

Attention Based Auto Image Cropping

Fred Stentiford

UCL Adastral Park Campus, Martlesham Heath, Ipswich, Suffolk, IP5 3RE, UK
f.stentiford@adastral.ucl.ac.uk

Abstract. Many images contain salient regions that are surrounded by too much uninteresting background material and are not as enlightening as a sensibly cropped version. The choice of the best picture window both at capture time and during subsequent processing is normally subjective and a wholly manual task. This paper proposes a method of automatically cropping visual material based upon a new measure of visual attention that reflects the informativeness of the image.

1 Introduction

Hugh quantities of digital images and video are being created which are not informative simply because the interesting parts are not immediately apparent to a human observer. This may be because there is an overwhelming amount of distracting background material or because the images are being viewed on a small display. In either case there is a need to identify those regions in an image which are salient and to reshape the image for maximum impact and improved composition without tedious manual involvement.

This problem has been addressed by Chen et al [1] for viewing large images on small displays. Itti's ([2] 1998) visual attention measure was used in combination with models for face and text detection to measure the saliency of Regions Of Interest (ROI) for inclusion in a cropped version of the image. This work was extended by Liu et al [3] to include variations in time with application to determining optimal browsing paths across the ROIs. However, different weights were necessary for fusing the different models into a single attention value for different categories of images. Zhang [4] used a cropping technique that made use of face detection and an attention model. The approach appeared to work well but employed many empirically determined parameters. Digital camera images were first segmented by Ma et al [5] and then ROIs were assessed according to their entropy, the size and nearness to the centre of the image. Again empirical parameters needed to be set according to the characteristics of the camera and the habits of the user. Boutell [6] has shown that creating more training data simply by blindly cropping 10% off sides of images significantly boosted scores on test sets. This provided indirect evidence that cropped images can be more informative.

Suh et al [7] summed saliency values within candidate cropped regions to determine the best selections. This approach also used Itti's measure of saliency and strategies for face detection, but required a threshold to limit the size of the cropped

region that varied from image to image. The approach taken in this paper makes use of a measure of visual attention that was successfully applied to the control of the focusing of image forming devices [8]. It was found that the subject was in focus when the average saliency score was maximised over the whole image. This measure could be considered as an indication of image “informativeness” and is applied here to the related problem of selecting an optimal cropping window.

2 Visual Attention

Itti [2] makes use of colour, intensity and orientation filters followed by centre-surround computations to determine saliency. Local contrast is employed by Ma et al [9] and Hu et al [10] together with fuzzy growing to identify ROIs. In this paper salient regions are detected through a process that compares small regions with others within the image. A region that does not match most other regions in the image is very likely to be anomalous and will stand out as foreground material. For example, the edges of large objects and the whole of small objects normally attract high attention scores mainly because of colour adjacencies or textures that only occur rarely in the image. Self similar backgrounds that display a translational symmetry are assigned low attention scores.

Region matching requires a few pixels (a fork) within that region to match in a translated position in another region. If the difference in colour of one pixel pair exceeds a certain threshold a mismatch is counted and the attention score is incremented.

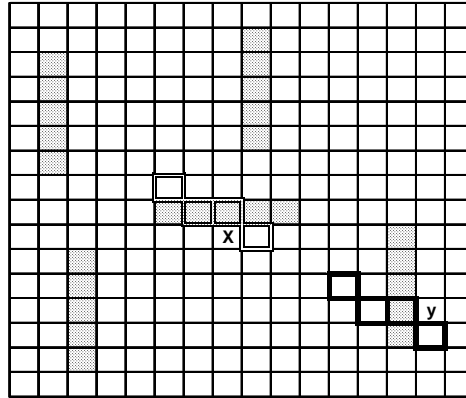


Fig. 1. Fork at x mismatching at y .

Let a pixel x in a pattern correspond to colour components a where

$$\mathbf{x} = (x_1, x_2) \quad \text{and} \quad \mathbf{a} = (a_1, a_2, a_3) \quad (1)$$

Let $F(x) = a$

Consider a neighbourhood N of x where

$$\{\mathbf{x}' \in N \text{ iff } |x_i - x'_i| \leq \epsilon_i \quad \forall i\}$$

Select a fork S_A of m random pixels where

$$S_A = \{\mathbf{x}'_1, \mathbf{x}'_2, \mathbf{x}'_3, \dots, \mathbf{x}'_m\}. \quad (2)$$

Construct a fork of m pixels S_B where

$$S_B = \{\mathbf{y}_1, \mathbf{y}_2, \mathbf{y}_3, \dots, \mathbf{y}_m\} \quad (3)$$

$$\text{and } \mathbf{x}_i - \mathbf{y}_i = \boldsymbol{\delta}_i$$

S_B is a translated version of S_A . The fork S_A mismatches the fork S_B if

$$|F_j(\mathbf{x}_i) - F_j(\mathbf{y}_i)| \geq \tau_j \quad \text{for any } i, j \quad (4)$$

In general the colour component threshold τ_j is not a constant and will be dependent upon the colour space of the measurements under comparison i.e.

$$\tau_j = f_j(\mathbf{F}(\mathbf{x}), \mathbf{F}(\mathbf{y})) \quad (5)$$

A location \mathbf{x} will be worthy of attention if a sequence of t forks matches only a few other neighbourhoods in the space. In Fig. 1 a fork with $m = 3$ pixels is selected in the neighbourhood of a pixel \mathbf{x} and is shown mismatching at \mathbf{y} . The neighbourhood of the second pixel \mathbf{y} matches the first if the colour intensities of all the corresponding pixels have values within τ of each other. The attention score $V(\mathbf{x})$ for each pixel \mathbf{x} is incremented each time a mismatch occurs in the fork comparisons with a sequence of \mathbf{y} .

This measure of visual attention has been applied to image compression Stentiford [11] and focus control Shilston [12]. It was found that the maximum M_I of the sum of the attention scores of all the pixels in an image I at different focal distances tended to occur when the principal subject is in focus. M_I may be considered to be a measure of image informativeness.

$$M_I = \sum_{\mathbf{x} \in I} V(\mathbf{x}) \quad (6)$$

This paper makes use of the same measure to assess the informativeness of cropped images with a view to selecting the best version. In this case the optimum cropping window W_I is taken to be the one having the highest average pixel attention score.

$$W_I = \arg \max_{W \in I} \sum_{\mathbf{x} \in W} V(\mathbf{x}) / \|W\| \quad (7)$$

Parameter values used in the experiments below were $t = 100, m = 3, \varepsilon = 1, \tau = 40$.

3 Results

3.1 Image Cropping

The approach begins by generating a saliency map $V(\mathbf{x})$ and then searching for the window which possesses the highest average pixel attention score according to



Fig. 2. Original image and saliency map

equation (7). Fig. 2 shows an image together with its saliency map in which the brighter colours indicate higher saliency.

The window W may be any shape or size, but for many image forming devices a fixed aspect ratio may be required and the method is illustrated here with windows restricted to a zoom factor of 2. Fig. 3 shows a series of images together with their cropped versions.

3.2 Image Zooming

The results described above open the possibility of an optimal zoom window for any view using the informativeness measure in (7). However, this would mean that all possible zoomed windows for each scene would need to be evaluated which is neither practical nor probably necessary. The approach was applied to the problem of selecting the most informative sub-window with the same aspect ratio within a static image for a range of sub-window sizes. This is equivalent to extracting the location of the most informative crop at each size. Fig. 4 shows two original images I and a succession of reducing windows W at increasing zoom factors selected according to equation (7).

3.3 Zoom Factor

The informativeness of an image of a distant object should increase as the object comes closer and then decrease again when perhaps a featureless surface of the object occupies a large proportion of the image. A series of 320x240 photos in Fig. 5 were taken of a red rectangle at various focal distances. The average pixel attention score $\sum_{x \in W} V(x) / \|W\|$ is plotted in Fig. 6. A peak is obtained when the rectangle enlarged but with background still present and edges and corners clearly separated from the image boundary.



Fig. 3. Original images and cropped versions

6 Fred Stentiford

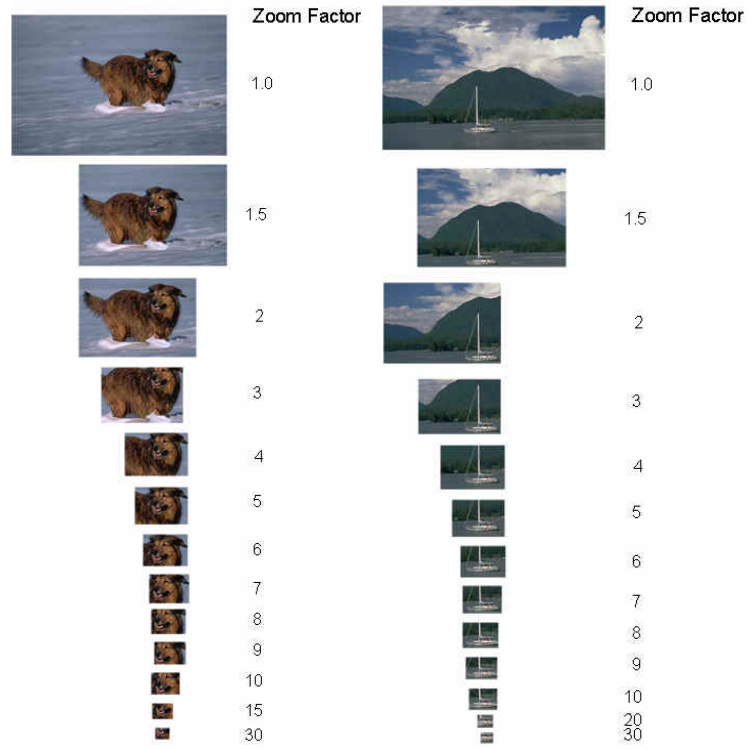


Fig. 4. Optimal cropping windows at various sizes.. a) dog, b) boat

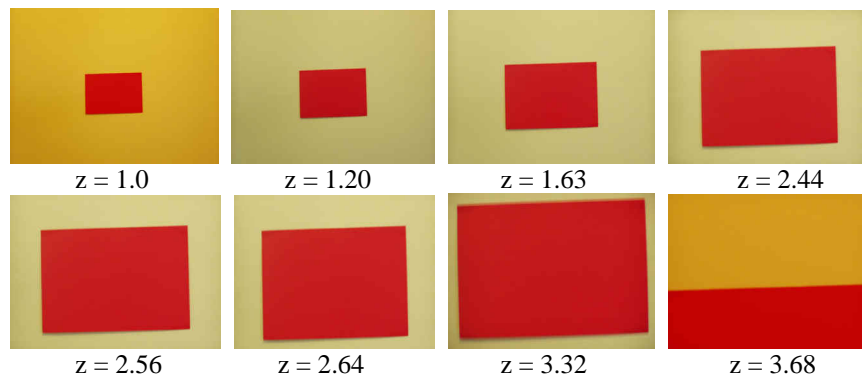


Fig. 5. Images of a rectangle taken at different zoom factors z .

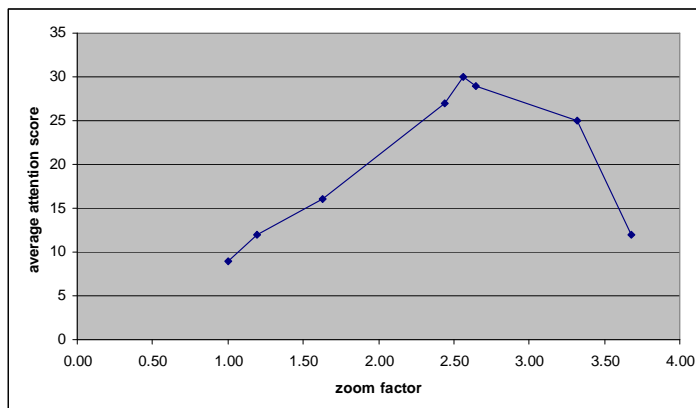


Fig. 6. Average pixel attention scores for images in Fig. 5.

Closer views of the object may expose more attentive detail and a sharp maximum would not be reached. Fig. 7 shows a wide angle view of a car and three successive zoomed x3 images (300x225) controlled by the attention measure in (7). In this example the new windows are derived from the previous image which possesses greater resolution in the region of interest than the original. The final image is a cropped version of the previous image. Zooming only on the original image does not produce the same results owing to the presence of small salient regions not relevant to the principal subject.



Fig. 7. Zooming x3 based upon attention measure.

4 Discussion

The cropping results in Fig. 2 are satisfactory for a zoom factor of 2. However, an acceptable cropped image may not exist for all window sizes and therefore some judgment is necessary for selection.

The succession of windows in Fig. 4 demonstrate a convergence towards the regions of highest attention value. The window at zoom factor 3 in Fig. 4a is not well cropped because the single window cannot encompass both the dog's head and the salient area of snow at the bottom of the image. This may be handled in future by branching and allowing more than one window to be formed at the same zoom factor. This would be a natural approach to images that contained more than one salient object.

Although this mechanism may be applied to picture framing in cameras, the results in Fig. 4 are derived from a single image and no additional definition is obtained in the smaller windows. In the case of images in Fig. 5 the resolution increases but no new detail is exposed and the saliency of the series of images passes a peak. On the other hand the images in Fig. 7 show that the additional resolution obtained by zooming reveals new salient material that influences the window convergence path.

Computation requirements can be quite high especially if every pixel region in the image is scored. However, the attention score only requires simple operations of matching and counting and may be carried out in parallel across the whole image. The calculations for controlling the focusing of a video camera have been implemented on a Texas Instruments DM642 DSP platform that operates at 25 fps.

As the informativeness measure in (7) is identical to that used to select the best focal distance [12] it is conceivable that the operations of focusing and cropping could take place at the same time. The focusing would naturally concentrate on the most salient region within each window. Objects that were blurred, either due to motion, camera shake, or the wrong focal distance, would tend to discourage focusing and cropping in those regions, but subjects that were at the correct focal distance and not moving relative to the camera would be targeted by the mechanism.

5 Conclusions

This paper has described the application of an attention measure to the automatic cropping of images. The method may be applied to cropping single images or guiding a series of zoom operations. More subjective evaluation is necessary on a greater range of images and strategies developed for identifying multiple cropping windows. These experiments could also compare human selected crops with those obtained by the model.

This research has been conducted within the framework of the European Commission funded Network of Excellence "Multimedia Understanding through Semantics, Computation and Learning" (MUSCLE) [11].

References

1. Chen, L., Xie, X., Fan, X., Ma, W., Zhang, H., Zhou, H. : A Visual Attention Model for Adapting Images on Small Displays. *Multimedia Systems*. **9** (2003) 353-364
2. Itti, L., Koch, C., Niebur, E.: A Model of Saliency-Based Visual Attention for Rapid Scene Analysis. *IEEE Trans Pattern Analysis Mach Intell*. **20** (1998) 1254-1259
3. Liu, H., Xie, X., Ma, W., Zhang, H. : Automatic Browsing of Large Pictures on Mobile Devices. *11th ACM Int. Conf. on Multimedia*. November (2003)
4. Zhang, M., Zhang, L., Sun, Y., Feng, L., Ma, W.: Auto Cropping for Digital Photographs. *IEEE Conf. on Multimedia and Expo*. July (2005)
5. Ma, M., Guo, J.K.: Automatic Image Cropping for Mobile Devices with Built-in Camera. *Proc. Consumer Communication & Networking Conf*. January (2004) 710-711
6. Boutell, M., Luo, J., Gray, R.T.: Sunset Scene Classification using Simulated Image Recomposition. *Proc. IEEE Conf. on Multimedia and Expo*. 1 (2003) 37-40
7. Suh, B., Ling, H., Bederson B.B., Jacobs, D.W.: Automatic Thumbnail Cropping and its Effectiveness. *Proc. 16th ACM Symposium on User Interface Software and Technology*. (2003) 95-104
8. Shilston, R., Stentiford, F.W.M.: An Attention-Based Focus Control System," *International Conf. on Image Processing*. October (2006)
9. Ma, Y., Zhang, H.: Contrast-Based Image Attention Analysis by Using Fuzzy Growing. *Proc.11th ACM Conf. on Multimedia*. (2003) 347-381
10. Hu, Y., Xie, X., Chen, Z., Ma, W. : Attention Model Based Progressive Image Transmission. *Proc. IEEE Int. Conf. On Multimedia and Expo*. (2004) 1079-1082
11. *Multimedia Understanding through Semantics, Computation and Learning*, Network of Excellence. EC 6th Framework Programme. FP6-507752. <http://www.muscle-noe.org/>