Colour Image Segmentation by Non-Parametric Density Estimation in Colour Space

P. A. Bromiley, N.A.Thacker and P. Courtney
Imaging Science and Biomedical Engineering Division
Medical School, University of Manchester
Manchester, M13 9PT, UK
paul.bromiley@man.ac.uk

Abstract

A novel colour image segmentation routine, based on clustering pixels in colour space using non-parametric density estimation, is described. Although the basic methodology is well known, several important improvements to the previous work in this area are introduced. The density is estimated at a series of knot points in the colour space, and clustering is performed by hill climbing on this density function. The hill climbing is constrained such that no step crosses an intermediate Voronoid cell, ensuring that all salient clusters are detected. Most importantly, the problem of scale selection has been addressed using a statistically motivated approach, by placing the knot points according to an estimate of the noise in the original images, taking full account of error propagation in the algorithm. The algorithm has been evaluated both on synthetic data and in the context of its application in a machine vision system, specifically the calibration of velocity estimates extracted from a novel infrared sensor used in a fall detector. The application of the technique to medical images and texture recognition is also discussed.

1 Introduction

Image segmentation is a fundamental task in computer vision, and the application of segmentation to colour images is used in a wide range of tasks, including content-based image retrieval for multimedia libraries [7], skin detection [2], object recognition [3], and robot control [9]. A variety of approaches to this problem have been adopted in the past, which can be divided into four groups: pixel-based techniques, such as clustering [7]; area-based techniques, such as split-and-merge algorithms [9]; edge-detection, including the use of colour-invariant snakes [4]; and physics-based segmentation [6]. A review of the methods and applications of colour segmentation is given in [9].

The approach to image segmentation adopted in this work relies on clustering of pixels in feature space using non-parametric density estimation. The subject of clustering, or unsupervised learning, has received considerable attention in the past and the clustering technique used here is not original [7]. However, much of the work in this area has focused on the determination of suitable criteria for defining the “correct” clustering. We have adopted a statistically motivated approach to this question, defining the size required

BMVC 2001 doi:10.5244/C.15.30

for a peak in feature space to be considered an independent cluster in terms of the noise in the underlying image. The method follows directly from our previous work on the formation of self-generating representations using knowledge of data accuracy [10]. Thus we can maximise the information extracted from the images without introducing artefacts due to noise, and also define an optimal clustering without the need for testing the results with other, more subjective criteria.

The segmentation process maps pixels from an arbitrary number n grey-scale images into an n-dimensional grey-level space, and calculates a density function in that space. A colour image can be represented as three greyscale images, showing for instance the red, green and blue components of the image, although many alternative three-dimensional schemes have been proposed [11]. Therefore a colour image will generate a 3D space. An image showing a number of well-defined, distinct colours will generate a number of compact and separate peaks in the space, each centered on the coordinates given by the average red, green and blue values for one of the colours. The algorithm then uses the troughs between these peaks as decision boundaries, thus classifying each pixel in the image as belonging to one of the peaks. Each peak is given a label, which is then assigned to the pixels clustered onto that peak, and an image of these labels is generated as output.

Performance evaluation for colour segmentation algorithms is problematic, largely because the desired result is ill-defined. The aim is often to extract semantically meaningful regions of images, but it then becomes difficult to evaluate algorithmic performance in the absence of ground truth data. Several authors have proposed evaluation techniques, using statistical [2], semantic [7], or ground-truth based [8] metrics. Here we evaluate the performance of the algorithm in terms of its integration into a larger machine vision system. The algorithm was developed as part of a system that used a novel differential infrared sensor as a fall detector. The development and evaluation of the fall detector algorithms are described in more detail elsewhere [1]. The fall detector used an MLP neural network to automatically recognise the temporal patterns of vertical velocities produced by falling human subjects, and so a vital step in the development of the system involved measuring the correlation between velocity estimates extracted from the infrared sensor and the physical velocities present in the scene. Therefore an actress was employed to perform a set of simulated falls, which were recorded with both the infrared sensor and a colour CCD video camera. The colour segmentation routine was used to extract the position of the actress in frames of the colour video, allowing her velocity to be calculated. The ability of the algorithm to provide gold-standard data for this correlation analysis gave a subjective measure of its performance. However, the basic statistical properties of the algorithm were also evaluated through the use of synthetic data.

The ultimate measure of clustering quality can be defined with reference to the Bayes optimal clustering, the clustering that would be achieved if the underlying distributions that generated the data set were known. The algorithm described here places a decision boundary at the point of lowest data density between any pair of peaks i.e. the point where the sum of the two distributions is a minimum. This is equivalent to the assumption that the local generators of the data are equally likely at this point. In fact, there are an infinite number of pairs of functions that could be summed to produce any given distribution, and so no non-parametric method can replicate the Bayes optimal classification for all data. However, if the local data density at the position of the decision boundary is low, the number of errors introduced will be small. Therefore, this approach may be considered as close as any non-paramentric method can approach to a Bayes optimal clustering.

2 Method

The aim of the colour segmentation routine described here was to identify distinctly coloured regions in an image that corresponded to physical objects present in the scene. However, an image will contain both chromatic and achromatic information, so an object of a single colour that is partly in shadow may appear to consist of two regions of different colour, and so may be split into two regions by the segmentation. Some colour spaces (e.g. HSI, YIQ) separate the achromatic (I, Y) and chromatic (HS, IQ) information onto different axes. Therefore, the achromatic information can be discarded and the segmentation performed on the remaining two chromatic dimensions. This confers the additional advantage of reducing the dimensionality of the problem, and so reducing the processor time required. This approach was tried with limited success [1]. A more effective way of removing the intensity information was found to be normalising the RGB values prior to any colour space conversions, using $r = R/(R+G+B)$, $g = G/(R+G+B)$, and $b = B/(R+G+B)$ [5]. This is equivalent to finding the intersection of the colour vectors in RGB space with the plane of constant intensity passing through (1,0,0), (0,1,0) and (0,0,1). It also retains the advantage of reducing the dimensionality of the colour space from three to two, since $r+g+b=1$ and so any two of these components is sufficient to describe the normalised colour vector. However, in order to form statistical groupings of data points we must also have knowledge regarding data accuracy. It is therefore desirable to use the colour space that has the simplest possible error propagation from RGB. Therefore a new colour space referred to as IJK was developed, a simple rotation of the RGB colour space with no scaling, such that one axis lay along the vector $R = G = B$, and so represented the intensity I. The second axis J lay along the projection of the R axis onto the plane normal to the intensity axis, and the third axis K was perpendicular to the others. The conversion from RGB was performed using the rotation matrix

$$\begin{pmatrix} I \\ J \\ K \end{pmatrix} = \begin{pmatrix} \frac{1}{\sqrt{3}} & \frac{1}{\sqrt{3}} & \frac{1}{\sqrt{3}} \\ \frac{2}{\sqrt{6}} & \frac{-1}{\sqrt{6}} & \frac{-1}{\sqrt{6}} \\ 0 & \frac{1}{\sqrt{2}} & \frac{-1}{\sqrt{2}} \end{pmatrix} \begin{pmatrix} R \\ G \\ B \end{pmatrix}.$$

When this rotation was applied to the normalised RGB space, the values for the intensity axis I were uniform across the image as expected. In practice, any arbitrary set of perpendicular axes in the normalised colour space can be used in the segmentation. The algorithm was tested with all combinations of pairs of r, g and b; with the the hue and saturation fields from HSI; with the I and Q fields from YIQ; and with the J and K fields from the new colour space. Ignoring differences due to error propagation, no significant advantage of one of these choices over the others was found.

The first step in the segmentation was to map the pixels $\mathbf{x}_{i=1,N_p}$ from the original images into colour space, producing an n-dimensional scattergram $S(\mathbf{g}_{1,n})$ where grey levels $\mathbf{g}_{1,N_p}$ of pixels at the same positions in the images are used as coordinates in feature space. To map out the whole space would be prohibitively expensive in terms of memory and processor time in high dimensional problems, and so the data itself was used to define a list of N_k knot points $\mathbf{G}_{i=1,N_k<N_p}$ that span the space. Clustering algorithms have been extensively studied in the past, and part of the original work in this algorithm concerns the technique used to generate this list. The traditional problem has been the definition of the "correct" clustering i.e. the correct scale at which to define peaks in the colour space. If clustering is performed at too small a scale, then artificial peaks due to noise

will be generated. Conversely, if the scale is too large then small but salient peaks will be absorbed into nearby, larger peaks. One approach to this problem has been to perform the segmentation at a range of scales, and then select the correct scale by applying some measure of the quality of clustering to the resultant images [7]. However, this does not solve the underlying problem of selecting the correct scale: it simply redefines the problem as one of selecting the correct measure of the quality of clustering.

A statistically motivated approach to the problem of scale selection was adopted, attempting to define the correct clustering in terms of the properties of the noise in the underlying images. The noise in the red, green and blue fields from the original colour images was estimated using the width of vertical and horizontal gradient histograms. It was assumed that the noise in each field was uniform across the image, and that the errors in each field were uncorrelated. The errors were propagated through the normalisation and colour space conversion equations, using the usual equation

$$\sigma_f{}^2 = \frac{\partial f}{\partial x}^2 \sigma_x{}^2 + \frac{\partial f}{\partial y}^2 \sigma_y{}^2 + ...,$$

where $x, y...$ are the variables in the equation for f through which errors are being propagated. Then, all distance measurements in the final colour space were scaled by the local error along each axis, such that the noise was approximately unitary across the whole of the space. The knot points could therefore be placed a distance of $0.5\sigma = 0.5$ apart, ensuring that smaller peaks due to noise artefacts were ignored, whilst all salient peaks were mapped with enough knot points that they remained discriminable. Although this method resulted in only a rough approximation to the noise in the final colour space, it was found to be accurate enough for the purposes of the algorithm. In order to minimise the number of knot points N_k, the data itself was used to generate them. The algorithm looped over the data points, maintaining a list of knot points, and adding the data point to the list if it was further from all the existing knot points than the threshold distance of 0.5σ.

Once the space had been mapped in this way, the density at each knot point was estimated using a common non-parametric techinique [7]. The density function $f(\mathbf{g})$ was obtained by convolving the data set $\mathbf{g}_i$ with a unimodal density kernel K_{σ_B},

$$f(\mathbf{g}) = \frac{1}{N_p} \sum_{i=1,N_p} K_{\sigma_B}(\mathbf{g} - \mathbf{g}_i).$$

The convolution kernel used was a Gaussian,

$$K_{\sigma_B}(\mathbf{g}) = \frac{1}{\sqrt{2\pi}\sigma_B} e^{-\frac{\mathbf{g}^2}{2\sigma_B^2}},$$

where σ_B was the width of the kernel. The noise in the scaled colour space was unitary, so if the width of the kernel was set to unity the space was blurred at the scale of the measurement accuracy, ensuring that the approximation function for the density was smooth. However, the width of the kernel could be increased to introduce additional blurring.

The next step in the algorithm was to loop over the data points $\mathbf{g}_i$, find the nearest knot point $\mathbf{G}_i$ to each, and then hill climb in the list of knot points to find the local peak. A list of peaks was constructed, maintaining a record of the number of data points assigned to each, and then sorted so that the peaks could be labelled in order of significance. The hill climbing was then performed a second time to assign these labels to the data points. A new image showing the labels was generated as output.

The hill climbing was further subdivided into several stages. The first was to run through the list of knot points, find the m nearest neighbours to each, and calculate the gradient in between the point $\mathbf{G}_i$ and each of its neighbours $\mathbf{G}_{j=1,m}$. A pointer was then added to the knot point leading to the neighbour in the direction of greatest gradient. In order to make this step more statistically rigorous, rather than using a simple gradient

$$\nabla f_{i,j} = \frac{f(\mathbf{G}_j) - f(\mathbf{G}_i)}{\mathbf{G}_j - \mathbf{G}_i},$$

the contribution

$$\nabla f_{i,j} = \frac{f(\mathbf{G}_j)}{\sigma_B\sqrt{2\pi}} e^{-\frac{(\mathbf{G}_j - \mathbf{G}_i)^2}{2\sigma_B{}^2}},$$

that a Gaussian function with standard deviation equal to the blurring parameter used in the previous stage, centered at the neighbouring knot point $\mathbf{G}_j$ and normalised to its height $f(\mathbf{G}_j)$, would have had to the point being considered was used. This was equivalent to a gradient calculation with an exponentially weighted distance measure, and so biased the choice in favour of closer points, preventing small peaks from being eliminated.

In order to perform the hill climbing reliably a constraint was placed on the neighbouring knot points considered. The knot points tesselate the space with Voronoid cells, and no step should cross over an intermediate cell. If this were allowed, it would be possible for the hill climbing to cross valleys in the density function and link together statistically discriminable peaks. Therefore, any step that crossed an intermediate cell was rejected. Referring to Fig. 1, if the step **AB** was being considered, it was compared to all other steps from the same point e.g. **AC**. The planes bisecting these vectors are necessarily the Voronoid boundaries if no intermediate cells lie between the knot points. Therefore, if the plane bisecting **AC** intersects **AB** (point I) at less than half of its length (to the mid-point M), **AB** must cross an intermediate cell. The angle **AB**∠**AC** is given by

$$\theta = cos^{-1}\frac{\mathbf{AB}.\mathbf{AC}}{\mid \mathbf{AB} \mid\mid \mathbf{AC} \mid}.$$

Making use of the fact that **AI** is the hypotenuse of a right-angled triangle, the condition for a step to be rejected is

$$\mid \mathbf{AM} \mid > \mid \mathbf{AI} \mid \Rightarrow 1 > \frac{\mid \mathbf{AC} \mid^2}{(\mathbf{AB}.\mathbf{AC})}.$$

The processor time required by the algorithm is dictated by the number of knot points used to map the space, and this also controls the accuracy of the segmentation. The most difficult synthetic problems presented to the algorithm were found to require a threshold distance of 0.5σ. For images of real scenes, producing feature spaces with large numbers of peaks, this threshold distance would require an impractical amount of processor time. However, the algorithm produced semantically meaningful segmentations for all images of real scenes presented to it with a threshold distance of σ, requiring approximately 110 minutes of processor time on a Pentium III 800 MHz PC. It was also found that for the majority of images the threshold distance could be increased to as much as 5σ with virtually no loss in performance, reducing the time taken to 5 minutes.

3 Results

In order to demonstrate the basic properties of the algorithm, a synthetic data set that exhibited the most difficult problem it could be expected to solve was generated. This data set consisted of the two images shown in Fig. 2. Each image had a dynamic range of 50, with the intensity varying smoothly such that they generated an extended, horseshoe-shaped cluster in feature space. A circular region was removed from the same position in both images to generate a second, compact cluster. Finally, uncorrelated Gaussian noise with $\sigma_A = 6$ was added to each. Fig. 2c shows the scattergram for these two images, overlaid with the map of knot points produced by the segmentation algorithm. Due to the difficulty of this segmentation problem it was found necessary to place the knot points 0.5σ apart in order to separate the two clusters. Fig. 2d shows the output of the algorithm, which clearly detected the smoothly varying background and the circular regions as separate clusters in feature space. At the point of closest approach the two features overlapped at approximately the $2\sigma_A$ points. In this simple example the result of the segmentation was dictated almost entirely by the knot points along the trough in the density function between the two peaks in feature space. Since the knot points were extracted from the data itself, the position of the decision boundary could vary by up to $0.5D$ from the trough in the underlying data density, where D was the threshold distance in units of σ between the knot points. The segmentation routine therefore produced approximately 5% to 10% misclassification for this data, which can be seen in the segmentation result. The important feature of this data is that it demonstrates the ability of the algorithm to link together the extended "horseshoe" cluster without linking it to the enclosed circular cluster, even at the points in the former which lay closer in colour space to the peak of the latter. This has important implications for the segmentation of colour images, where inter-reflections between objects of different colour can cause the apparent colour of an object to vary smoothly over its surface, even if its true colour is uniform. This can be considered an important criterion for any colour segmentation algorithm, since those algorithms which assume compact, circular clusters in colour space typically fail under these conditions.

The synthetic images showed that the algorithm had the expected statistical properties. In order to evaluate its ability to extract semantically meaningful regions from images, it was applied to a set of colour video sequences of an actress simulating falls, recorded as part of the fall detector development project described above. Fig. 3a shows one frame from one of the sequences, and the J and K fields of the normalised and rotated colour space are shown in Figs. 3b and 3c respectively. Fig. 3d shows the segmentation, which was performed with a threshold distance of 5σ between the knot points, and detected twelve regions. The grey levels of the regions correspond to the number of pixels they contain. The shirt region was successfully identified, allowing thresolding to be used to extract it, as shown in Fig. 4a. Figs. 4b to 4e show the results of thresholding for selected frames from the rest of the fall. During the course of the development of the algorithm, the ability to extract the shirt region from these images was used as a system-level measure of its performance. It was found that, in order to successfully perform these segmentations, the use of a normalised colour space was needed to remove illumination effects from the images; that consideration of the error propagation was needed in order to map the data density in such a way to prevent either over- or under-segmentation; and that the Voronoid checking step was vital in order to prevent the hill climibing from linking together discriminable peaks in the density function. The final algorithm, incorporating these three features, was used successfully to calculate velocities in each of 780 frames.

Finally, the algorithm was applied to two standard images collected from the University of Southern California Signal and Image Processing Institute web site [1]. Fig. 5 shows an image of jellybeans. and its segmentation with a threshold of 5σ. The segmentation detected 24 regions: 2 covering the background, and 4 covering the different colours of jellybeans. The remaining 18 covered minority regions in the image, consisting of shadows and inter-reflections at the boundaries of objects. Figs. 6a-d show the results of thresholding to extract the four different colours of jellybeans. Fig. 7 shows an image of a house and its segmentation with a threshold distance of 2σ, detecting eleven regions: sky; walls; shadowed regions of the walls; roof; and white regions (windowsills, guttering etc.). The remaining six again accounted for minority regions in the image. Fig. 7e shows the result of thresholding to extract the white regions of the image.

It is clear from both the jellybeans and house images that the new segmentation rotuine is able to extract semantically meaningful regions from images. The colour normalisation and the ability of the algorithm to link extended but continuous clusters in feature space allow the algorithm to cope with illumination changes and inter-reflections respectively. In the case of the SIPI images, colour correction had been applied at the time the original photographs were scanned, equivalent to offsetting the origin of the colour space. The normalisation was therefore unable to completely remove illumination effects from the images, accounting for the minority regions in the segmentation result and the appearance of the shadowed areas of the walls in the house image as a separate region. The jellybeans images also illustrates that, as expected, the algorithm is unable to cope with highlights, since saturation in these regions destroys the colour information.

4 Conclusions

This paper has described a novel colour segmentation algorithm, which operates by clustering pixels in colour space using non-parametric denisty estimation followed by hill-climbing. The pixels from the original images are clustered onto local peaks, which are then labelled, and an image of these labels is returned as output. The traditional problem of selecting the correct scale to generate a segmentation showing all of the salient peaks, neither artificially splitting peaks due to noise artefacts, nor artificially merging statistically significant peaks due to selecting too high a scale, has been addressed using a statistically motivated approach. This consisted of mapping the space at a resolution determined using the statistical properites of the underlying images. This ensured that localised peaks at small scales due to noise were smoothed out, without destroying significant peaks through oversmoothing. Nevertheless, the algorithm includes the ability to apply additional blurring to the feature space, and thus generate a bifurction diagram showing the significance of peaks detected at various scales. An example of the successful segmentation of the most challenging synthetic data set that the algorithm could be expected to cope with has been given. Further examples showing the performance of the algorithm on colour video images showing simulated falls have also been given. This task has been used as the evaluation criterion for the development of the algorithm and we believe that a system-level measure of segmentation performance is perhaps the only objective definition. The algorithm successfully extracted the pixels in the actresses shirt,

[1] http://sipi.usc.edu/services/database/Database.html

allowing the centroid of the shirt and thus the velocity of the actress to be calculated. Examples of the application of the algorithm to some standard images have also been given, showing its ability to extract semantically meaningful regions from images.

The algorithm has been designed to work with an arbitrary number of dimensions, and several 2D examples have been given. It could also be applied to higher dimensional problems consisting of a stack of medical images of the same region generated using different techniques e.g. MRI scans with varying modalities, PET and CT scans etc. Each of these techniques has various strengths and weaknesses in terms of their ability to differentiate between different tissues. The segmentation algorithm could combine the information content of each image in a statistically rigorous way to achieve near-optimal tissue classification, maximising the information content in the output image. The method could also be applied to texture segmentation and recognition, using suitable texture measures in the place of the original pixel values to generate the feature space.

Acknowledgements

The authors would like to acknowledge the support of the MEDLINK programme, grant no. P169, in funding part of this work. All software is freely available from the TINA website www.niac.man.ac.uk/Tina.

References

[1] P.A. Bromiley, P. Courtney and N.A. Thacker. *A Case Study in the use of ROC Curves for Algorithm Design.* In Proc. BMVC 2001, BMVA, 2001.

[2] Y. Fang and T. Tan, *A Novel Adaptive Colour Segmentation Algorithm and its Application to Skin Detection.* In Proc. BMVC 2000, pp. 23-31. BMVA, 2000.

[3] G.D. Finlayson and G.Y. Tian. *Colour Normalisation for Colour Object Recognition.* International Journal of Pattern Recognition and Artificial Intelligence, pp. 1271-1285, 1999.

[4] T. Gevers, S. Ghebreab, and A.W.M. Smeulders. *Colour Invariant Snakes.* In Proc. BMVC 1998, pp. 578-588. BMVA, 1998.

[5] T. Gevers, A.W.M. Smeulders and H. Stokman. *Photometric Invariant Region Detection.* In Proc. BMVC 1998, BMVA, 1998.

[6] G.J. Klinker, A. Shafer, and T. Kanada. *A Physical Approach to Colour Image Understanding.* International Journal of Computer Vision, **4**, pp. 7-38, 1990.

[7] E. Pauwles and G. Frederix. *Non-Parametric Clustering for Image Segmentation and Grouping.* Computer Vision and Image Understanding, **75** nos. 1/2, pp. 73-85, 1999.

[8] G. Rees and P. Greenway. *Metrics for Image Segmentation.* In ICVS Performance Characterisation Workshop 1999, pp. 20-37, BMVA, 1999.

[9] W. Skarbek and A. Koschan. *Colour Image Segmentation: A Survey.* Technischer Bericht 94-32, Technical University of Berlin, 1994.

[10] N.A. Thacker, I. Abraham and P. Courtney. *Supervised Learning Extensions to the CLAM Network.* Neural Networks, **10** no. 2, pp 315-326, 1997.

[11] G. Wyszecki and W.S. Stiles. *Color Science: Concepts and Methods, Quantitative Data and Formulae* (2nd Edition) John Wiley and Sons, New York, 1982.

5 Figures

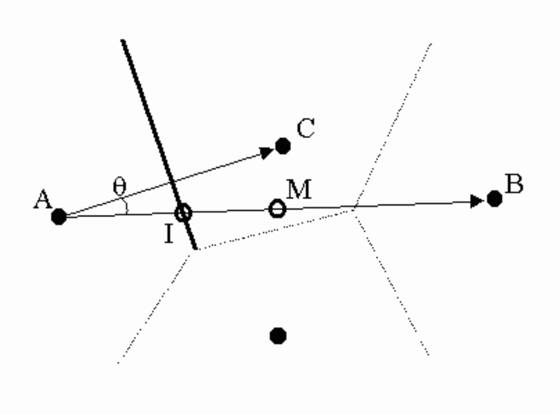
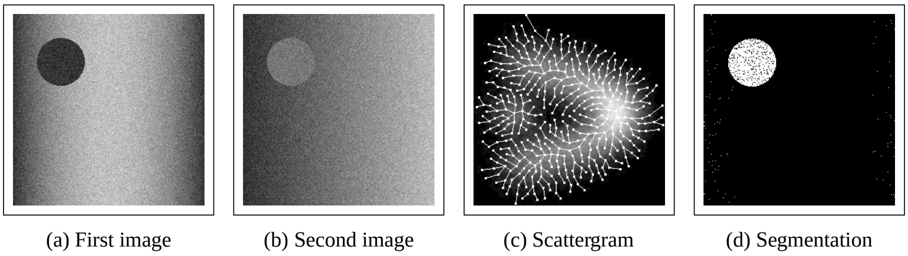


Figure 1: The derivation of the Voronoid cell crossing constraint for the hill climbing. The solid points are knot points and the dashed lines are the cell boundaries. The darker line is the boundary crossed by the allowed step **AC**. The constraint tests that the intersection I of this plane with the step **AB** is at less than half the length of **AB**, to the mid-point M.



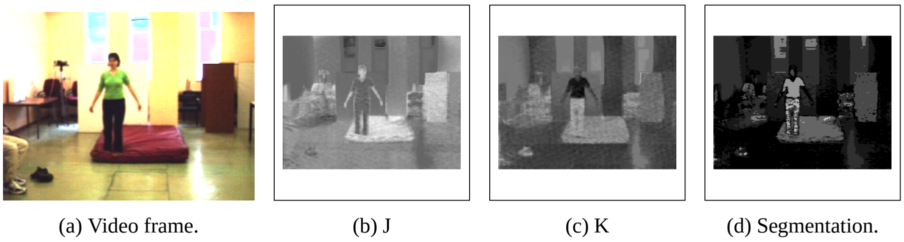
(a) First image

(b) Second image

(c) Scattergram

(d) Segmentation

Figure 2: The “horseshoe” synthetic data (a, b), scattergram (c), showing the map of knot points produced by the algorithm with a threshold distance of 0.5σ, and segmentation (d).



(a) Video frame.

(b) J

(c) K

(d) Segmentation.

Figure 3: Segmentation of a frame of colour video.

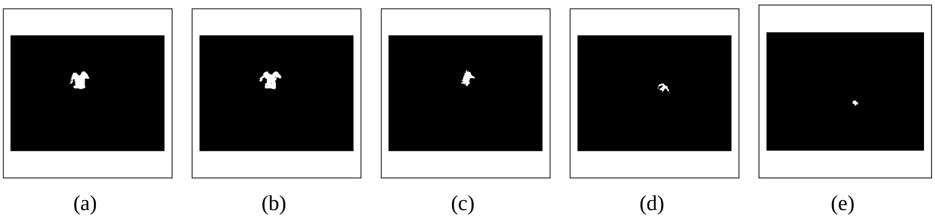

(a) (b) (c) (d) (e)

Figure 4: Results of thresholding the segmented colour video sequence to extract the shirt region. The result for the frame used in Fig. 4 is given in (a), and (b) to (e) show selected frames from the rest of the fall. The actress is falling away from the camera, and so the shirt region gradually becomes obstructed by her lower body.

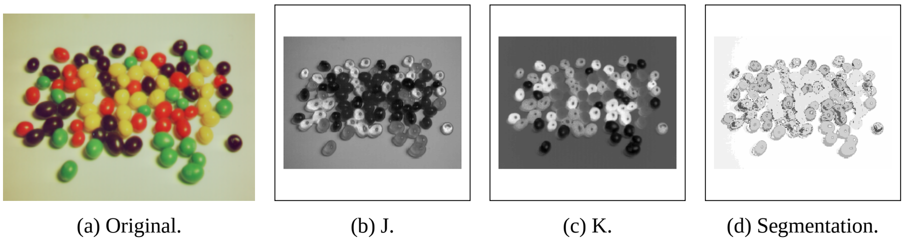

(a) Original. (b) J. (c) K. (d) Segmentation.

Figure 5: The jellybeans image.

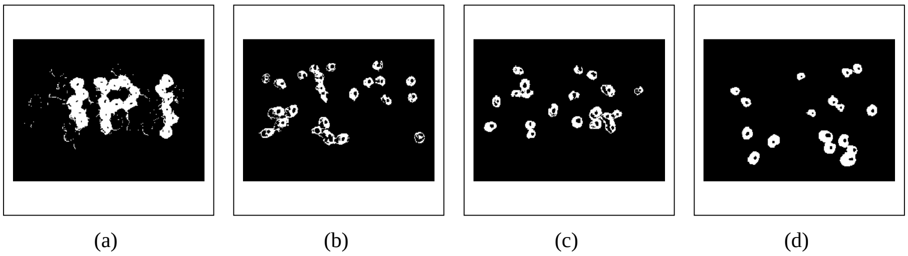

(a) (b) (c) (d)

Figure 6: Result of thresholding the jellybeans image segmentation to extract the yellow (a), black (b), red (c), and green (d) jellybeans.

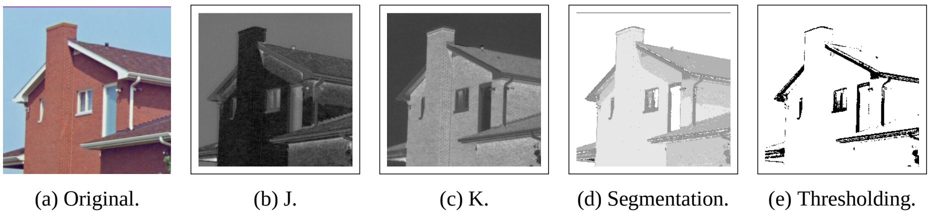

(a) Original. (b) J. (c) K. (d) Segmentation. (e) Thresholding.

Figure 7: The house image, and the result of thresholding the segmentation to extract the white regions.