# Image saliency mapping and ranking using an extensible visual attention model based on MPEG-7 feature descriptors

Heiko Wolf and Da Deng *
*Department of Information Science, University of Otago*
*PO Box 56, Dunedin, New Zealand*
*Email: {hwolf1, ddeng}@infoscience.otago.ac.nz*

December 2, 2005

**Abstract**

In visual perception, finding regions of interest in a scene is very important in the carrying out visual tasks. Recently there have been a number of works proposing saliency detectors and visual attention models. In this paper, we propose an extensible visual attention framework based on MPEG-7 descriptors. Hotspots in an image are detected from the combined saliency map obtained from multiple feature maps of multi-scales. The saliency concept is then further extended and we propose a saliency index for the ranking of images on their interestingness. Simulations on hotspots detection and automatic image ranking are conducted and statistically tested with a user test. Results show that our method captures more important regions of interest and the automatic ranking positively agrees to user rankings.

## 1 INTRODUCTION

Selective visual attention is one of the most important function of the human vision system. Our gaze can be directly oriented towards salient objects in a cluttered visual scene. This is surely an attractive characteristic for artificial vision systems, as selected attention gaining and shifting will enable efficient pre-processing of the image and fast locating of the most important regions for further processing. Automatic detection of salient regions within an image is important to a range of scene analysis applications, such as landmark detection, traffic and road sign recognition, and video surveillance. An area that has gained a research focus is the modelling of visual attention in early vision, as shown in

*Corresponding author. This work is supported by FRST Grant UOOX0208, New Zealand, and SoB Grant of University of Otago.

e.g. [1][2][3]. It has been suggested that saliency is computed in a pre-attentive, bottom-up manner across the entire scene. Various feature schemes of colour, intensity and texture contrast have been proposed to produce a *saliency map* of the visual scene. On the other hand, it has been pointed out that visual attention can also be biased by top-down, task-dependent cues [1].

The idea of using pre-attentive visual features to compute saliency maps has been adopted in rapid scene analysis [4] and automatic video summarisation [5]. However, the concept of 'saliency' has been limited as image-based and defined locally, and the saliency map merely consists of an array of local saliency values.

On the other hand, the rapid progress of content-based image retrieval (CBIR) [6] research has resulted in a set of rigorously tested visual feature descriptors obtained from very large-scale MPEG-7 core experiments carried out world-wide. These MPEG-7 feature descriptors, defined on colour, texture, shape and motion features, have demonstrated very good capabilities in modeling low-level visual similarity [7] as well as high-level semantics of image content [8].

In this paper, we propose a visual attention model for image analysis built on MPEG-7 colour and texture descriptors. We also extend the use of saliency maps to calculate a global "interestingness" for an image so as to achieve image rankings according to their visual interestingness. Such kind of 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 reviews briefly the concept of visual attention and summarises other research in the field of image saliency. Section 3 introduces our visual attention model using MPEG-7 visual features. The extension of the visual attention model for image ranking is proposed in Section 4. Section 5 presents the simulation and the simulation results. The paper is concluded in Section 6 with some discussion on future direction. 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 image ranking methods using 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. It helps with efficient information extraction from an image and guides further object recognition processes. This is especially valuable as early vision processing is context independent and therefore allows to derive image content descriptions without domain knowledge. In one of the most influential work by Itti et al. [1], recent works on computational modeling of visual attention were reviewed and a bottom-up saliency-based framework was presented. The framework incorporates early vision features such as intensity contrast, colour contrast, orientation differences, and direction of motion.

They also hypothesised that these features are integrated into one saliency map, in which the combination of features determines the points that draw the most attention. In [4] a system was implemented based on the framework, where saliency points were presented in attended order. It was found that this model can identify salient regions even when strong noise is introduced.

Among the recent works, Kadir and Brady [2] proposed to find corners in the image and extract the optimal salient regions by maximising the local entropy around the corner areas. A set of intuitive saliency features and weights were used in [9] to extract regions of interest, but the integration of features was not attempted. In [10], an evolutionary programming approach was introduced and proven to work effectively even though with a high cost on computing time. In [3] it was proposed to use 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. In [11], salient regions in an image are computed using the Euclidean distance between the RGB values of a pixel and its neighbourhood. The salient regions are used 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 [12].

Although visual attention modeling usually poses an emphasis on the bottom-up processes of feature extraction and saliency location, it has become obvious that a more complete model of attention control must include top-down cues generated from object or scene recognition [1]. Some recent works such as [13] have been investigating the modeling of top-down attention bias in vision systems.

## 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. To our knowledge there is little research on automatic image ranking or prioritisation based on image saliency. In a NASA study, the prioritisation of Mars Rover images was validated [14]. 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. The ranking criterion however is a "scientific value" which is specifically defined and judged by experts. Correlation analysis was carried out so as to assess if rankings given by different experts agree to each other.

As an important application, image ranking is closely related to image retrieval, where often a large number of images returned by image search engines can be automatically ranked according to query conditions or some other criteria. In a previous work, we used MPEG-7 feature descriptors to organise results returned from contemporary image search engines according to visual similarity [15]. In [16], an attention model was used which consists of saliency, face detection and query-dependent attention objects. Images were then cropped to the most interesting regions and ranked according to the similarity of those regions.

Another application of image ranking based on a visual attention model is to select interesting frames in video. The visual attention model proposed by [1] has been employed as part of a system for video summarisation in [5].

# 3 THE VISUAL ATTENTION MODEL

The visual attention model aims to describe the attention or saliency that an image produces. Various features have been used in the literature to extract regions of interest. Here we propose an extensible framework based on MPEG-7 visual feature descriptors. The adoption of MPEG-7 features is based on several considerations. First, these MPEG-7 feature descriptors have been extensively tested in content-based image retrieval studies, and their capabilities in modeling low level visual similarity as well as semantic modeling have been proven by CBIR researches. The generalisation ability of these feature descriptors is therefore very promising. Secondly, such a framework can include bottom-up early vision features as well as top-down task-dependent components such as face detection and object recognition, since MPEG-7 includes visual feature descriptors to support both feature extraction strategies. Being an extensible framework, other features that can be used to create a feature map might be included as well. The seperate feature maps are then combined and a global saliency map is created. The framework is shown in Figure 1.

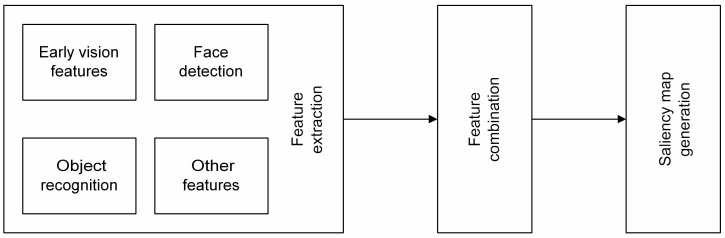



Figure 1: *Extensible visual attention framework*

Hereafter in this section, we describe the mechanisms of feature extraction, feature map amplification and combination in our visual attention model.

## 3.1 Feature extraction

The bottom-up visual attention is guided by early visual features such as intensity contrast, colour opponency, orientation, and direction and velocity of motion. It is also noted that it is not the feature characteristics themselves but the difference between a feature region and its neighbourhood that generates attention. The MPEG-7 visual feature descriptors include multiple descriptors for both colour and orientation features that have been rigorously tested in its standardisation process. However, there is no feature descriptor that specifically describes intensities, so we introduce our own intensity histogram descriptor as

part of the visual attention model. Differing from [4], which calculates visual features per pixel, MPEG-7 feature descriptors in our model are calculated from images or image regions and the feature maps are obtained based on the difference between image regions.

For each feature descriptor, the saliency of a region is defined as the average difference 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 apply 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 $8 \times 8$ pixels 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.

Given a region $R_c$, denote its neighbours as $R_n \in \Omega$, $n = 1, 2, \ldots, N$. Denote the feature code of a region as $f(R)$, where $f$ corresponds to the feature descriptor being used. The saliency value of Region $R_c$, defined on $f_S$, is:

$$\theta_f(R_c) = \frac{\sum_{i=1}^{N} \mathrm{DIST}\{f(R_c), f(R_i)\}}{N}. \quad (1)$$

Here the distance measure DIST$\{.\}$ is dependent on the feature descriptor being used and will be defined individually for the descriptors given as follows. More details about these MPEG-7 feature descriptors can be found in [17].

#### 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 [17, pp.198–201] which also includes a schematic diagram of the SCD generation.

As described in Eq.(1), the saliency of a region is calculated as the average difference between the region's SCD and the SCD of regions in its 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 Orientational features - Edge Histogram Descriptor

Orientations within images are observable as edges and textures. In this implementation, we use the MPEG-7 Edge Histogram Descriptor (EHD) as orientational features.

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.

To calculate the EHD feature map, we apply Eq.(1) to the EHD of each region on all scales. The distance measure between two Edge Histogram Descriptors 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 $\times$ 5 edge types), 5 bins globally and 65 bins semiglobally (13 semiglobal groupings of subimages $\times$ 5 edge types). To equalise weights, the global histogram distance is multiplied by a factor of 5, resulting in the following distance metrics:

$$\begin{aligned} \mathrm{DIST}(\mathrm{EHD}_A, \mathrm{EHD}_B) = \quad & \textstyle\sum_{i=0}^{79} |h_A(i) - h_B(i)| \\ & +5 \times \textstyle\sum_{j=0}^{4} |h_A^g(j) - h_B^g(j)| \\ & + \textstyle\sum_{k=0}^{64} |h_A^S(k) - h_B^S(k)| \end{aligned} \tag{2}$$

### 3.1.3 Intensity features - Intensity Histogram Descriptor

As the MPEG-7 visual feature descriptors do not include intensity descriptors, we need to define our own intensity histogram descriptor (IHD) for the visual attention model. Intensity differences can be calculated based on different features such as luminance or brightness. For the sake of simplicity, we compute intensity as the pixel value of the grey-scale transform of an input image. However, it can be obtained during the process of calculating other colour descriptors such as SCD, as the V value in the HSV colour space already gives the intensity value.

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 IHD feature map is created by using IHD features for Eq.(1). The distance measure between intensity histograms $\mathrm{IHD}_A$ and $\mathrm{IHD}_B$, similar to the distance measure for the EHD, is defined using the city-block distance:

$$\mathrm{DIST}(\mathrm{IHD}_A, \mathrm{IHD}_B) = \sum_{i=0}^{15} |\mathrm{IHD}_A(i) - \mathrm{IHD}_B(i)| \tag{3}$$

### 3.2 Feature map amplification

Early vision attention is triggered by feature contrast. 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. In [18], four feature combination strategies were compared, 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 intensity features as described in the previous sections, these maps can now be combined in order to generate a saliency map to represent the attention that parts of this image trigger.

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. Upon evaluating the created feature map, we can now locate the maxima and extract the most interesting spots from the image. The overall computational diagram is shown in Figure 2.

An example of the different feature maps and their combination to one saliency map, using a "coke can" image for example, can be found in Figure 3.

## 4 AUTOMATIC IMAGE RANKING

An "interest value" based on image saliency can be used as a ranking criterion for image sets. Previously, it has been proposed to rank images based on their saliency maps [5]. Ma's method [5] proposes to use a saliency map to calculate an attention value that takes position, size and brightness of salient regions into account. There is, however, a problem in this approach, since feature maps maps are created for each feature channel and normalised to grey-scale images dependent on the feature map. The final saliency map is obtained by integrating different normalised feature maps, hence there is no guarantee that saliency maps of different images can be accurately compared. Also, a Gaussian template was used to assign lesser weight to the outer regions of the image. Although it is generally accepted that humans perceive the centre 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.

In order to create a saliency value that represents the interestingness of an image relative to the other images within a given image set, we calculate a global scalar value $\Theta$ for each (not-normalised) feature map of the image by

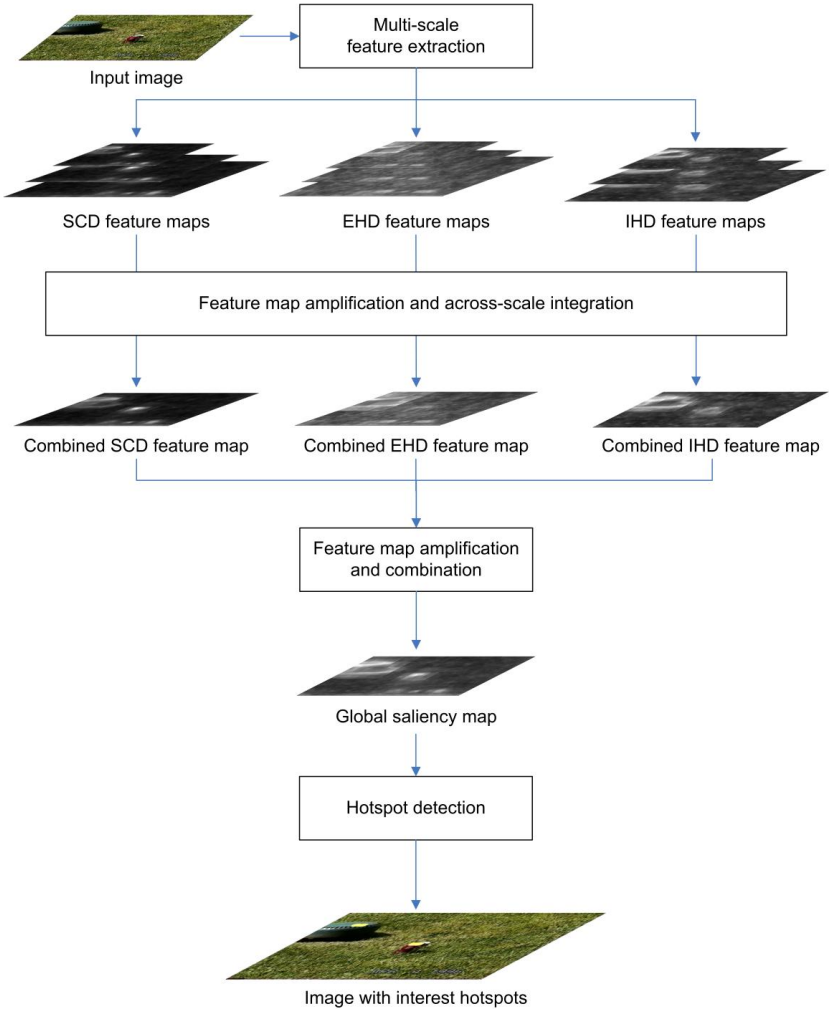
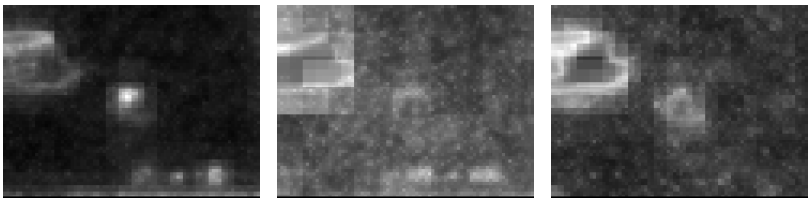
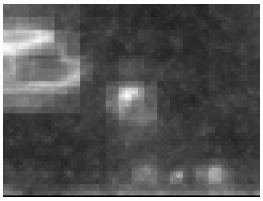



Figure 2: *Diagram for hotspots detection.*



(a) SCD feature map (b) EHD feature map (c) IHD feature map



(d) Saliency map

Figure 3: Feature maps for "coke can" test image

first summing up the average regional distances and dividing them by the total number of regions, and then averaging the attention values for each scale over the number of scales:

$$\Theta_f = \frac{1}{M} \sum_{s=1}^{M} \frac{\sum_{i=1}^{N_s} \theta_f(s)(i)}{N_s}. \tag{4}$$

Here the feature code $f$ is one of SCD, EHD and IHD, $\theta(s)(i)$ is the saliency of region $i$ on the respective feature descriptor $f$ and scale $s$, while $i$ runs from 1 to $N_s$ (the number of regions for scale $s$) and $s$ runs from 1 to $M$ (the number of scales).

After calculating saliency values on each feature channel respectively, we can then normalise each channel over the whole image set. Each channel is normalised between 0 and 1. 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, the single feature values are combined to an overall Saliency Index (SI) $\Theta$:

$$\Theta = \frac{\Theta_{\text{SCD}} + \Theta_{\text{EHD}} + \Theta_{\text{IHD}}}{3} \tag{5}$$

# 5 SIMULATIONS

## 5.1 Simulation setup

A system is constructed to detect hotspots in images as well as to rank images according to their saliency values. The system is implemented in C++ based on the MPEG-7 eXperimentation Model reference implementation [19]. 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 4 and second using the combined saliency map.

We use a test image database that includes 37 images, of which 13 were taken from the University of Otago Library image series [20], and 24 from the iLab Database [21]. Of the 24 iLab images, 6 images were part of the "autobahn", "coke" and "triangle" image sets [1] respectively, and 6 images were taken from the "outdoor" image set [4].

The ranking of images according to their perceived interestingness can be highly subjective. An ideal test bed should therefore involve a good user study. It has to be noted that user tests are characterised by limitations such as the number of available subjects, their time constraints and fatigue as well as finding a suitable test design. We have undertaken a user study with 26 subjects for image ranking evaluation, which gave some promising results that we present as follows.

## 5.2 Evaluation of interesting spot detection

To assess the detected interesting spots, results obtained using our method are compared with those of *ezvision* [22], an implementation of Itti's model of visual attention [1]. As this model has been validated by extensive user tests, we can use the regions of attention found by ezvision as a good reference point.

We calculated the first four hotspots 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 attended regions from ezvision. A hotspot was counted as agreeing with an attended region if they cover the same or similar regions.

As we see, the agreement is quite variable over the image sets as shown in Table 1. The average agreement of the hotspots with the first four attended ezvision regions is at 46%. The average agreement rises to 58% when we compare the hotspots with the first eight attended regions. The "library1" set shows a significantly lower agreement than the average value. This might be due to the complexity of the images. In particular, it contains a large number of objects of similar visual characteristics. Figure 4 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 that our approach finds more significant spots, for example groups of people. If the "library1" set is taken out as an 'outlier', the average agreement with the first 8 attended regions will

| Image sets | Agreement (First 4 Regions) | Agreement (First 8 Regions) |
|---|---|---|
| "coke can" | 79% | 79% |
| "library1" | 19% | 19% |
| "library2" | 42% | 56% |
| "outdoor" | 46% | 71% |
| "traffic" | 38% | 54% |
| "triangle" | 54% | 67% |
| Average | 46% | 58% |

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

rise to 65%, which suggests a reasonable agreement between ezvision and our method in general.

Interesting results are obtained also on images from other sets, as shown in Figure 5, where the first four hotspots are compared with the first four attended regions reported by ezvision side-by-side. For the "traffic" image, our approach manages to capture more traffic signs. For the "triangle" image, the red triangle was missed by the ezvision method, but picked up by our method. These differences can be explained with the different methods of calculating the saliency map, especially the feature schemes used. Without aiming at biological plausibility, we do not simulate inhibition of return for attended regions or alter the saliency map when locating the next hotspot.

As the perception of the "interestingness" of an image is very subjective, it is hard to assess the accuracy of our method in locating interesting spots solely based on the comparison to another method. However, in comparison with ezvision, our approach based on MPEG-7 features does give comparable and in cases even better performance, presenting salient regions of higher semantic significance, as indicated in the examples given.

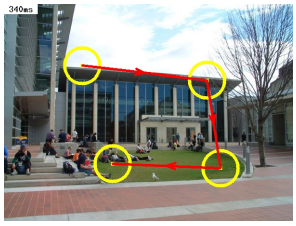
The robustness of our approach is also tested with noise-added images. This is shown in Figure 6, where an example is given for an "alps" image from the "outdoor" image set. Noise was generated from a uniform distribution between -30 and 30 and added onto the colour values of image pixels.

## 5.3 Evaluation of image ranking

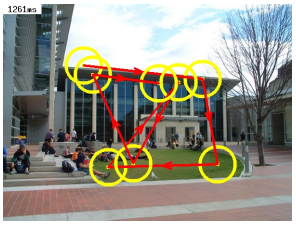
In a second test case, we compare the image rankings derived from two different methods with the rankings from users. The image test sets are the same as used to assess the interesting spots. In our main study, 26 people were interviewed. The users were given the 34 images, which were divided into six groups, and asked to rank them according to the perceived interest.

To compare the results from the user test with computed rankings, we need to establish whether viewers agree on a common ranking and how significant this common ranking is. This problem is known in many situations where sets

(a) First four hotspots

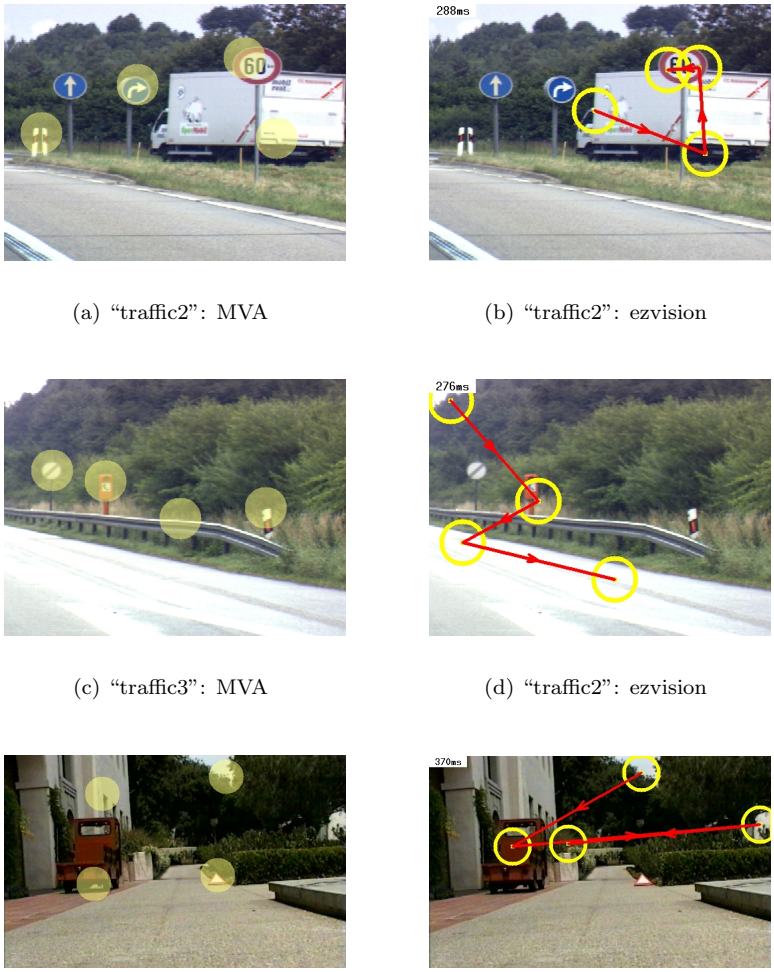



(b) First four attended regions




(c) First eight attended regions

Figure 4: lib1d (from the "library1" set)




(a) "traffic2": MVA

(b) "traffic2": ezvision

(c) "traffic3": MVA

(d) "traffic2": ezvision

(e) "triangle1": MVA

(f) "triangle1": ezvision

Figure 5: Comparison of the first four hotspots.

(a) alps

(b) alps with noise

Figure 6: Robustness of hotspots location in presence of image noise.

of objects are rated by multiple judges, for example figure skating competitions or product comparisons. The overall mean of average ranks is defined as $R = \frac{1}{2}(s+1)$, with $s$ being the number of judged elements. One way to measure the overall closeness of a calculated average ranking to the overall mean is Friedman's statistic [23], which for example has been applied to assess the statistical significance of wine taste rankings [24]. Friedman's statistic is defined as

$$Q = \frac{12N}{s(s+1)} \sum_{i=1}^{s} [R_i - \frac{1}{2}(s+1)]^2, \tag{6}$$

where $N$ is the number of judges, $s$ the number of judged cases and $R_i$ the average rank of case $i$. $Q$ is high when the average ranks $R_i$ are significantly different from each other. This leads to our first hypothesis to be verified:
**H1 The average rankings of each image in the test set are significantly different from each other.**

Table 2 shows $Q$ for the six image sets used in our user test calculated for $N = 26$ and $s = 6$ for "coke", "outdoor", "autobahn" and "triangle", $s = 4$ for "lib1" and $s = 9$ for "lib2". For large numbers of $N$, the critical values $c$ for Friedman's statistic can be approximated with a $\chi^2$ distribution with $s - 1$

degrees of freedom. The probabilities of $Q$ being greater than or equal to the critical value $c$ are stated in Table 2. Hypothesis H1 is true for all image sets except "triangle" with values for $Q$ that are significant at the 99.9% level. The "triangle" set has a low value for $Q$, which also has a significance of lower than 90%.

| Image set | "coke" | "lib1" | "lib2" | "outdoor" | "autobahn" | "triangle" |
|---|---|---|---|---|---|---|
| $Q$ | 25.605 | 16.292 | 62.561 | 40.879 | 92.374 | 9.07 |
| $W$ | 0.197 | 0.209 | 0.301 | 0.314 | 0.711 | 0.07 |
| $P(Q \geq c)$ | 0.000 | 0.001 | 0.000 | 0.000 | 0.000 | 0.106 |

Table 2: Friedman's statistic and Kendall's W for user rankings of test image sets

These results indicate that the images in each image set are rather 'rankable', except those from the "triangle" set. After confirming this, we can proceed further to examine how well the average rankings represent the opinion of single users. A test for the overall agreement of judges' rankings is Kendall's coefficient of concordance $W$, which is statistically equivalent to Friedman's statistic [25]. Kendall's coefficient is computed as

$$W = \sum_{i=1}^{s} \frac{(R_i - R)^2}{\frac{N}{12}(N^2 - 1)}, \tag{7}$$

where $R_i$ is the average rank for case $i$, $R$ the overall mean of average ranks and $N$ the number of judges. $W$ is ranging between 0 and 1 with 0 being complete disagreement and 1 being complete agreement of judges. Like $Q$, $W$ is distributed as $\chi^2$ with $s - 1$ degrees of freedom for large $N$.

Using the values for $W$ in Table 2, we evaluate the second hypothesis:

**H2 Users strongly agree on the image rankings.**

Simular to $Q$ values, the values of $W$ for almost all test sets indicate a strong agreement among users and are significant at the 99.9% level. Therefore Hypothesis H2 can be accepted for these cases. The only exception is again the "triangle" set. This is however not surprising, given the finding that Hypothesis H1 does not hold for the "triangle" set. On the other hand, it was expected to show a low agreement as its images present a red triangle in very different settings. This makes it very hard for a viewer to establish a metric on which to rank the images. One user also mentioned how he associated the triangles with accident scenes, which influenced his ranking. While it is true for any test case that interestingness is a very subjective concept and that every viewer has his own interpretation of image content influenced by cultural and educational background, personal interest or aesthetic perception, this seems to be particularly strong for the "triangle" test set. This set of images therefore is omitted in the comparison with computed rankings.

Two image ranking methods were compared with the user rankings. To compute the saliency ranking of images and compare it with the average user

rankings, we use the method as described in Section 4 to rank on the overall saliency indeces calculated from multiple feature maps. This method is then contrasted with the Ma's ranking method based on saliency maps. These saliency maps are generated by ezvision.

The third hypothesis we want to test on is:

**H3 The computed rankings are positively correlated with the user rankings.**

Both methods of automatic ranking are assessed on how well the rankings generated agree with the average user ranking. We use the Spearman Rank Correlation Coefficient [23] 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 rank of $x_i$ according to Ranking $A$ and $B$, respectively. The Spearman rank correlation coefficient is defined as

$$r = 1 - 6 \sum_{i=1}^{n} \frac{(A_i - B_i)^2}{n(n^2 - 1)} \tag{8}$$

The correlation coefficient ranges from -1 to 1, with -1 indicating complete disagreement and 1 complete agreement of two rankings.

The correlation coefficients between the mean user rankings and the two automatic methods are listed in Table 3. On average, the method based on the saliency index of our visual attention model (noted as "*SI-MPEG7*") shows a correlation with the user rankings of 0.55, while "*MAP-ezvision*", which is based on the ezvision saliency map, shows an average correlation of 0.35. The difficulty of simulating user rankings computationally is shown by the fact that only two correlations are significant at the 99% and one at the 95% level. Correlations in Table 3 that are significant at the 95% and 99% level are marked with $^*$ and $^{**}$, respectively. Despite there are a few weaker cases such as "library1" and "outdoors", we believe Hypothesis H3 can be accepted for our SI-MPEG7 method with an average correlation coefficient of 0.55.

| Image set | MAP-ezvision | SI-MPEG7 |
|---|---|---|
| "coke" | $-0.029$ | $0.928^{**}$ |
| "library1" | 0.8 | 0.2 |
| "library2" | 0 | 0.65 |
| "outdoor" | 0.086 | 0.143 |
| "autobahn" | $0.943^{**}$ | $0.829^{*}$ |
| average | 0.36 | 0.55 |

Table 3: Correlation of mean user rankings with the automatic ranking methods

# 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. A comprehensive comparison of user rankings with rankings achieved by two computational methods indicates that our method ranks images closely to the users' perception of relative interestingness. Our approach of image ranking based on the attention model shows 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 -0.029 to 0.943, while the performance of SI-MPEG7 was all positive and stabler.

An interesting observation in the user tests was the time taken to rank images. Although not measured, the time needed to rank images seemed to increase significantly when people were asked to rank the nine images of the "lib2" image set. Psychology research has indicated that people's minds can only hold about four objects they have just seen [26] and that this capacity varies across individuals [27] which will affect ranking of larger image sets. Further, it will be become harder to establish differences and rank them with growing numbers of images. This is where the great advantage of automatic image ranking lies - in organising large collections that would be too time- and work-intensive for humans to process.

So far we have only made use of several colour and texture descriptors in our model, but this model can be extended using more MPEG-7 descriptors, e.g. of shape and motion features. Other task-related or customly defined features can also be introduced into the extensible visual attention model, hence enhancing the reliability of hotspots detection and image ranking. Object recognition can be conducted to simulate the top-down bias for visual attention.

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. In further development we plan to parallelise the computation process of the attention model. This will hopefully leverage the efficiency of our visual attention model so that it can be used to detect and rank video key-frames in real-time for the task of automatic video summarisation.

# References


[1] L. Itti and C. Koch. Computational modelling of visual attention. *Nature Reviews Neuroscience*, 2(3):194–203, March 2001.

[2] T. Kadir and M. Brady. Saliency, scale and image description. *International Journal of Computer Vision*, 45(2):83–105, 2001.

[3] W.D. Ferreira and D.L. Borges. Detecting and ranking saliency for scene description. In *Lecture Notes in Computer Science*, volume 3287, pages 76–83, 2004.

[4] L. Itti, C. Koch, and E. Niebur. A model of saliency-based visual attention for rapid scene analysis. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 20(11):1254–1259, November 1998.

[5] Y.F. Ma, L. Lu, H.J. Zhang, and M.J. Li. A user attention model for video summarization. In *MULTIMEDIA '02: Proceedings of the tenth ACM international conference on Multimedia*, pages 533–542, Juan-les-Pins, France, 2002.

[6] A. Smeulders, M. Worring, S. Santini, A. Gupta, and Jain R. Content-based image retrieval at the end of the early years. *IEEE Transaction on Pattern Analysis and Machine Intelligence*, 22(12):1349–1380, 2000.

[7] B.S. Manjunath, J.-R. Ohm, and V.V. Vasudevan. MPEG-7 color and texture descriptors. *IEEE Trans. Circuits Syst. Video Technol.*, 11:703–715, June 2001.

[8] L. Wang and B.S. Manjunath. A semantic representation for image retrieval. In *Proc. ICIP 2003*. IEEE, 2003.

[9] J. Luo and A. Singhal. On measuring low-level saliency in photographic images. In *Proc. IEEE Conf. on CVPR*, pages 1084–1089, 2000.

[10] F.W.M. Stentiford. An evolutionary programming approach to the simulation of visual attention. In *Proc. IEEE Congress on Evolutionary Computation*, pages 851–858, 2001.

[11] E. Celaya and P. Jiménez. Salience detection in time-evolving image sequences. In *Design and Application of Hybrid Intelligent Systems*, pages 852–860, Amsterdam, The Netherlands, 2003. IOS Press.

[12] Y.F. Ma and H.J. Zhang. Contrast-based image attention analysis by using fuzzy growing. In *MULTIMEDIA '03: Proceedings of the eleventh ACM international conference on Multimedia*, pages 374–381, Berkeley, CA, USA, 2003.

[13] Silvia Corchs, Martin Stetter, and Gustavo Deco. Systems-level neuronal modeling of visual attentional mechanisms. *Artif. Intell. Rev.*, 20:143–160, 2003.

[14] R. Castano, K. Wagstaff, L. Song, and R.C. Anderson. Validating rover image prioritizations. *The Interplanetary Network Progress Report*, 42(160), February 2005.

[15] D. Deng and H. Wolf. POISE - Achieving content-based picture organisation for image search engines. In *Lecture Notes in Computer Science*, volume 3682, pages 1–7, August 2005.

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

[17] B.S. Manjunath, P. Salembier, and T. Sikora, editors. *Introduction to MPEG-7: Multimedia Content Description Interface.* John Wiley & Sons, Inc., New York, NY, USA, 2002.

[18] L. Itti and C. Koch. Comparison of feature combination strategies for saliency-based visual attention systems. In *Proc. SPIE Vol. 3644, Human Vision and Electronic Imaging IV*, pages 473–482, May 1999.

[19] TU Munich. Mpeg-7 experimentation model website. URL http://www.lis.e-technik.tu-muenchen.de/research/bv/topics/mmdb/e_mpeg7.html, 2005. Retrieved 30 October 2005.

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

[21] anon. iLab image databases. URL http://ilab.usc.edu/imgdbs. Retrieved 30 October 2005.

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

[23] E.L. Lehmann. *Nonparametrics: Statistical Methods Based on Ranks.* Holden-Day, Inc., San Francisco, CA, USA, 1975.

[24] R.E. Quandt. Measurement and inference in wine tasting. In *Meetings of the Vineyard Data Quantification Society*, Corsica, 1 - 3 October 1998.

[25] J.J. Higgins. *Introduction to modern nonparametric statistics.* Brooks/Cole, Pacific Grove, CA, USA, 2004.

[26] S.J. Luck and E.K. Vogel. The capacity of visual working memory for features and conjunctions. *Nature*, 390:279–281, 20 November 1997.

[27] E.K. Vogel and M.G. Machizawa. Neural activity predicts individual differences in visual working memory capacity. *Nature*, 428:748–751, 15 April 2004.