How interesting is this? Finding interest hotspots and ranking images using an MPEG-7 visual attention model

Heiko Wolf & Da Deng

Department of Information Science
University of Otago. Dunedin, New Zealand
Email: {hwolf1, ddeng}@infoscience.otago.ac.nz

Submitted to SIRC 2005 - The 17th Annual Colloquium of the Spatial Information Research Centre
University of Otago, Dunedin, New Zealand
November 24th-25th 2005

ABSTRACT

A lively Dunedin street scene, and a panoramic view of the Southern Alps - two images that might appeal and interest a viewer. But where do people look, and which of those images appears more interesting? In this paper, we are introducing a visual attention model based on MPEG-7 descriptors that creates multi-scale feature maps to detect interest hotspots in images. Further, we are assessing three methods that use attention models for image ranking and compare them to results gathered in a user test. Preliminary results indicate that rankings created by our model show a high agreement with rankings obtained in a pilot user study.

Keywords and phrases: Visual attention model, visual saliency, scene analysis, image ranking, MPEG-7

1 INTRODUCTION

The automatic detection of interesting areas within an image is important to a range of scene analysis applications, such as landmark detection, traffic and road sign recognition or extraction of scene descriptions. Furthermore, it allows to efficiently preprocess an image and filter the most important areas for further processing. But how can we define which areas will be perceived as "interesting" by humans, and automatically detect these regions?

One possible approach is to simulate charateristics of human vision. An area that has gained a lot of research interest is the modelling of visual attention in early vision. Itti & Koch (2001) suggest that "subjects selectively direct attention to objects in a scene using both bottom-up, image-based saliency cues and top-down, task-dependent cues" (Itti & Koch 2001, p.194). The first analysis of images is based on the "pre-attentive computation of early visual features" (Itti & Koch 2001, p.196). This concept has been used in rapid scene analysis (Itti et al. 1998) and automatic video summarisation (Ma et al. 2002).

In this paper, we are describing a visual attention model for image analysis built on MPEG-7 feature descriptors. We are also extending the concept of salient areas to calculate a global "interest value" for an image to achieve rankings according to the interest of the content. Such image rankings will be useful to organise large image collections and web search result sets, or to prioritize image data for further analysis or processing. The paper is organised in six sections. Section 2 summarises other research in the field of image saliency, briefly covers the concept of visual attention and gives an overview of the MPEG-7 visual feature descriptors. Section 3 introduces our visual attention model, before we describe the implementation of our system in Section 4. Section 5 shows some preliminary results and Section 6 summarises this paper. The main contributions of this paper are the MPEG-7 based visual attention model, a novel approach to image ranking based on this model and an evaluation of three ranking methods in a user test.

2 RELATED WORK

2.1 Detection of interesting regions

The extraction of "interesting regions" using saliency has been used to model characteristics of early human vision and can help to derive information about the interest of content within an image. This is especially valuable as early vision processing is context independent and therefore allows to derive image content descriptions without domain knowledge. Kadir & Brady (2001) use the local entropy of grey-level values in an image to calculate the most salient regions. Ferreira & Borges (2004) propose using directional features extracted by Gabor filtering to find the most significant directions and select salient regions according to them. They extend the saliency concept by ranking the salient regions, an approach that has been used in image compression. Celaya & Jiménez (2003) compute salient regions in an image using the Euclidean distance between the RGB values of a pixel and its neighbourhood to find regions that deserve further processing. They use the salient regions to guide a robot's vision to interesting objects in its field of view. Colour contrasts in the LUV colour space have been used to create a saliency map as well (Ma & Zhang 2003).

Important work on modelling visual attention as a neurobiological concept has been published by Itti & Koch (2001). They propose a framework incorporating early vision features such as intensity contrast, colour contrast, orientation differences, and direction of motion. They also hypothesise that these features are integrated into one saliency map, in which the combination of features determines the points that draw the most attention. Itti et al. (1998) implemented a system based on the early visual features that can determine which points in an image are attended in which order. They find that this model can identify "informative" regions even when strong noise is introduced.

2.2 Image ranking

Apart from finding interesting regions, the calculated saliency of an image can also be used to rank the interestingness of images. There is little research on automatic image ranking or prioritisation based on image saliency that we know of. NASA conducted a study to validate the prioritisation of Mars Rover images (Castano et al. 2005). As there is only a limited bandwidth to send images from other planets back to Earth, an automatic ranking of images can help to send back the images with the highest scientific value. However, in Castano et al.'s (2005) study the ranking criterium is a "scientific value" which is exactly defined and judged by experts, while we are ranking general images by their preattentive interestingness.

An important area for image ranking lies within extensive image collections, with the World Wide Web being the largest. In previous work, we used MPEG-7 feature descriptors to organise results returned from contemporary image search engines according to content similarity (Deng & Wolf 2005). Liu et al. (2004) use an attention model consisting of saliency, face detection and query-dependent attention objects to crop images to the most interesting regions and then rank them according to the similarity of those regions.

Finally, image ranking based on a visual attention model can be used to select interesting frames in video. The method proposed in this paper is part of a research effort towards a system for automatic video summary (Wolf 2005). The visual attention model proposed by Itti & Koch (2001) has been used as part of a system for video summarisation (Ma et al. 2002).

2.3 MPEG-7 Visual Feature Descriptors

MPEG-7 is a standard that provides a representation of multimedia metadata in XML (Martinez 2004). Consisting of seven main parts, it also includes a set of well-tested visual feature descriptors. A superb overview of the MPEG-7 standard, its goals and applications as well as the visual descriptors, can be found in the standard book by Manjunath et al. (2002). MPEG-7 descriptors have been successfully used in image retrieval systems (Koskela et al. 2001). Saberi et al. (2005) extend the clustering-based video browsing system introduced by Koprinska et al. (2004) and report that the use of MPEG-7 descriptors results in a better hierarchical representation of the video content than the use of colour histograms.

3 THE VISUAL ATTENTION MODEL

The visual attention model aims to describe the attention or interest the content of an image produces. Possible uses for a static attention model are the detection of interesting regions in images as described by Itti et al. (1998) or the ranking of images in query result sets from image databases or image search engines. The following sections will describe the feature extraction, feature map amplification and combination in our attention model.

3.1 Feature extraction

According to Itti & Koch (2001), human bottom-up attention is guided by early visual features such as intensity contrast, colour opponency, orientation, and direction and velocity of motion[1]. They also note that it is not the feature characteristics themselves but the difference between a feature region and its neighbourhood that generates

[1] In the design of this static visual attention model, we ignore direction and velocity of motion which are integrated in the video summary system currently under development (Wolf 2005).

attention. The MPEG-7 visual feature descriptors include multiple descriptors for both colour and orientation features that have been rigorously tested in the standardisation process. However, there is no feature descriptor that specifically describes intensities, so we introduce our own intensity histogram descriptor as part of the static visual attention model. Differing from Itti et al.'s (1998) work, who calculate their features per pixel, we will make use of the properties of the MPEG-7 feature descriptors that are calculated from images or image regions and base our feature maps on differences between image regions.

For each feature descriptor, the saliency of a region is defined as the average distance of a region to its neighbouring regions. In our attention model, regions have a rectangular shape and the neighbourhood of a region is defined as the four regions sharing an edge with the current region plus the four regions sharing only a corner with the current region. To capture contrast on different scales within the image, we are applying a multi-scale approach. For each scale, the image is divided into regions of different size. In our implemenation, we start at a region size of 8x8 pixel and create smaller scales by enlarging the region size by a factor of two. The scaling process stops if the next scale would contain less than 8 regions in x or y direction.

The average distance of the current region $D_{avg_{R_c}}$ is defined as the sum of distances D divided by the number of neighbouring regions N as shown in Equation 1. The distance measure D is depending on the used feature descriptor and will be defined for each of them specifically in the according section. Each distance is calculated between the feature descriptor of the current region FD_{R_c} and the feature descriptor of a neighbouring region FD_{R_i}, $(i = 1, 2, \ldots, N)$.

$$D_{avg_{R_c}} = \frac{\sum_{i=1}^{N} D\{FD_{R_c} - FD_{R_i}\}}{N} \tag{1}$$

3.1.1 Colour features - Scalable Colour Descriptor

The colour descriptor used in our model, the SCD, is a colour histogram in the HSV colour space that is normalised and encoded by a Haar transform. Finally, adjacent bins are summed up to create a 128-bin histogram. Detailed information about the extraction process and the distance computation between two Scalable Colour Descriptors as defined in the MPEG-7 standard can be obtained from (Manjunath et al. 2002, pp.198–201) which also includes a schematic diagram of the SCD generation.

As described in Equation 1, the average distance is computed between each region's SCD and the SCD of regions in the neighbourhood. For this process, the image is divided into rectangular regions on each scale. The SCD feature maps of each scale are then amplified using the process described in Section 3.2.

3.1.2 Orientation features - Edge Histogram Descriptor

The second group of early vision features used in bottom-up attention calculation according to Itti & Koch (2001) are differences of orientations between a region and its neighbourhood. Orientations within images are observable as edges and textures. In this implementation, we use the MPEG-7 Edge Histogram Descriptor (EHD).

The EHD describes the local edge distribution within an image or image region. It detects non-directional edges as well as four directional edge categories (vertical, horizontal, 45° and 90°). To achieve information about localised edge distribution, each input image or region is divided into 4x4 subimages. For each subimage, edges that fall in one of the five categories above are counted into five bins, which are then normalised by the total number of edge and non-edge pixels within the subimage. Also, one global and 13 semiglobal edge histograms are calculated from the local histograms to capture global edge distribution as well. For more information on the creation of subimages and detection of edges, refer to (Manjunath et al. 2002, pp.223–224).

To calculate the EHD feature map, we apply Equation 1 to the EHD of each region on all scales. The distance measure between two Edge Histogram Descriptors is shown in Equation 2 which takes into account the bin values for the local edge histograms $h_A(i)$ and $h_B(i)$, the global edge histograms $h_A^g(i)$ and $h_B^g(i)$ and the semiglobal edge histograms $h_A^S(i)$ and $h_B^S(i)$ which are all calculated from region A and B, respectively. The indices equal the number of bins, which means there are 80 bins locally (16 subimages × 5 edge types), 5 bins globally and 65 bins semiglobally (13 semiglobal groupings of subimages × 5 edge types). To equalise weights, the global histogram distance is multiplied by a factor of 5, resulting in the following distance metrics:

$$D(A, B) = \sum_{i=0}^{79} |h_A(i) - h_B(i)| + 5 \times \sum_{i=0}^{4} |h_A^g(i) - h_B^g(i)| + \sum_{i=0}^{64} |h_A^S(i) - h_B^S(i)| \tag{2}$$

3.1.3 Intensity features - Intensity Histogram Descriptor

Intensity features are the third feature group used for botton-up attention computation (Itti & Koch 2001). As the MPEG-7 visual feature descriptors do not include intensity descriptors, we are defining our own intensity histogram descriptor (IHD) to use in the visual attention model. Intensity differences can be calculated based on different features such as luminance or brightness. For our IHD we define intensity as the pixel value of the grey-scale transform of an input image.

A grey-scale image is divided into regions for each scale and the IHD histogram is calculated for each region. The intensity histogram consists of 16 bins that represent equal parts of the grey-scale value space between 0 and 255. For each region, the pixels are sorted into the according bins and then normalised by the total number of pixels in this region, so that each IHD bin describes the percentage of grey values in this scale within the region. The calculation of the intensity histogram is shown in Equation 3 with bin describing the current bin which is chosen according to the current pixel's grey value P_{Grey} by integer division ($bin = P_{Grey}/16$), and with the index i running from 1 to N image pixels.

$$IHD(bin) = \sum_{i=1}^{N} P_{Grey}(i), \quad (bin = P_{Grey}/16) \tag{3}$$

The IHD feature map is created by calculating the average distance between the intensity histogram descriptors of neighbouring regions according to Equation 1. The distance measure between intensity histograms IHD_A and IHD_B is, similar to the distance measure for the EHD, the city block distance as described in Equation 4.

$$D(IHD_A, IHD_B) = \sum_{i=0}^{15} |IHD_A(i) - IHD_B(i)| \tag{4}$$

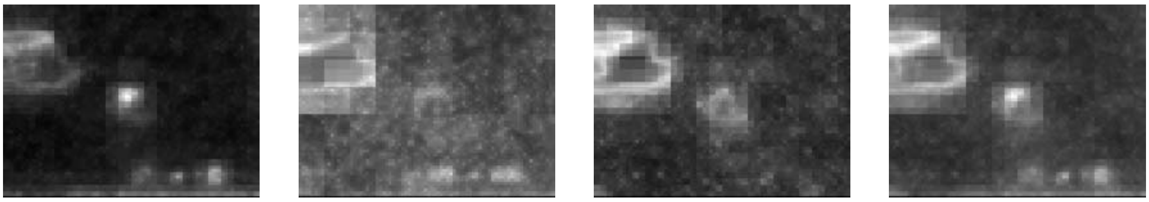
3.2 Feature map amplification

Early vision attention is triggered by feature contrast (Itti & Koch 2001). The stronger the contrast, the stronger an area of an image "pops out". As we are extracting multiple feature maps on different scales and for different feature descriptors, we will obtain different maxima for each feature map. To accurately simulate the popping out of areas of strong contrast, we want to promote feature maps with strong maxima. Itti & Koch (1999) compare four feature combination strategies, one of them being global non-linear normalisation with following summarisation. This strategy was described as computationally simple yet a good approximation to human saliency and is used in our implementation.

3.3 Feature combination

After obtaining the multiple-scale feature maps for colour, orientation, and intensiy features as described in the previous sections, these maps can now be combined to generate a saliency map to represent the attention that parts of this image trigger. This is in accordance with Itti & Koch (2001) who propose that "various feature maps feed into a unique 'saliency [...] map' " (Itti & Koch 2001, p.198).

First, the saliency maps from all scales are integrated into one map for each descriptor using the feature map amplification described in Section 3.2. After that, the three descriptor feature maps are amplified again and then summed to one global saliency map. In evaluating the created feature map, we can now find maxima to extract the most interesting spots from the image. An example of the different feature maps and their combination to one saliency map for the coke can image in Figure 5(a) can be found in Figure 1.



(a) SCD feature map

(b) EHD feature map

(c) IHD feature map

(d) Combined saliency map

Figure 1: Feature maps for "coke can" test image

3.4 Image ranking

An "interest value" based on image saliency can be used as a ranking criterium for image sets as proposed by Ma et al. (2002). They use a saliency map to calculate an attention value that takes position, size and brightness of salient regions into account. While this creates a single attention value, it leaves the problem that feature maps are normalised without reference to other images we want to compare the current image with. Also, Ma et al. (2002) use a Gaussian template to assign lesser weight to the outer regions of the image. Although it is generally accepted that humans perceive the center of the image as more important, there are applications in which the outer regions can bear just as much information, for example surveillance. In this study, we assume that all image regions are equally important.

As mentioned above, the normalisation process while integrating the different feature maps loses important information about how the attention values of multiple images or frames compare to each other. In order to create an interest value that represents the importance of an image relative to the other images within a given image set, we calculate a single feature value FV for each (not-normalised) feature map of the image by 1.) summing up the average regional distances and dividing them by the total number of regions and 2.) summing up the attention values for each scale and dividing them by the number of scales, as shown in Equation5. In this equations, $D(s)(i)$ is the average distance of region i to its neighbourhood for the respective feature descriptor and scale s, while i is running from 1 to N_s (the number of regions for scale s) and s is running from 1 to M (the number of scales).

$$FV_{SCD/EHD/IHD} = \frac{1}{M}\sum_{s=1}^{M}\frac{\sum_{i=1}^{N_s} D_{SCD/EHD/IHD}(s)(i)}{N_s} \tag{5}$$

After calculating single values for each feature channel, FV_{SCD}, FV_{EHD} and FV_{IHD}, respectively, we can now normalise each channel over the whole image set. Each channel is normalised between 0 and 1. The values 0 and 1 are assigned to the image with the lowest and highest feature value in the current feature channel, respectively, while all other feature values are normalised accordingly. This step allows us to combine the feature channels which use different value spaces but also to keep the original distance ratios between images. Finally, we are combining the single feature values to one global feature value FV_{global}. The combination is following Equation 6.

$$FV_{global} = \frac{FV_{SCD} + FV_{EHD} + FV_{IHD}}{3} \tag{6}$$

4 THE SYSTEM

The system detecting both spots of interest as well as ranking images has been implemented in C++ based on the the MPEG-7 eXperimentation model reference implementation (TU Munich 2005). A diagram of the system is shown in Figure 2. Image ranking is implemented in two ways, first using the global saliency values as described in Section 3.4 and second employing Ma et al.'s (2002) method using the combined saliency map.

5 PRELIMINARY RESULTS

The definition of interesting spots as well as the ranking of images according to their interest value is highly subjective. The ideal test bed therefore involves a large user study which we have not undertaken to date. However, some other tests brought preliminary results that are encouraging for further studies.

5.1 Evaluation of interesting spot detection

To assess the detected interesting spots, we compared our findings with the results of ezvision (Itti 2004), an implementation of Itti & Koch's (2001) model of visual attention. As this model has been validated by extensive user tests, we can use the regions of attention found by ezvision as a point of reference.

Our test set included 37 images, of which 13 were taken from the University of Otago (2005) library image series and 24 from the iLab Image Databases (2005). Of the 24 images, 6 images were part of the "autobahn", "coke" and "triangle" (Itti & Koch 2001) image sets, respectively, and 6 images were taken from the "outdoor" (Itti et al. 1998) image set. We calculated the first four "interesting spots" using our model and then had ezvision calculate two sets of results which we used as a reference for comparison, 1.) the first four attended regions and 2.) the first eight attended regions. We used the two sets to compare how many of our interesting spots agree with the first four attended regions and the first eight attended regions, respectively. An interesting spot was counted as agreeing with an attended region when they were describing similar regions.

The agreement was very variable over the image sets as shown in Table 1. Between 19% (Library1 test set) and 79% (traffic test set) of the interesting spots agreed with the first four attended regions, averaging at 46%. The

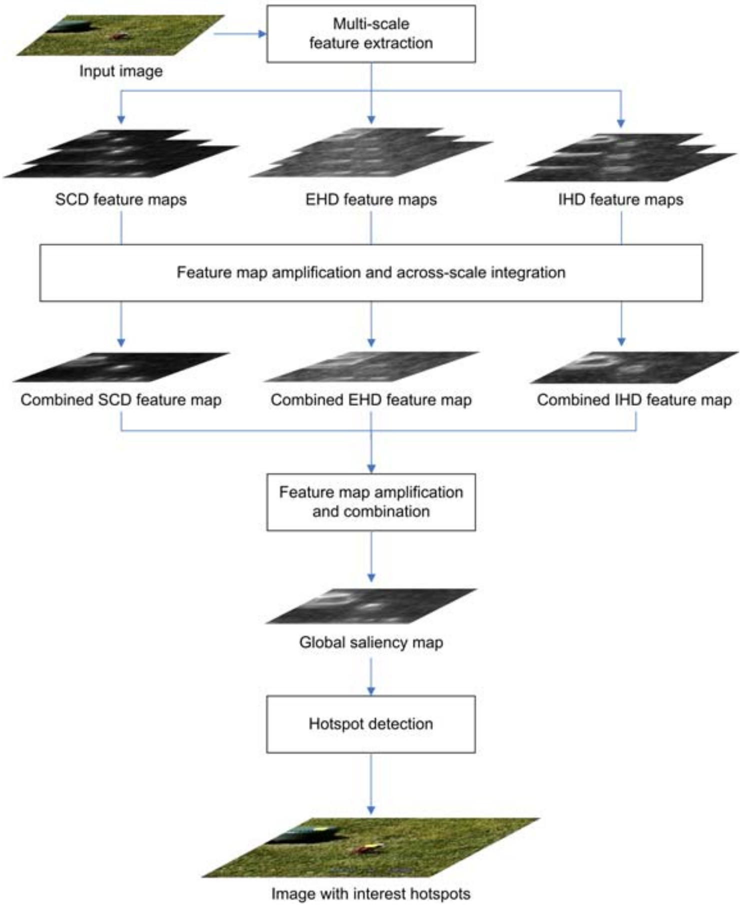


Figure 2: *System diagram*

average agreement rises to 54% when we compare the interesting spots with the first eight attended regions. The "library1" test case shows a significantely lower than average agreement of our interesting spots with the attended regions. This might be due to the complexity of these images. Figure 3 shows an example from the "library1" test set where the interesting spots show only little agreement with the attended regions calculated by ezvision. It can be seen though that we find significant spots, for example groups of people.

This was true in other cases as well where interesting spots were detected on significant points that had not been found by ezvision, for example our method found the red triangle in the test image in Figure 4 while it was missed by the ezvision method, which also did not visit the time stamp at the bottom of the image. These differences can be explained with the different methods of calculating the saliency map and the most interesting spots as well as the purpose of the application. While Itti et al. (1998) are calculating the attended regions in order of attendance, we are looking for the globally most interesting spots. We are not simulating inhibition of return for attended regions or alter the saliency when locating the next interesting spot. The differences between the two methods are visible especially in the cluttered images of the "library" test sets. When excluding "library1" and "library2", we achieve an average agreement of 67% when comparing with the first eight attended regions.

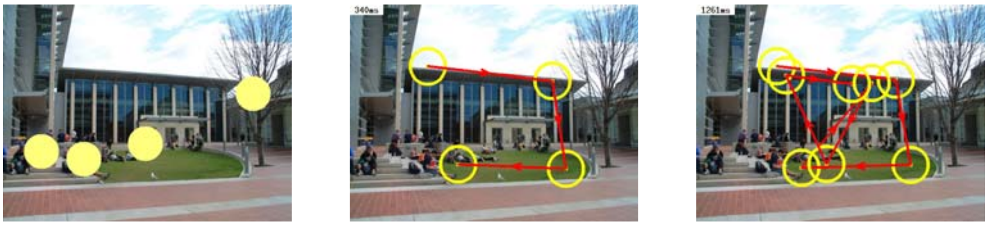
As the perception of the "interestingness" of an image is very subjective, it is hard to judge the efficiency of our method in locating interesting spots solely based on the scomparison to another method. The results indicate that, despite the difference in detected regions, our significant points are a plausible description of the scene, as more examples in Figure 5 show. This assumption should be tested in further experiments, preferably in a user study.

5.2 Evaluation of image ranking

In a second test case, we compared the image rankings derived from three different methods with image rankings done by users. The image test sets were the same as used to assess the interesting spots. We use the Spearman Rank Correlation Coefficient (Lehmann 1975) to calculate the agreement between two rankings of the same data set, in our case a set of images. Let $X = \{x_1, \ldots, x_n\}$ be a set of n images, then A_i and B_i are the

Image set	Comparision with 4 attended regions	Comparison with 8 attended regions
“coke can”	79%	79%
“library1”	19%	19%
“library2”	42%	56%
“outdour”	46%	71%
“traffic”	38%	54%
“triangle”	54%	67%
average	46%	58%
average without “library” image sets	54%	68%

Table 1: Agreement of interesting spots with attended regions calculated by ezvision



(a) Interest hotspots

(b) First four attended regions

(c) First eight attended regions

Figure 3: lib1d (“library1” test set)

rank of x_i according to ranking A and B, respectively. The Spearman Rank Correlation Coefficient is defined in Equation 7, where d is the rank difference between corresponding variables. The Spearman Rank Correlation Coefficient ranges from -1 to 1, with -1 indicating complete disagreement and 1 complete agreement between two rankings.

$$r = 1 - 6 \sum \frac{d^2}{N(N^2-1)} \tag{7}$$

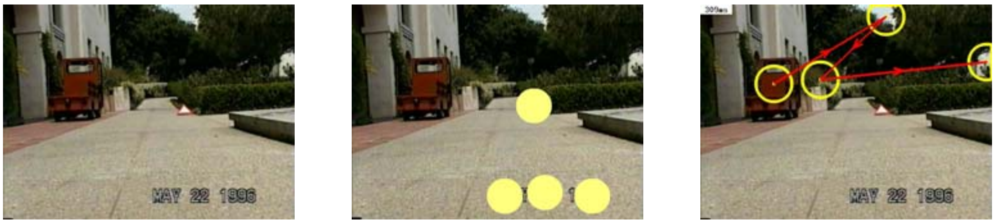
Three image ranking methods were compared:

1. “*VALUES*”, an image ranking using single feature values derived directly from the region differences on the feature maps, as described in Section 3.4
2. Ma et al.’s (2002) ranking, extracting the feature value from the global saliency map as calculated by our method and described in Section 3.3 (“*MAP_MPEG*”),
3. and Ma et al.’s (2002) ranking method applied to the saliency maps created by ezvision while calculating the first attended region (“*MAP_EZVISION*”).

Rankings calculated by these methods were compared with rankings derived from a user test. It has to be noted that for this test only four users were interviewed, which affects the statistical significance. However, the results give an indication of the agreement of the ranking methods with human judges and will be used as a pilot study for a larger user test. The users were given the 34 images, divided into six groups, and ranked them according to the perceived interest or importance. They were also asked how “rankable” they perceived the images to find out whether there were differences between image sets that were easy to rank and image sets that had very similar images. The “rankability” was to be judged as one of three categories: “easy to rank”, “difficult to rank” and “not rankable”.

First, we calculated the pairwise agreement among the four tested users employing the Spearman Rank Correlation Coefficient. The highest average agreement could be seen on the “coke can” test set with 0.79, ranging from 0.77 to 0.94 for pairwise agreement. The lowest average agreement was on the “outdoor” image set with 0.4, ranging from 0.14 to 0.83 for pairwise agreement. The pairwise agreement over all image sets was calculated as the average of agreements between two users. This ranged from 0.42 to 0.74, averaging at 0.6.

Looking at the user answers of how “rankable” they perceived the images, we get an interesting picture: The three test sets that were voted to be “easy to rank” by the majority of users (“coke can”, “library1” and “traffic”)

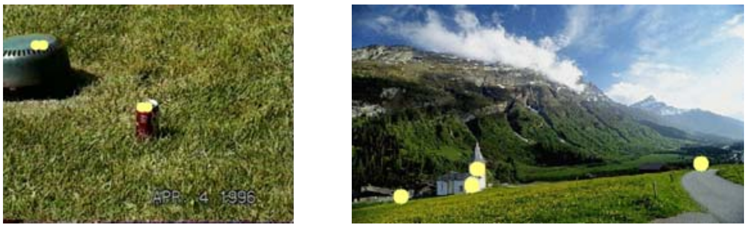


(a) Original image

(b) Interest hotspots

(c) Attended regions as calculated by ezvision

Figure 4: ctrig016 (“triangle” test set)



(a) ccoke0024 (“coke can” test set)

(b) alps (“outdoor” test set)

Figure 5: Interest hotspots for sample test images

also showed the highest average agreement, ranging from 0.72 to 0.79. On the other hand, the image sets voted to be “difficult to rank” only achieve average agreements between 0.4 and 0.54. The results indicate that there is a common opinion among the users interviewed of how comparatively interesting the images are.

For comparison with the three ranking methods, we calculated the mean ranking for all users. The ranks were then interpolated to fit the original ranking scheme, e.g. the lowest mean rank was changed to 1 and the highest was changed to 6 for a set of six images. The correlations between the mean user rankings and the three automatic methods are listed in Table 2. On average, both methods based on our MPEG-7 visual attention model (“*MAP_MPEG*” and “*VALUES*”) showed a correlation with the user rankings of 0.67 and 0.58, respectively, while “*MAP_EZVISION*” which is based on the ezvision saliency map shows an average correlation of 0.39. It is further interesting that all three methods show very similar average agreement with the test sets voted to be “easy to rank”, while only “*MAP_MPEG*” and “*VALUES*” achieve high average agreements on the “difficult to rank” image sets.

Image set	“*MAP_EZVISION*”	“*MAP_MPEG*”	“*VALUES*”
“coke can”	0.34	0.63	0.69
“library1”	0.8	0.6	0.2
“library2”	0.1	0.76	0.76
“outdour”	0.43	0.89	0.83
“traffic”	0.86	0.74	0.8
“triangle”	-0.2	0.43	0.2
average	0.39	0.67	0.58
“easy to rank” average	0.67	0.66	0.56
“difficult to rank” average	0.11	0.69	0.6

Table 2: Correlations of mean user rankings with three automatic ranking methods

To compare the performance of the single feature descriptors with the combination of all three, we calculated the rankings of all six image test sets according to the single feature descriptors folling the “*VALUES*” method and their correlation with the user rankings. In no case, the ranking according to the combined feature descriptors achieved a lower correlation with the user rankings than any ranking of a single feature descriptor. For four image sets, the agreement of the combined feature ranking with the user ranking was higher than the lowest agreement of single feature rankings. Further, for different test sets, different feature descriptors achieved very low correlations for their rankings with the user ranking, for example the correlation between the EHD ranking and the user ranking for the "coke" test set was -0.06 and the correlation between the SCD ranking and the user ranking for the "traffic" test set was -0.57. These features were balanced by the other feature descriptors and resulted in positive correlations between rankings based on the combined feature descriptors and the user rankings as shown in Table 2. This is evidence that a combination of feature descriptors results in rankings closer to our tested users' perception of interestingness than any single feature descriptor, especially in cases where one feature descriptor performs much worse than the others.

6 CONCLUSIONS

In this paper we introduced an MPEG-7 based visual attention model which we used to select interest hotspots and to rank images according to perceived interestingness. First preliminary tests showed promising results. When comparing the selected interest hotspots with attented regions calculated by another well-tested model (Itti et al. 1998), on average 57% of the four most interesting hotspots were within the first eight attended regions calculated by the other method. While this is an average agreement, we think our model calculates plausible scene descriptions as it often selected subjectively interesting spots not detected by ezvision. However, this should be substantiated in a user test as the "interestingness" of areas is very subjective.

A pilot comparison of user rankings with rankings achieved by three computational methods further indicates that our method ranks images closely to the users' perception of relative interestingness. The two methods based on our model show a higher average agreement with user rankings than the method based on ezvision's saliency map. Also, the correlations between "*MAP_EZVISION*" and user rankings showed higher variances between the different image sets, ranging from -0.2 to 0.86. Very interesting is also the comparison of "*MAP_MPEG*" and "*VALUES*", the first building its rankings on the saliency map created by our model, while the latter one ranks images according to the region differences of single feature maps on all scales. While we initially thought the normalisation and intergration of feature maps into one saliency map might affect the comparability of images, the results show that "*MAP_MPEG*" has a higher average correlation and also a lower variance of correlation between the image sets than "*VALUES*". This is an indication that the saliency maps created by our model are a good description of the image content's interest value and that the amplification of feature maps rather enhances the comparability of the interest of images. This is supported by our finding that using single feature descriptors with the "*VALUES*" ranking method results in lower correlations with the user rankings than the combination of all feature descriptors. In combination, descriptors resulting in rankings that show low correlations with user rankings are balanced by other, better performing descriptors. This balance effect will show even stronger when using the "*MAP_MPEG*" method that builds on a saliency map created from amplified feature maps. Strong feature maps will be promoted in this method, which results in "*MAP_MPEG*" rankings having the strongest correlation with user rankings. It has to be noted that our user study was only a small pilot with little statistical significance. However, a larger user study to test our findings against a statistically significant set of user rankings is underway.

In further developments, we are planning to parallelise the computation of the attention model as the calculation of MPEG-7 descriptors is a time consuming task that is repetitively performed on multiple image regions. This makes our model well suited for parallel calculation either multi-threaded on a single machine or distributed on a cluster of computers. Another step that will enhance the reliability of interest hotspot detection and image ranking is the introduction of task-related feature extraction and object recognition. If a user searches for certain objects, images including those objects must rank higher than others. It has yet to be shown how far our pre-attentive approach to ranking based on interestingess is applicable as users will always incorporate their knowledge about the scene in their rankings. Finally, we are applying this attention model to rank video frames in a research project aimed at automatic video summary.

References

Castano, R., Wagstaff, K., Song, L. & Anderson, R. C. (2005). "Validating Rover Image Prioritizations." *The Interplanetary Network Progress Report*. **42**(160).

Celaya, E. & Jiménez, P. (2003). "Salience detection in time-evolving image sequences." *Design and Application of Hybrid Intelligent Systems*. IOS Press Amsterdam, The Netherlands pp. 852–860.

Deng, D. & Wolf, H. (2005). "POISE - Achieving content-based picture organisation for image search engines." *Lecture Notes in Computer Science*. Vol. 3682. pp. 1–7.

Ferreira, W. D. & Borges, D. L. (2004). "Detecting and Ranking Saliency for Scene Description." *Lecture Notes in Computer Science*. Vol. 3287. pp. 76–83.

iLab Image Databases (2005). *Retrieved 30 October 2005.*
[http://ilab.usc.edu/imgdbs]

Itti, L. (2004). "The iLab Neuromorphic Vision C++ Toolkit: Free tools for the next generation of vision algorithms." *The Neuromorphic Engineer.* **1**(1): 10.

Itti, L. & Koch, C. (1999). "Comparison of feature combination strategies for saliency-based visual attention systems." *Proc. SPIE Vol. 3644, Human Vision and Electronic Imaging IV.* pp. 473–482.

Itti, L. & Koch, C. (2001). "Computational modelling of visual attention." *Nature Reviews Neuroscience.* **2**(3): 194–203.

Itti, L., Koch, C. & Niebur, E. (1998). "A Model of Saliency-Based Visual Attention for Rapid Scene Analysis." *IEEE Transactions on Pattern Analysis and Machine Intelligence.* **20**(11): 1254–1259.

Kadir, T. & Brady, M. (2001). "Saliency, Scale and Image Description." *International Journal of Computer Vision.* **45**(2): 83–105.

Koprinska, I., Clark, J. & Carrato, S. (2004). "VideoGCS - a clustering-based system for video summarization and browsing." *6th COST 276 Workshop on Information and Knowledge Management for Integrated Media Communication.* Thessaloniki, Greece pp. 34–40.

Koskela, J., Laaksonen, J. & Oja, E. (2001). "Self-organizing image retrieval with MPEG-7 descriptors." *Proceedings of Infotech Oulu International Conference on Information Retrieval (IR'2001).* Oulu, Finland.

Lehmann, E. (ed.) (1975). *Nonparametrics: Statistical Methods Based on Ranks.* Holden-Day, Inc. San Francisco, CA, USA.

Liu, H., Xie, X., Tang, X., Li, Z.-W. & Ma, W.-Y. (2004). "Effective browsing of web image search results." *MIR '04: Proceedings of the 6th ACM SIGMM international workshop on Multimedia information retrieval.* New York, NY, USA pp. 84–90.

Ma, Y., Lu, L., Zhang, H. & Li, M. (2002). "A User Attention Model for Video Summarization." *MULTIMEDIA '02: Proceedings of the tenth ACM international conference on Multimedia.* Juan-les-Pins, France pp. 533–542.

Ma, Y. & Zhang, H. (2003). "Contrast-based image attention analysis by using fuzzy growing." *MULTIMEDIA '03: Proceedings of the eleventh ACM international conference on Multimedia.* Berkeley, CA, USA pp. 374–381.

Manjunath, B., Salembier, P. & Sikora, T. (eds) (2002). *Introduction to MPEG-7: Multimedia Content Description Interface.* John Wiley & Sons, Inc. New York, NY, USA.

Martinez, J. M. (2004). "Overview of the MPEG-7 standard." *ISO/IEC JTC1/SC29/WCll N4509.*
[http://www.chiariglione.org/mpeg/standards/mpeg-7/mpeg-7.htm]

Saberi, M., Carrato, S., Koprinska, I. & Clark, J. (2005). "Estimation of the hierarchical structure of a video sequence using MPEG-7 descriptors and GCS." *9th International Conference on Knowledge-based Intelligent Information & Engineering Systems 2005, Special Session on Machine Learning Techniques for Image and Video Processing.* Sydney, Australia.

TU Munich (2005). "MPEG-7 experimentation model website." *Retrieved 30 October 2005.*
[http://www.lis.e-technik.tu-muenchen.de/research/bv/topics/mmdb/e_mpeg7.html]

University of Otago (2005). "Photos of the ISB." *Retrieved 30 October 2005.*
[http://www.library.otago.ac.nz/admin/photos.html]

Wolf, H. (2005). "Automatic video summary using a visual attention model." *Otago University Student's Association Postgraduate Symposium.* Dunedin, New Zealand.
[http://www.covic.otago.ac.nz/~hwolf/pubs/wolf_posterOUSA.pdf]