi}}\left(1+\frac{(x-\mu)^2}{v\sigma^2}\right)^{-\frac{v+1}{2}} \tag{8}$$

to avoid evaluating the Gamma functions, as they are the most computationally demanding part. The normalising constant can be expressed in terms of a Beta function

$$\frac{\Gamma(\frac{v+1}{2})}{\Gamma(\frac{v}{2})\sqrt{v\pi}} = \frac{1}{\beta(\frac{1}{2},\frac{v}{2})\sqrt{v}} \tag{9}$$

to which Stirling's approximation for one parameter of Beta fixed and the other large,

$$\beta(x,y) \sim \Gamma(x)y^{-x}, \tag{10}$$

can be applied. This gives

$$\frac{1}{\beta(\frac{1}{2},\frac{v}{2})\sqrt{v}} \sim \frac{1}{\Gamma(0.5)v^{-0.5}2^{0.5}\sqrt{v}} \tag{11}$$

which can be simplified

$$\frac{1}{\Gamma(0.5)v^{-0.5}2^{0.5}\sqrt{v}} = \frac{1}{\sqrt{2\pi}} \sim 0.399\ldots \tag{12}$$

noting that the dependence on v has gone, creating a constant. The error for using this approximation peaks at $v=1$, with the approximation 1.1 times its correct value - it quickly drops either side. It does not appear to have a negative effect when solving the current problem and avoids over 80% of the otherwise required computation.

3 Experiments: Wallflower

The *wallflower* [17] data set tests one frame only for each problem, by counting the number of mistakes made[2]. Testing on a single frame is hardly ideal, and the resolution (160×120) is very low; its value is that many algorithms have been run on it. There are seven tests, given in Figure 2 for a qualitative evaluation:

2. For the purpose of comparison the error metrics used by previous papers [17] have been used.

Barnich [2]	KDE with a spherical kernel. Uses a stochastic history, that removes information randomly so that old but useful samples can stay around longer.
Collins [3]	Hybrid frame differencing / background model.
Culibrk [4]	Neural network variant of Gaussian KDE.
Evangelio [5]	Runs two copies of Stauffer [6] at different learning rates and moves mixture components between them to short-cut responding to transitions between foreground and background.
Hofmann [7]	History based; equivalent to a KDE based approach with square kernels. Learns both a learning rate and acceptance threshold dynamically for each pixel.
Kim [8]	'Codebook' based; almost KDE with a cuboid kernel.
Li 1 [9]	Histogram based, includes co-occurrence statistics. Lots of heuristics.
Li 2 [10]	Refinement of the above.
Maddalena 1 [11]	Uses a self-organising map, passes information between adjacent pixels.
Maddalena 2 [12], [13]	As above but adds spatial coherence by biasing the bg/fg threshold in favour of neighbourhood consistency.
Morde [14]	Takes a very basic background subtraction approach and adds in lots of modules - includes motion detection, recurrent motion image [15], shadow detection, object tracking and object classification.
Schick [16]	A postprocessor only - takes a preexisting background subtraction algorithm and regularises it using a Markov random field over k-means generated superpixels.
Stauffer [6]	Classic GMM approach. Assigns mixture components to bg/fg.
Toyama [17]	History based, with region growing. Has explicit light switch detection.
Wren [18]	Incremental spatio-colourmetric clustering (tracking) with change detection.
Zivkovic [19]	Refinement of Stauffer [6]. Has an adaptive learning rate.
Seidel [20]	Models the background as a low dimensional subspace using PCA with a smoothed l_0 norm.

TABLE 1
Brief summaries of the key competitors.

- *moved object*, where a chair is moved part way through the sequence, being sat on for a short while.
- *time of day*, where the sun moving in the sky is simulated by fading lights indoors.
- *light switch*, where a light switch is toggled in a room.
- *waving trees*, where an outdoor shot of trees waving in the wind is used.
- *camouflage*, where a person wanders in front of a flickering CRT monitor, to test if the flicker can camouflage the person.
- *bootstrap*, where there is no training period.
- *foreground aperture*, where a sleeping person wakes and transitions from background to foreground, without entirely moving from his starting spot.

Quantitative results are given in Table 2. Previously published results have been tuned for each problem, so

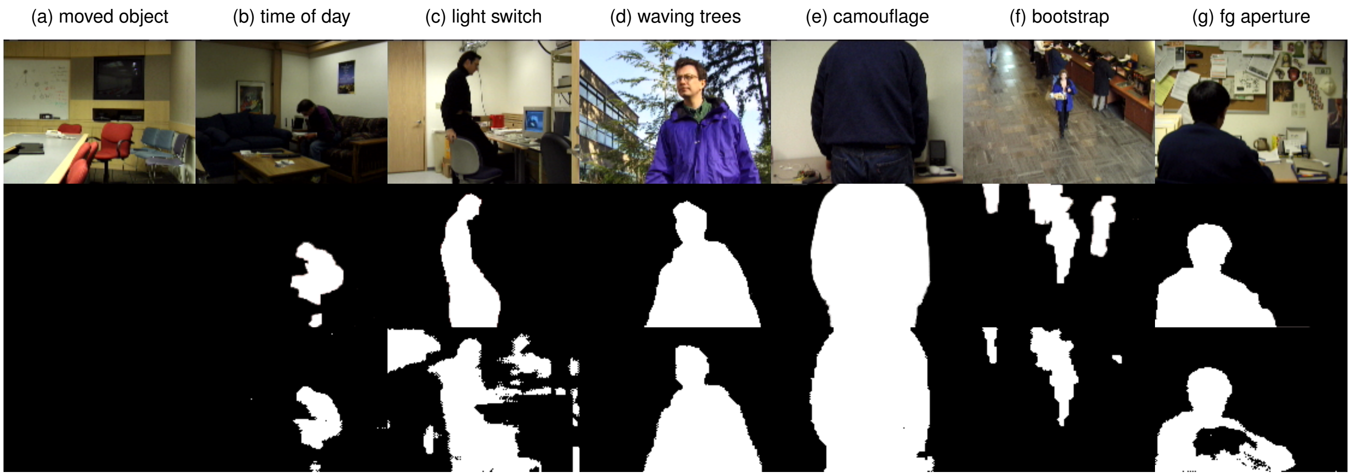

Fig. 2. *Wallflower* results: On the top row is the image, on the second row the ground truth and on the third row the output of the presented algorithm. Toyama et al. [17] provide the outputs for other algorithms.

method	moved object	time of day	light switch	waving trees	camouflage	bootstrap	foreground aperture	*mean*
Temporal derivative [17], [21]	1563 (12)	12993 (16)	16083 (10)	2742 (10)	5231 (9)	2645 (7)	4910 (12)	6595 (12)
Mean + covariance [17], [18]	**0** (1)	1484 (13)	16980 (12)	3467 (14)	6141 (10)	2307 (5)	4754 (11)	5019 (11)
Bayesian decision [17], [22]	**0** (1)	1580 (14)	15819 (9)	963 (7)	3668 (8)	4907 (10)	4485 (10)	4489 (10)
Mean + threshold [17]	**0** (1)	2593 (15)	16232 (11)	3285 (13)	1832 (3)	3236 (9)	2818 (5)	4285 (9)
Frame difference [17]	**0** (1)	1358 (12)	2565 (4)	6789 (16)	10070 (12)	2175 (4)	4354 (9)	3902 (8)
Mixture of Gaussians [17], [23]	**0** (1)	1028 (10)	15802 (8)	1664 (8)	3496 (6)	2091 (3)	2972 (6)	3865 (7)
Linear prediction [17]	**0** (1)	986 (8)	15161 (7)	1864 (9)	3558 (7)	2390 (6)	3068 (8)	3861 (6)
Block correlation [17], [24]	1200 (11)	1165 (11)	3802 (5)	3771 (15)	6670 (11)	2673 (8)	2402 (4)	3098 (5)
Eigen-background [17], [25]	1065 (10)	895 (7)	1324 (2)	3084 (12)	1898 (4)	6433 (11)	2978 (7)	2525 (4)
Toyama [17]	**0** (1)	986 (8)	**1322** (1)	2876 (11)	2935 (5)	2390 (6)	**969** (1)	1640 (3)
Collins [3]		≈653 (5)		≈430 (6)				
Wren [18]		≈654 (6)		≈298 (4)				
Kim [8]		≈492 (4)		≈353 (5)				
Maddalena [11]		≈453 (3)		≈293 (3)				
DP-GMM	**0** (1)	439 (2)	4442 (6)	234 (2)	1291 (2)	1823 (2)	1946 (3)	1454 (2)
DP-GMM, tuned	**0** (1)	**301 (1)**	2502 (3)	**178** (1)	**384** (1)	**1236** (1)	1534 (2)	**1033** (1)

TABLE 2

Wallflower [17] results: Given as the number of pixels that have been assigned the wrong class. On average the presented approach makes 37% less mistakes than its nearest competitor. The four algorithms with only two results have been inferred from the numbers given by Maddalena & Petrosino [11], and are subject to some inaccuracy. The approaches above the first separator are from Toyama et al. [17] - most are based on other approaches however, which have also been cited.

we do the same in the *DP, tuned* row, but results using a single set of parameters are also shown, in the *DP* row, to demonstrate its high degree of robustness to parameter selection. For 5 of the 7 tests we take 1^{st}, albeit shared for *moved object*, and its overall mean error puts it in first, with 37% less errors than its nearest competitor.

On foreground aperture it takes 2^{nd}, beaten by the Toyama [17] algorithm. This shot consists of a sleeping person waking up, at which point they are expected to transition from background to foreground. He is wearing black and do not entirely move from his resting spot, so the algorithm continues to think they are background in that area. The regularisation helps to shrink this spot, but the area remains. It performs relatively poorly on the light switch test, which is interesting as no issue occurs with the synthetic equivalent. For the presented approach lighting correction consists of estimating a single multiplicative constant - this works outdoors where it is a reasonable model of the sun, but indoors where light bounces around and has a highly non-linear effect on the scene it fails. It is therefore not surprising that the synthetic approach, which simulates a sun, works, whilst the indoor approach, which includes light coming through a door and the glow from a computer monitor, forces it to relearn most of the model from scratch. Examining the output in Figure 2 it can be noted that it is in the process of relearning - the test frame is the 13^{th} after the light is switched on.

4 Parameter Sweeps

Figure 3 presents parameter sweeps for the key parameters, in terms of the f-measure for the *change detection* data set. All of them show a similar pattern - a peak, with a slow decline as the parameter is increased but a sharp decline as the parameter heads towards zero. They demonstrate how the approach is robust to the choice of parameters, which allows it to obtain consistantly high performance across the extensive set of videos it has been tested on.

REFERENCES

[1] A. Elgammal, D. Harwood, and L. Davis, "Non-parametric model for background subtraction," *Frame-rate Workshop*, pp. 751–767, 2000.

[2] O. Barnich and M. V. Droogenbroeck, "Vibe: A powerful random technique to estimate the background in video sequences," *Acoustics, Speech and Signal Processing*, pp. 945 – 948, 2009.

[3] R. T. Collins, A. J. Lipton, T. Kanade, H. Fujiyoshi, D. Duggins, Y. Tsin, D. Tolliver, N. Enomoto, O. Hasegawa, P. Burt1, and L. Wixson, "A system for video surveillance and monitoring," CMU, Tech. Rep., 2000.

[4] D. Culibrk, O. Marques, D. Socek, H. Kalva, and B. Furht, "Neural network approach to background modeling for video object segmentation," *Neural Networks*, vol. 18(6), pp. 1614–1627, 2007.

[5] R. H. Evangelio and T. Sikora, "Complementary background models for the detection of static and moving objects in crowded environments," *Adv. Video and Signal-Based Surveillance*, pp. 71–76, 2011.

[6] C. Stauffer and W. E. L. Grimson, "Adaptive background mixture models for real-time tracking," *CVPR*, vol. 2, pp. 637–663, 1999.

[7] M. Hofmann, P. Tiefenbacher, and G. Rigoll, "Background segmentation with feedback: The pixel-based adaptive segmenter," *Workshop on Change Detection, CVPR*, pp. 38–43, 2012.

[8] K. Kim, T. H. Chalidabhongse, D. Harwood, and L. Davis, "Background modeling and subtraction by codebook construction," *ICIP*, vol. 5, pp. 3061–3064, 2004.

[9] L. Li, W. Huang, I. Y. H. Gu, and Q. Tian, "Foreground object detection from videos containing complex background," *Proc. Multimedia*, pp. 2–10, 2003.

[10] L. Li, W. Huang, I. Y.-H. Gu, and Q. Tian, "Statistical modeling of complex backgrounds for foreground object detection," *IEEE Tran. IP*, vol. 13(11), pp. 1459–1472, 2004.

[11] L. Maddalena and A. Petrosino, "A self-organizing approach to background subtraction for visual surveillance applications," *IEEE Tran. IP*, vol. 17(7), pp. 1168–1177, 2008.

[12] ——, "A fuzzy spatial coherence-based approach to background/foreground separation for moving object detection," *Neural Computing and Applications*, vol. 19(2), pp. 179–186, 2010.

[13] ——, "The sobs algorithm: what are the limits?" *Workshop on Change Detection, CVPR*, pp. 21–26, 2012.

[14] A. Morde, X. Ma, and S. Guler, "Learning a background model for change detection," *Workshop on Change Detection, CVPR*, pp. 15–20, 2012.

[15] O. Javed and M. Shah, "Tracking and object classification for automated surveillance," *ECCV*, pp. 343–357, 2002.

[16] A. Schick, M. Bauml, and R. Stiefelhagen, "Improving foreground segmentation with probabilistic superpixel markov random fields," *Workshop on Change Detection, CVPR*, pp. 27–31, 2012.

[17] K. Toyama, J. Krumm, B. Brumitt, and B. Meyers, "Wallflower: Principles and practise of background maintenance," *ICCV*, pp. 255–261, 1999.

[18] C. R. Wren, A. Azarbayejani, T. Darrell, and A. P. Pentland, "Pfinder: Real-time tracking of the human body," *PAMI*, vol. 19(7), pp. 780–785, 1997.

[19] Z. Zivkovica and F. Heijden, "Efficient adaptive density estimation per image pixel for the task of background subtraction," *Pattern Recognition Letters*, pp. 773–780, 2006.

[20] F. Seidel, C. Hage, and M. Kleinsteuber, "prost : A smoothed lp-norm robust online subspace tracking method for realtime background subtraction in video," *Unpublished, preprint in CoRR*.

[21] I. Haritaoglu, D. Harwood, and L. S. Davis, "W^4S: A real-time system for detecting and tracking people in $2\frac{1}{2}D$," *ECCV*, vol. 5, pp. 877–892, 1998.

[22] H. Nakai, "Non-parameterized bayes decision method for moving object detection," *ACCV*, 1995.

[23] W. E. L. Grimson, C. Stauffer, R. Romano, and L. Lee, "Using adaptive tracking to classify and monitor activities in a site," *CVPR*, pp. 22–29, 1998.

[24] T. Matsuyama, T. Ohya, and H. Habe, "Background subtraction for non-stationary scenes," *ACCV*, 2000.

[25] N. Oliver, B. Rosario, and A. P. Pentland, "A bayesian computer vision system for modeling human interactions," *PAMI*, vol. 22(8), pp. 831–843, 2000.

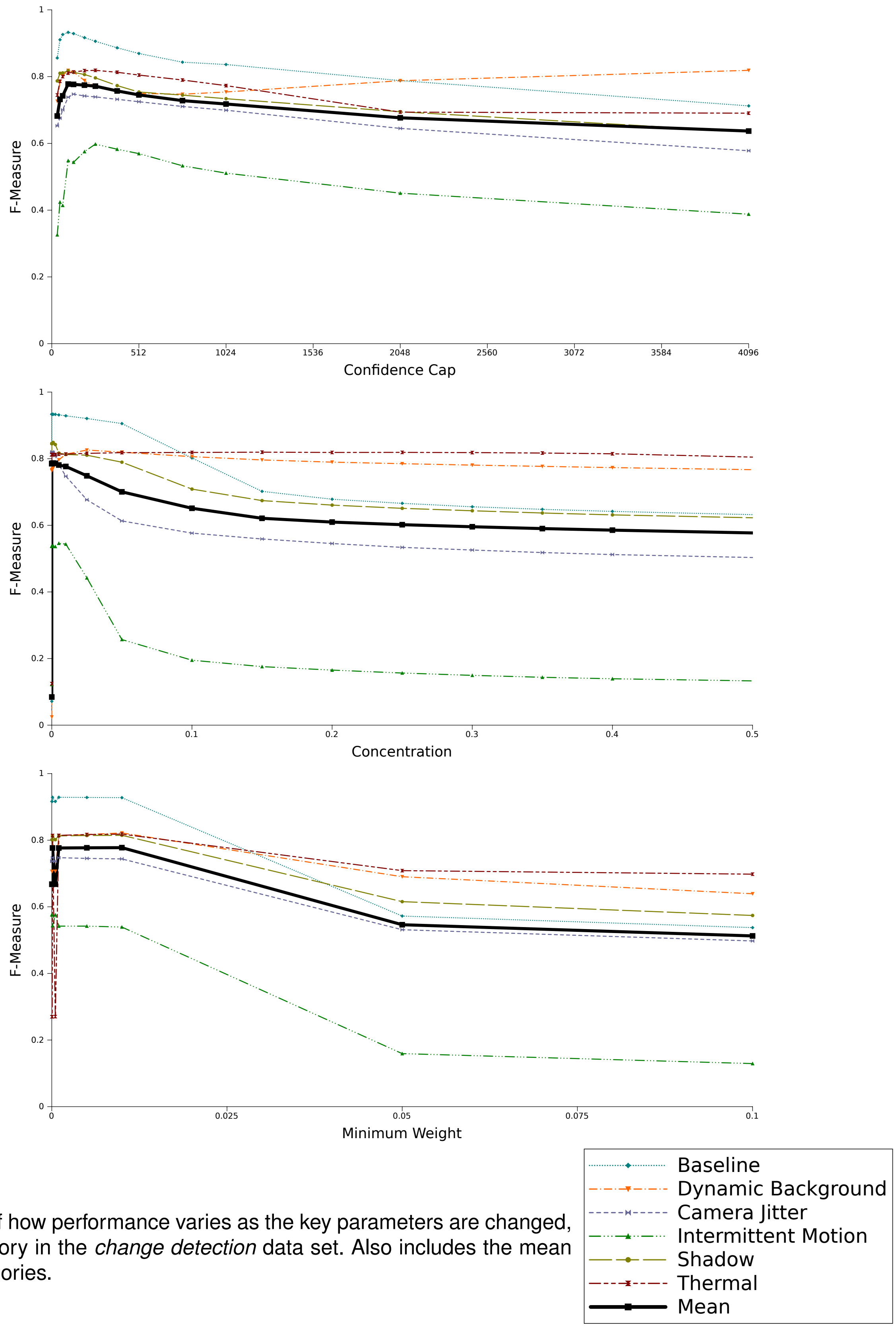

Fig. 3. Plots of how performance varies as the key parameters are changed, for each category in the *change detection* data set. Also includes the mean of all the categories.