Detector adaptation by maximising agreement between independent data sources

Ciarán Ó Conaire, Noel E. O'Connor, Alan F. Smeaton
Centre for Digital Video Processing, Dublin City University, Dublin, Ireland
{oconaire,oconnorn}@eeng.dcu.ie, asmeaton@computing.dcu.ie

Abstract

Traditional methods for creating classifiers have two main disadvantages. Firstly, it is time consuming to acquire, or manually annotate, the training collection. Secondly, the data on which the classifier is trained may be over-generalised or too specific. This paper presents our investigations into overcoming both of these drawbacks simultaneously, by providing example applications where two data sources train each other. This removes both the need for supervised annotation or feedback, and allows rapid adaptation of the classifier to different data. Two applications are presented: one using thermal infrared and visual imagery to robustly learn changing skin models, and another using changes in saturation and luminance to learn shadow appearance parameters.

1. Introduction

In traditional machine-learning, classifiers and detectors are created by training them on ground-truthed data. The creation of ground-truth is time-consuming, usually involving manual annotation of each example. A further difficulty lies in the amount of ground-truth used for training. If the training data contains only a small number of examples, the classifier will be very specific and will not generalise well to other unseen data. If the training data is extensive and contains a large number of examples, the classifier may perform well on a broad variety of data, but will not be optimal for specific datasets. While there is no optimal classifier for all datasets, in this paper we show that in some cases it is possible to use two sources of information, and the redundant information they share, to dynamically create a classifier on new data automatically without any user annotation.

1.1. Related work

In the absense of appropriate training data, two independent sources of information could be used to, in a sense, *train each other*, by providing feedback on an appropriate configuration for optimal detection. Intuitively, two independent detectors, with their parameters selected by maximising agreement, will agree on detections that are correct, and false positives will be excluded since they are uncorrelated and including them would decrease agreement. Agreement between sources has often been measured using mutual information.

Kruppa and Schiele's approach is a good example [3]. Here, detector configurations that correspond to peaks of the agreement function are selected and used to fuse detector outputs in a hierarchical framework. In the face-detection application demonstrating their method, a simple ellipse-based shape detector is used to fuse the outputs of a template-based and colour-based face detector.

Sharma and Davis [8] use a mutual information approach to choose the contour segments in the visual modality in such a way as to maximise an agreement measure between these contours and the detected contours in the corresponding thermal image. Results on segmenting people from the background are quantitatively evaluated using manually segmented ground truth and their method is shown to outperform either visual or infrared analysis alone.

In our previous work [5], we adaptively computed thresholds for foreground detection for multi-spectral video frames so as to maximise the mutual information between the foreground maps of visual and thermal infrared images. A dynamic programming algorithm was described to efficiently investigate the search-space of all possible pairs of thresholds.

In this paper, we further this work by adapting the dynamic programming algorithm to cater for bounded ranges, instead of simple thresholds. Additionally, we generalise the notion of agreement between binary signals, of which mutual information is a special case. Encouraging results are shown on shadow pixel and skin pixel detection without the use of training data or user-specified parameters.

This paper is organised as follows: Firstly, we give a simple illustrative example of shadow pixel detection and show how maximising agreement between complementary data sources provides good detection parameters. Next our algorithm for dynamic bounding that efficiently searches the parameter space to maximise agreement is presented. We then

1-4244-1180-7/07/$25.00 ©2007 IEEE

describe our system for adaptive skin detection, using thermal infrared and visual images, which does not need pre-annotated training data or user-selected thresholds. We give details of a number of experimental trials, demonstrating the benefits of our technique and the importance of adapting detectors to the data. Finally, we present our conclusions and some directions for future work.

2. Illustrative example

Before describing our contribution, we first describe an example application where two data sources could potentially assist each other in determining appropriate parameters for object or event detection. Here, the target application is shadow pixel detection.

Shadow detection is a useful component in background modelling algorithms, as it eliminates foreground pixel errors caused by colour changes due to shadows cast by moving objects. Shadow pixels can be modelled as a bounded decrease in brightness:

$$l_3 \leq \hat{V}_i \leq l_4 \tag{1}$$

where $\hat{V}_i$ is the relative change in luminance of pixel i compared to the background pixel, and is given by $\hat{V}_i = V_i/V_i^{BG}$. The selection of appropriate bounds can be done empirically, or can be trained on pre-annotated data. However, if we make the *assumption* that shadows also cause a decrease in the pixel's colour saturation [6], we then have a second source of data that can assist in our parameter selection. This assumption may not be true in general, but is a useful means of illustrating the approach. We model the shadow-pixel in saturation space as a bounded range given by

$$l_1 \leq \hat{S}_i \leq l_2 \tag{2}$$

where $\hat{S}_i$ is the relative change in saturation, and is given by $\hat{S}_i = S_i/S_i^{BG}$. Given an image containing a cast shadow, applying equation (1) to the luminance change image produces a binary image. A binary image is similarly obtained by applying equation (2) to the associated saturation change image. If the parameters $\{l_1, l_2, l_3, l_4\}$ are selected correctly, we expect there to be a strong agreement between the two binary masks. We propose to dynamically set these parameters so that they maximise agreement. To measure agreement between binary images from different modalities, we previously used mutual information as an agreement measure [5]. This measure returns high values when there is significant agreement and avoids the trivial case of *complete agreement* which could be achieved by setting the parameters to classify every pixel as shadow. We now discuss two possible agreement measures: mutual information and Kendall's tau (τ).

Since we are dealing only with binary images, a 4-value co-occurrence histogram is all that is needed to compute agreement. Given 2 binary images, X and Y, with N pixels each, we let u and v be binary-valued variables, with $C_{u,v}$ equal to the number of pixels whose classification is u in image X and v in image Y. The mutual information, μ_{XY}, between the pair of binary images, X and Y, is computed as follows:

$$p_{XY}(u,v) = \frac{C_{u,v}}{N} \tag{3}$$

$$p_X(u) = p_{XY}(u,0) + p_{XY}(u,1) \tag{4}$$

$$p_Y(v) = p_{XY}(0,v) + p_{XY}(1,v) \tag{5}$$

$$\mu_{XY} = \sum_{u\in\{0,1\}} \sum_{v\in\{0,1\}} p_{XY}(u,v) log \frac{p_{XY}(u,v)}{p_X(u)p_Y(v)} \tag{6}$$

As well as mutual information, another measure that has been frequently used to determine correlation between signals is Kendall's τ [2]. This measure can be computed using the same histogram counts:

$$\tau = \frac{p_{XY}(0,0)p_{XY}(1,1) - p_{XY}(0,1)p_{XY}(1,0)}{\sqrt{p_X(0)p_Y(0)p_X(1)p_Y(1)}}. \tag{7}$$

Alternative agreement measures, other than the two given here, are also possible and are all functions of the four values of $C_{u,v}$. Regardless of the choice of agreement measure, maximising this measure requires finding the optimum parameters in high-dimensional space, 4-dimensions in the case of shadow detection. As with most complex high-dimensional problems, finding a global maximum cannot be guaranteed. However, the Simplex algorithm [4] or some other gradient ascent method could be used to find a good local maximum. We propose instead to use a dynamic programming-based solution, similar to that used in [5], to optimise two of the parameters at a time, iterating between data sources until we converge on a solution. In the next section, our dynamic bounding algorithm is explained in detail.

3. Dynamic bounding algorithm

In order to choose the optimum pair of bounds that will maximise the agreement between the bounded image and the binary source, a brute-force search could be employed. Trying all pairs of thresholds from a discrete set of K elements has complexity in the order of $O(NK^2)$, where N is the number of pixels in the image. The dynamic programming algorithm described here is of order $O(K^2 + N)$ and evaluates all possible pairs of bounds in a discrete set.

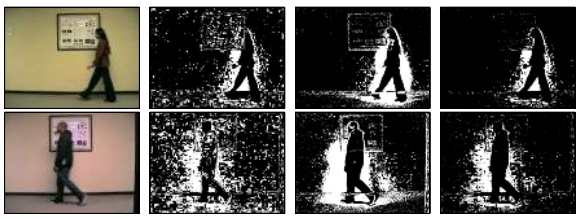
The input to the algorithm is a discrete set of thresholds, $\hat{A} = \{a_1, a_2, ..., a_K\}$, a binary signal, X, and a real-valued signal, Y, of the same size as X. The goal is to select bounds for signal Y, such that when a binary signal, Y^*, is created using these bounds, its agreement with signal X is maximised. The output is a mapping array, $C_{p,q}(i,j)$,

Input: Threshold list $\hat{A}$ and signals X and Y
with $X = \{x_1, x_2, ..., x_L\}, Y = \{y_1, y_2, ..., y_L\}$
Initialise count maps to zero: $C_{*,*}(*,*) = 0$
$c_0 = \#\{k; x_k = 0\}$ // count zeros in binary signal
$c_1 = \#\{k; x_k = 1\}$ // count ones in binary signal
For all data points (x_k, y_k)
  Find largest $a_i \in \hat{A}$ such that $a_i \leq y_k$
  Find smallest $a_j \in \hat{A}$ such that $y_k \leq a_j$
  $C_{x_k,0}(1,1) + +$
  if (a_i and a_j exist)
    $C_{x_k,0}(1,j) - -$
    $C_{x_k,0}(i+1, j+1) + +$
  end
$C_{0,0} = integralImage(C_{0,0})$ // integrate markers
$C_{1,0} = integralImage(C_{1,0})$ // integrate markers
$C_{0,1} = c_0 - C_{0,0}$
$C_{1,1} = c_1 - C_{1,0}$

Figure 1. Pseudocode for algorithm in section 3

which gives the number of binary pairings of $x_k = p$ and $y_k^* = q$ when the bounds selected are a_i and a_j, with $i \leq j$. These counts can then be normalised and used in equation (6) or (7) to create an *agreement surface*, providing the agreement score for all possible bounding parameter selections. The bounds a_i and a_j that give the maximum agreement can then be selected. The pseudocode for the algorithm is given in figure 1. The $integralImage()$ function refers to the standard dynamic programming method that efficiently replaces each pixel with the sum of all pixels in the rectangle whose opposite corners are this pixel and the pixel in $(1,1)$ [11].

4. Shadow detection



(a) image (b) bounded S (c) bounded V (d) shadow

Figure 2. Shadow parameter selection using Kendall's τ

Figure 2 shows the results of shadow detection on two images, with Kendall's τ used as the agreement measure. We used data from the Terrascope dataset [1]. For our experiments, we used a median background image, and 256 equally spaced thresholds between $1/255$ and 1. The four parameters are selected so as to maximise the agreement between the binary images obtained by bounding the saturation and luminance images, as in equations (1) and (2). For all tests, the initial parameters were set at $\{0.3, 0.97, 0.3, 0.97\}$, though other reasonable initialisations produced similar results. Parameters $\{l_1, l_2\}$ were optimised first, and then $\{l_3, l_4\}$. This continued until convergence.

Image 1 (Gupta) converges in 7 iterations to $\{0.3686, 0.9294, 0.3889, 0.9500\}$ with $\tau = 0.3680$. Image 2 (Crasto) converges in 5 iterations to $\{0.4549, 0.9333, 0.5725, 0.9490\}$ with $\tau = 0.3088$. Using either Kendall's τ or mutual information as the agreement measure provides good results for images in this dataset. Additionally, our method is much more efficient than a Simplex search, which required over 150 iterations.

Overall, shadow detection using this method did not perform well on other data we investigated, such as the ground-truthed shadow data provided by [6]. Our assumption that saturation decreases is often not true as many backgrounds do not have strong colour content. Additionally, the two sources (luminance and saturation) cannot really be considered independent, as they come from the same sensor. In scenarios where the assumption is true, the method might be improved by first removing 'true foreground pixels'; such as those whose hue has changed significantly. We next describe a more practical application of our method, using thermo-visual information for adaptive skin detection.

5. Skin pixel detection

(a) (b)

Figure 3. Examples of (a) visible and (b) infrared input images

Figure 3 shows a colour image and its corresponding thermal infrared image. Skin pixels lie in a particular subspace in both the thermal and visible domains. Similar to our shadow detection example, we use simple bounds to model skin in both the colour and infrared domains, and can exploit the shared information between the modalities to compute the parameters for both these subspaces. In the visible domain, we select a certain bounded subspace of the HSV space to indicate a possible skin pixel. Using $\{l_1, l_2, l_3, l_4, l_5, l_6\}$ as the boundaries of the subspace, a pixel i belongs to this subspace if its colour components

in HSV space, (H_i, S_i, V_i) conform to:

$$l_1 \leq H_i \leq l_2 \quad (8)$$

$$l_3 \leq S_i \leq l_4 \quad (9)$$

$$l_5 \leq V_i \leq l_6. \quad (10)$$

Since the hue component can be considered circular, we set $H_i \leftarrow (H_i + 128)\ mod\ 256$, so that red, the dominant hue in skin pixels, is in the centre of the band. In the thermal infrared images, we use a similar model for the appearance of skin pixels, with pixel I_i being a potential skin pixel if

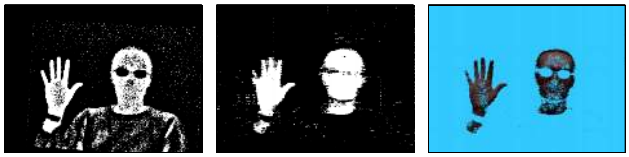
$$l_7 \leq I_i \leq l_8 \quad (11)$$

where $\{l_7, l_8\}$ are the thermal brightness boundaries. Therefore, the parameters for our models are fully represented by $L = \{L_{VIS}, L_{IR}\} = \{\{l_1, l_2, .., l_6\}, \{l_7, l_8\}\}$.

In figure 4, examples of the use of these models are shown in relation to figure 3. Setting $L = \{\{78, 159, 60, 255, 3, 139\}, \{67, 137\}\}$ maximises the Kendall's τ agreement measure. Pixels within the hue, saturation and value boundaries are shown in figure 4(a)-(c). Figure 4(e) combines (a)-(c), showing pixels that are within all the colour boundaries, and are considered possible skin pixels. Figure 4(d) shows infrared pixels that fall within the thermal boundary, and are therefore considered possible skin pixels.

Ideally, if there are skin regions present in the scene, and there are not many skin-like distractors present in visible or infrared, then there should be a high level of agreement between the binary images in figure 4(d) and (e). By selecting pixels that appear as skin in both modalities (binary AND fusion), figure 4(f) is produced.

(a) Bounded H. (b) Bounded S. (c) Bounded V.



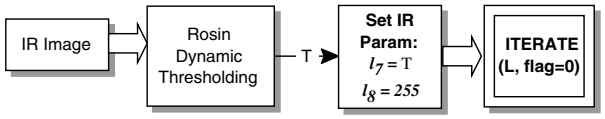
(d) Bounded IR. (e) Bounded HSV. (f) Likely skin pixels.

Figure 4. Examples of bounded (a) hue, (b) saturation, (c) value, (d) infrared and (e) HSV. Binary AND fusion of (d) and (e) produce the skin pixels in (f).

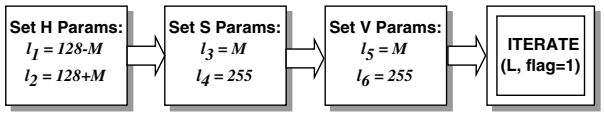
5.1. System overview

The input to our system is a colour image, the corresponding thermal infrared image and an initialisation method. The initialisation method provides a binary image either from the colour or thermal image. The other modality's bounds will be optimised to maximise agreement. Bounds are then alternatively optimised iteratively until convergence.

Figure 6 shows the two initialisation methods used in this work. The first method applies a dynamic threshold to the IR image using Rosin's method [7]. The second method uses predefined colour bounds to provide the initial binary image. In our tests, we set $M = 255/5$, which sets quite a broad range, so almost all skin-like pixels will be included. After initialisation, the system will iteratively optimise all the parameters until it converges, as illustrated in figure 5. The *flag* variable indicates whether the IR bounds should be optimised first. When we are optimising pairs of colour bounds, such as hue bounds $\{l_1, l_2\}$, some pixels may already be excluded since they are outside the other colour bounds. We cater for this by excluding these pixels from processing and adding them on to the appropriate counts at the end (either to $C_{0,0}$ or $C_{1,0}$). The final outputs are (i) the set of 8 parameters, L, (ii) 2 binary maps (one for each modality) and (iii) an agreement value score. While mutual information performs well, its value does not change if one of the binary images is inverted. This is not a desirable property, therefore we chose Kendall's τ as the agreement measure.



(a)



(b)

Figure 6. Initialisation Methods: (a) Infrared-based and (b) Colour-based initialisation

6. Experimental Results

6.1. Initialisation evaluation

In order to compare the initialisation methods of figure 6, we ran our algorithm on $6,697$ images from 7 thermo-visual video sequences. We investigated which method would cause convergence to the highest agreement value. The results are given in table 1. Both methods converged in a similar number of iterations on average, as shown in columns 3 and 5. Neither method showed superiority, with both methods having roughly equal performance on average, and converging to the same parameters about one-third

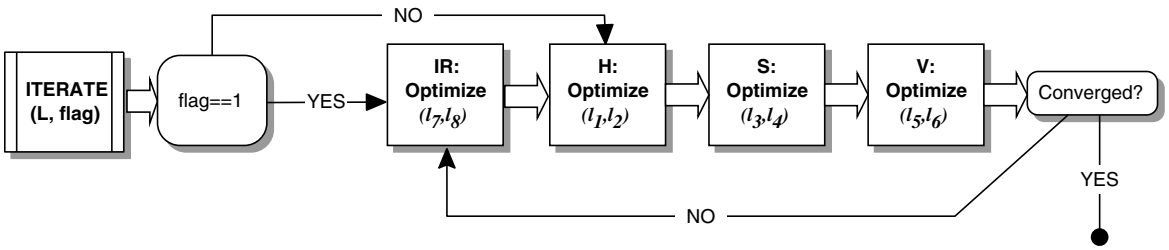


Figure 5. Iteration function

Seq	Frame count	Method 1 Iter.	Method 1 %	Method 2 Iter.	Method 2 %	Both %
A	235	3.42	14.89	3.91	16.60	68.51
B	406	3.63	19.46	4.43	3.20	77.34
C	615	4.05	9.27	4.16	79.67	11.06
D	2984	4.33	83.88	3.86	16.09	0.03
E	306	3.39	0.00	3.61	100.00	0.00
F	997	4.10	47.14	4.09	29.19	23.67
G	1154	3.91	24.78	4.10	12.65	62.57
ALL	6697	3.83	28.49	4.02	36.77	34.74

Table 1. Table above indicates the percentage of frames for which each initialisation method converged to the highest agreement score, for all seven sequences tested. The rightmost column indicates that both methods converged to very similar configurations, within a small tolerance.

of the time. Sequence D contains a lot of skin-like pixels, due to the colour of the floor, causing the colour-based initialisation to perform poorly in this sequence. On the other hand, sequence E contains many people and therefore a lot of 'hot' pixels, causing the infrared-based initialisation to perform poorly in this sequence. By running the algorithm with both methods, and selecting the set of parameters with greater agreement, high quality skin detection is obtained.

6.2. Fusion evaluation

After selecting appropriate parameters for the skin models, we have a binary image from visible and from infrared as sources of evidence as to whether or not a pixel is a skin-pixel. These binary masks can be fused for a final classification decision. We evaluated 5 simple fusion schemes on 16 ground-truthed skin-detection images. The fusion schemes were (i) binary AND, (ii) binary OR, (iii) Visible only, (iv) IR only and (v) region-based fusion. The region-based scheme examined all the connected-component regions in the binary OR image. If a region had 10% or more of its pixels also belonging to the binary AND image, then it was included. Otherwise, only the pixels in that region from the AND image were used. Although the threshold of 10% is ad-hoc, a range of thresholds were found to perform similarly. The results are given in table 2. As expected, the AND fusion achieves very high precision and the OR fusion achieves high recall. Using IR only performs well, compared to visible only, as there were fewer distractors at a similar brightness to skin in the dataset, compared to skin-colour-like distractors in the visual domain. Using the F_1 measure [10] to combine precision and recall, the region based fusion performed best overall.

	AND	OR	VIS	IR	REG
Precision	0.976	0.605	0.641	0.776	0.849
Recall	0.516	0.878	0.664	0.731	0.838
F_1	0.675	0.717	0.652	0.753	0.843

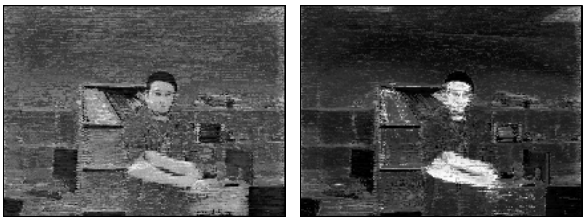
Table 2. Binary fusion methods evaluation.

6.3. Adaptive probabilistic model

The described method does not exploit any temporal information available in video sequences. However, we now show how our method can be used to automatically create probabilistic models of skin and background colour appearance and we compare this to a pre-learned human-annotated colour model. Manually annotated skin and background images are available online as part of Sigal et al.'s work on skin segmentation [9]. Using a similar approach to the original work, these samples were used to create $32 \times 32 \times 32$ RGB colour histograms for both skin and background appearance, and these histograms were normalised and used as probabilistic models of the skin and background. For a given colour image, Bayes' rule can be applied and these models create a log-likelihood image, giving each pixel a skin-likelihood value. The pre-trained (PT) model was created using 723 images which contained $8,929,954$ skin pixel samples and $129,642,003$ background pixel samples.

Our skin and background models were created in a similar fashion but the samples they are trained with were all automatically selected by our method. For each image in the video sequence, we detect skin pixels by maximising agreement and then performing binary AND fusion to achieve high precision. All these pixels are inserted into our skin model. All pixels which are classified as background by both IR and visible are inserted into our background model (NAND fusion). All other pixels are ambiguous, so are ignored. For each video image we tested, we used up to 100 of

the previous images for training our model. Figure 7 shows examples of the log-likelihood image created by our method versus the PT model. Figure 8 shows the ROC curve that indicates the improvement of using adaptive skin modelling over a pre-trained model.



(a) PT model (b) our model

Figure 7. Examples of log-likelihood images created by (a) the PT model and (b) our method

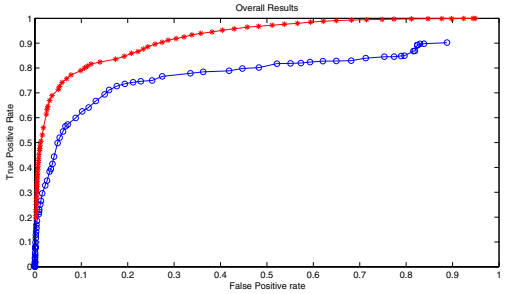


Figure 8. ROC curve for our automatically learned probabilistic model (red stars) vs. a PT model (blue circles)

Further experimental results can be found online at http://www.eeng.dcu.ie/~oconaire/otcbvs07/. These results include image and video results from subsections 6.1 and 6.3, graphs showing the adaptation of the skin bounds to changing environments and an illustrative example of using our method for skin detection on the OTCBVS benchmark dataset.

7. Conclusion and discussion

In this paper a method is described for automatically and efficiently choosing appropriate parameters to maximise agreement between two binarised sources of information. We have given examples of agreement measures for binary images, such as mutual information and Kendall's τ, and shown that both measures are functions of the same four counts of binary pairings. Using Kendall's τ as an agreement measure, experimental results were shown using our method for shadow detection, with the assumption of background saturation change, and for skin detection in thermo-visual imagery.

In cases where no skin is present our method fails since it relies on common information being present in the sources. Failure of the method is usually indicated by low agreement values, the hue bounds lying outside the normal skin range and the IR lower bound dropping below the Rosin threshold. Since our algorithm is a general method for finding pixel appearance subspaces that are in strong agreement between data sources, skin may not always correspond to the highest agreement peak. For example, a cold blue bottle might be distinct enough in both modalities to return a high agreement value, though the initialisation method we use targets skin and as such, may find the best *local peak* corresponding to skin.

Colour-spaces other than HSV may be better at separating the skin and background subspaces using our bounded model of skin. We believe that this method could be used to dynamically select the optimum colour-space, again by testing multiple colour-spaces and choosing the one that returns the highest agreement with infrared. This is targeted as future work.

References

[1] C. Jaynes, A. Kale, N. Sanders, and E. Grossmann. The terrascope dataset: A scripted multi-camera indoor video surveillance dataset with ground-truth. In *Procs of the IEEE Workshop on Visual Surveillance and Performance Evaluation of Tracking and Surveillance*, 2005. 3

[2] M. Kendall. A new measure of rank correlation. *Biometrika*, 30(1-2):81–93, 1938. 2

[3] H. Kruppa and B. Schiele. Hierarchical combination of object models using mutual information. In *BMVC*, 2001. 1

[4] J. Nelder and R. Mead. A simplex method for function minimization. *The Computer Journal*, 7:308–313, 1965. 2

[5] C. Ó Conaire, N. O'Connor, E. Cooke, and A. Smeaton. Detection thresholding using mutual information. In *VISAPP: International Conference on Computer Vision Theory and Applications, Setúbal, Portugal*, Feb 2006. 1, 2

[6] A. Prati, I. Mikic, M. Trivedi, and R. Cucchiara. Detecting moving shadows: Algorithms and evaluation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 25(7):918–923, July 2003. 2, 3

[7] P. L. Rosin. Unimodal thresholding. *Pattern Recognition*, 34(11):2083–2096, 2001. 4

[8] V. Sharma and J. W. Davis. Feature-level fusion for object segmentation using mutual information. In *IEEE Workshop on Object Tracking and Classification in and Beyond the Visible Spectrum*, June 2006. 1

[9] L. Sigal, S. Sclaroff, and V. Athitsos. Skin color-based video segmentation under time-varying illumination. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 26(7):862–877, July 2004. 5

[10] C. J. Van Rijsbergen. *Information Retrieval*. Dept. of Computer Science, University of Glasgow, Butterworths, London, 2nd edition edition, 1979. 5

[11] P. Viola, M. J. Jones, and D. Snow. Detecting pedestrians using patterns of motion and appearance. In *IEEE International Conference on Computer Vision (ICCV)*, volume 2, pages 734–741, Oct 2003. 3