FAST VIDEO OBJECT SELECTION FOR INTERACTIVE TELEVISION

Rémi Trichet, Bernard Merialdo

Institut Eurécom, Sophia Antipolis, France

ABSTRACT

In this paper, we study the problem of the fast selection of video objects, as an aid for the efficient semi-automatic annotation of video programs. In a regular system, the user has to draw a bounding box around the object, requiring at least two clicks on the image. We propose and experiment algorithms that allow the selection by indicating only one point inside the object, therefore requiring only one click. The problem is then to identify the correct bounding box. We use an attention model and a growing algorithm to construct the most plausible bounding box, based on the comparison of the interior, the border and the outside of the box. We present some experimentation that suggests that in many cases, our algorithm is able to propose a reasonable bounding box.

1. INTRODUCTION

With the rapid development of multimedia technologies and the convergence between broadcast and network communications, Interactive Television is becoming a more and more popular area of application. Within the scope of Interactive Television, the GMF4iTV project [1] has developed an application for Hypervideo television programs, where active video objects are associated to metadata information, imbedded in the program stream at production time, and can be selected by the user at run time to trigger the presentation of their associated metadata. Demonstrations of the current prototype have shown the interest of such interactivity. However, a crucial factor for the success of such systems is the extra cost necessary for the semi-automatic annotation of the video program. Therefore, it is very important to facilitate this annotation as much as possible.

In the current system [1], the video producer has to define a bounding box for each active object in a video sequence. This bounding box is defined only for one image, and then tracked automatically on all images of the same video shot. Defining the bounding box requires two clicks, or drawing a long stretch across the object. In order to speed up this process, we propose a mechanism where the user would only select one point inside the object, and the system would automatically suggest the bounding box. In practice, it is likely that users will select objects which are distinct from the background, either by their movement, their color or their contrast, so that we may hope that this combined information can lead to a reasonable bounding box.

In this paper, we investigate the use of a video attention model to define regions-of-interest (ROI) in the image. In section 2 we review some related work. Section 3 presents the framework of the system. Our visual attention model is described in section 4. Section 5 discusses the bounding box growing algorithms and

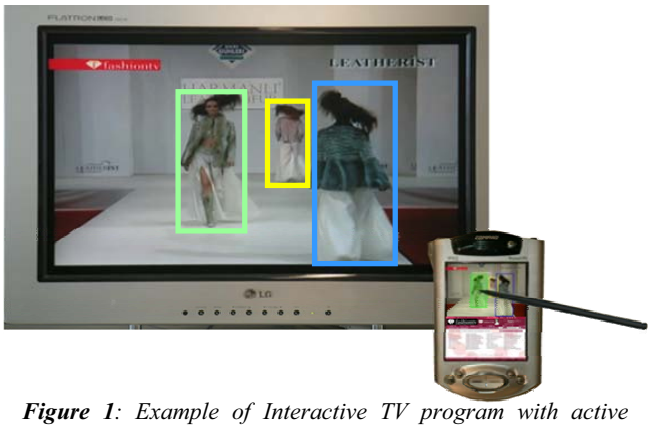

***Figure 1**: Example of Interactive TV program with active video objects*

presents some experimental results, together with an adaptive version of the base algorithm. We conclude in section 6.

2. RELATED WORK

Studies on visual attention and eye movements [17] have shown that, as long as we are looking to an image with the same knowledge and motivations, there is a strong correlation of our eye movements. Moreover, we do not scan the whole scene, our gaze is always attracted by the same areas of the image. Stelmach [11] found similar results for video. According to these findings, visual attention models try to automatically determine parts of an image or a video that a human is likely to gaze at.

One approach is to first segment the image into homogenous regions and score each area with intuitive measures [3][9]. The drawback is the result is highly dependent upon the quality of the segmentation.

Another approach argues that the most salient areas are those that contain distinctive and uncommon features [12][13]. This method relies upon the comparison of the features of representative points of the scene, the saliency of these points being inversely proportional to the occurrence of their features.

A last kind of approach [2][6][15][16], originally developed by Koch and Ullman [7], builds a feature map for each low-level feature (for example color, intensity, orientation) and combines these maps into a saliency map representing the visual interest of each part of the scene. It provides accurate perceptual analysis of the image at the cost of intensive preprocessing and exhaustive scene analysis. There are many variations, based on the possible combination of the features.

3. OVERVIEW OF THE FRAMEWORK

Our approach follows the feature map model. We use four visual features to build the saliency map. In addition to the standard color, motion, and contrast, we also include the distance to the user-selected position. The most plausible bounding box is found by a bounding box growing algorithm, which starts from the initial user-selected point and stops when the saliency of the extension is not sufficient. During this growing, only the saliency of the pixels surrounding the bounding box is calculated, saving computational time. We also suggest an adaptive version of the saliency.

The proposed framework is illustrated in the following figure. First, the color map is generated from the input image. Meanwhile, the motion map is build from the input image and his previous image. The motion feature is the only one computed before the selection process. The position, the contrast, and the color entropy of a pixel are dynamically calculated when its saliency is needed.

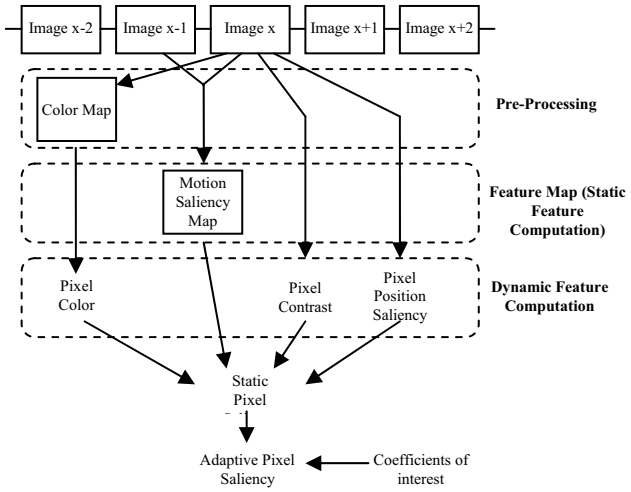

***Figure 2**: Organization of the saliency computation*

4. VISUAL ATTENTION MODEL

In order to determine the ROI of an image, we need to determine the factors which influence our attention. Previous work [2][5][15] have suggested features which could be represented in a visual attention model. In our work, we use four features (color entropy, motion, contrast, and position) to represent the visual attention map of an image.

Color Entropy

We first define a color map for the image by clustering the various colors which appear in the image. For simplicity, the criteria to compare colors is the variance of the (R,G,B) coordinates, as the difference of those variances will be an approximation of the difference in color hue and saturation. Then the color entropy *CE(i)* for a pixel is computed as the inverse of the occurrence of its color in the color map.

Contrast

Contrast is one of the most important features in a visual attention model. We define the contrast *C(i)* of a pixel as the inverse frequency of its color in a 5x5 neighborhood. (This technique gives us better results than a difference of Gaussians).

Motion

Motion is the only cue that is not being dynamically determined. For the sake of computational complexity, we adopt 3x3 block-motion vectors for motion analysis and user-attention representation. An approach similar to [4][8] is used to evaluate the motion saliency map. For each motion vector *mv*, we first compute its intensity $I(mv) = \sqrt{dx^2 + dy^2}$ and quantize its phase $P(mv) = \arctan(\frac{dy}{dx})$ into a 8 bin histogram. The spatial consistency is then defined as

$$C(mv) = -p_s(h(mv)) \times \log(p_s(h(mv)))$$

where $p_s(h(mv))$ is the probability of the bin *h(mv)* of the motion vector *mv*. The larger *C(mv)*, the more consistent the motion field of bin *h(mv)*.

The motion saliency of pixel *i* whose surrounding block has motion *mv(i)* is defined as :

$$M(i) = I(mv(i)) \times C(mv(i))$$

Position

We make the assumption that the point selected by the user is close to the center of the desired object, thus we add a position feature which is based on the distance of the current pixel to the user-selected point (P_x,P_y). The position feature *G(i)* is computed as a normalized Gaussian function centered at (P_x,P_y).

Static pixel saliency

The static saliency *S(i)* of a pixel *i* is then obtained by

$$S(i) = (\omega_M M(i) + \omega_C C(i) + \omega_{CE} CE(i)) \times G(i)$$

where ω_M, ω_C, ω_{CE}, are positive weights satisfying the constraint:

$$\omega_M + \omega_C + \omega_{CE} = 1$$

We then define the static saliency of a region as the average saliency of its pixels.

5. FAST SELECTION PROCESS

5.1. Algorithm

The ROI determination is achieved by growing a bounding box from the starting point of the user selected point. This algorithm uses the visual saliency to determine if a region should be added to the bounding box or not. The algorithm is initialized by placing a bounding box of $2 \times GS$ size centered at the user-selected pixel, where *GS* is the growth step size (in pixels) of the method. Then, for each iteration, the bounding box could extend by *GS* pixels into one of the four directions: left, right, up, or down. We choose the direction for which the extension has the highest saliency (we also tried the direction for which the extended bounding box has the highest saliency, but this gave lower performance). The growing stops when the extension saliency is under a given threshold θ.

5.2. Experiments

We tested our approach on two video programs, one fashion show and one music show. The active video objects considered are all human beings, appearing in total or in part in the videos. These objects have been manually selected, and users have drawn a bounding box around the objects that is being considered as the reference ground truth. The intuitive definition of the best bounding box was to consider the largest bounding box containing the object of interest with the minimum number of background pixels. The measures to evaluate the quality of a proposed

bounding box are the precision and recall rates at the pixel level when compared with the reference bounding box.

In a first experiment, we want to compare the effectiveness of each feature in the total combination. So we fix the parameter *GS*=10, and evaluate each feature independently, then a combined model with almost equal weights, for a range of values of the threshold θ. The results are displayed in the figure below:

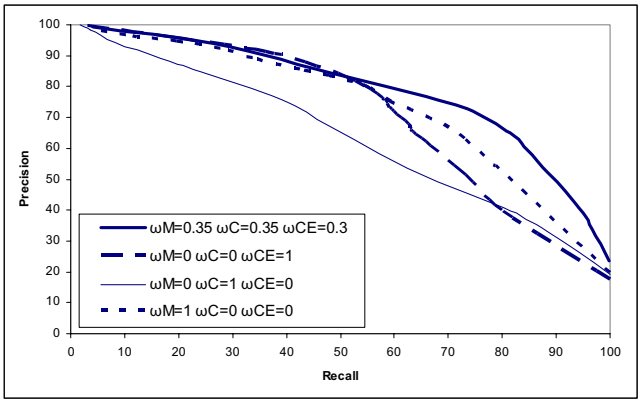

***Figure 3**: Precision and Recall graph for various combinations of features*

Thus, we can observe that the combined model provides a better performance than any of the single feature, although for low recall rates, the motion and color entropy are of similar quality. The contrast feature provides a lower performance than the others. We use the combined model in the following experiments.

5.3 Growth Step size determination

The growth step size parameter GS is the number of pixels by which a bounding box is extended in a given direction. The smaller *GS*, the more precise the boundary of the ROI will be, but the greater the risk that the bounding box boundary may be trapped in a low saliency region. We performed some tests with varying values for GS. The results are indicated in the following graph:

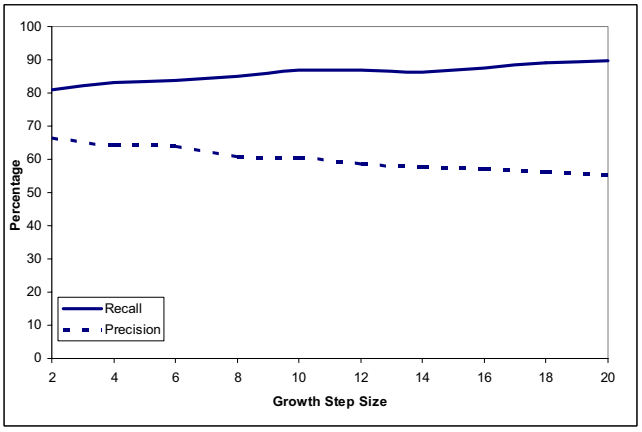

***Figure 4**: Precision and Recall for various Growth step sizes.*

This graph shows that there are monotonic relations between GS and the precision and recall rates. This motivates our choice for GS=10 as a reasonable value.

5.4. Adaptive pixel saliency

Some visual objects contain gradual color variations, for example, due to variable lighting. This suggests that an adaptive scheme could improve the accurate determination of the object boundary. We added to our model a coefficient of interest associated to each of the color layers and representing the confidence for this color to belong to the object. The coefficient of interest *CI(i)* is defined as:

$$CI(i) = \sqrt[x]{P(i) \times A(i)}$$

(where x is a user defined parameter, we experimentally found that x=4 is a reasonable value). *P(j)* and *A(j)* are respectively the proportion and the attraction of the layer j in the bounding box calculated by

$$P(i) = \frac{N(i)}{(\sum_{k=0}^{n} N(k))/n} \qquad A(i) = \frac{S(i)}{(\sum_{k=0}^{n} S(k))/n}$$

with n the number of color layers represented in the bounding box, *N(i)* and *S(i)* respectively the number of pixels and the average saliency of the color layer i in the bounding box. Thus *P(i)* is the relative frequency of color layer i in the bounding box. The attraction parameter *A(j)* represents the layer i relative average visual saliency when compared with the bounding box average saliency. Therefore frequent color layers with high saliency are favored by the coefficient of interest. The adaptive saliency *AS(i)* is computed from the original saliency *S(i)* by:

$$AS(i) = S(i) \times CI(c(i))$$

where *c(i)* is the color layer of pixel i.

The following figure shows the precision and recall rates compared between the standard and adaptive model, for a range of threshold values for θ. It can be seen that, for a given value of the threshold, the adaptive model provides a better precision, but lower recall than the standard model. The lower recall can be explained by the fact that when sometimes a color layer that is not part of the object is emphasized in the bounding box by the coefficient of interest, leading to an erroneous boundary selection.

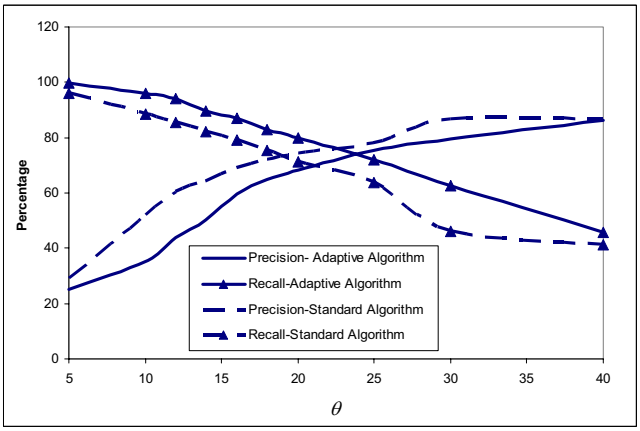

***Figure 5**: Comparison of Precision and Recall for the standard and the adaptive algorithms.*

This graph also shows that how the precision increases and the recall decreases as the value of the threshold increases.

5.5. Robustness

Finally, we studied how these results vary when the initial point selected by the user is not exactly at the center of the reference bounding box. For this purpose, we generated a number of initial points randomly within the bounding box, but we kept only those

within the object or very close to it (the bounding box may sometimes contain large areas outside the object). The results are given in the following table:

	Precision	Recall
Optimal user-selected pixel	59.7%	86.9%
Randomly user-selected pixel	61.9%	78.5%

Table 1*: Comparison between centered and random initial point.*

These results indicate that there is only a slight degradation when the point selected by the user is not exactly at the center of the bounding box. In the cases where our algorithm fails to recover the correct bounding box, we identified three types of problems. First, the layer segmentation is sometimes incorrect, leading to background zones which are very similar to object zones. Second, sometimes the user-selected point is close to a distractor (a background zone with a high saliency, which causes a large portion of the background to be included in the bounding box). Third, the thresholds are sometimes not optimal for a particular zone, and produce a bounding box which is too small or too large.

However, we have found that our algorithm is also very often able to do a reasonable job at selecting a reasonable bounding box. Our next step is to include this algorithm into our annotation workstation and experiment with users to evaluate the practical impact on the semi-automated annotation of video programs.

6. CONCLUSION

With the development of multimedia technologies, we expect that semi-automated annotation of video material is going to be a crucial bottleneck for the development of effective applications such as Interactive Television in the near future. In this paper, we have presented some research to speed up the process of selecting a video object in an image. Based on an attention model, the user may only select one point inside the object, and our algorithm proposes a reasonable bounding box. We have presented several experimentations, together with an adaptive version of our algorithm.

Our algorithm provides reasonable performance, yet still fails in some cases, for example in the case of the user-selected point close to a distractor. Our further objective is to improve this framework, and to integrate it into an annotation workstation for evaluation with real users.

7. REFERENCES

[1] B. Cardoso, F. de Carvalho, L. Carvalho, G. Fernàndez, P. Gouveia, B. Huet, J. Jiten, A.López, B. Merialdo, A. Navarro, H. Neuschmied, M. Noé, R. Salgado, G. Thallinger, "Hyperlinked Video with Moving Objects in Digital Television", *IEEE International Conference on Multimedia & Expo* July 6-8, 2005, Amsterdam, The Netherlands.

[2] Wen-Huang Cheng, Wei-Ta Chu, and Ja-Ling Wu, "A Visual Attention based Region-of-Interest Determination Framework for Video Sequences," *IEICE Transactions on Information and Systems Journal*, vol. E-88D, no. 7, pp. 1578-1586, 2005. (SCIE, EI)

[3] Junwei Han; Ngan, K.N, "Automatic segmentation of objects of interest in video: a unified framework", Intelligent Signal Processing and Communication Systems, pp 375 – 378, Nov 2004.

[4] Chia-Chiang Ho, Wen-Huang Cheng, Ting-Jian Pan, and Ja-Ling Wu, "A User-Attention Based Focus Detection Framework and Its Applications," *The Fourth IEEE Pacific-Rim Conference on Multimedia*, 15-18 December, 2003, Singapore, pp. 1341-1345, 2003.

[5] L. Itti, C. Gold, C. Koch, "Visual Attention and Target Detection in Cluttered Natural Scenes", *Optical Engineering,* Vol. 40, No. 9, pp. 1784-1793, Sep 2001.

[6] L. Itti, C. Koch, E. Niebur, "A Model of Saliency-Based Visual Attention for Rapid Scene Analysis", *IEEE Transactions on Pattern Analysis and Machine Intelligence,* Vol. 20, No. 11, pp. 1254-1259, Nov 1998.

[7] C. Koch and S. Ullman "Shifts in selective visual attention: towards the underlying neural circuitery", *Hum neurobiol*, Vol 4, pp 219-227, 1985.

[8] Y-F. Ma, L. Lu, H-J. Zhang, and M. Li, "A user attention model for video summarization", *in proc. ACM Multimedia (ACMMM'02)*, pp. 533-542, Dec. 2002.

[9] Osberger, W. Maeder, A.J., "Automatic identification of perceptually important regions in an image", *in proc Pattern Recognition*, Vol 1,pp: 701-704, Aug 1998.

[10] C. M. Privitera, Lawrence W. Stark "Algorithms for Defining Visual Regions-of-Interest: Comparison with Eye Fixations", *in proc IEEE Transactions on Pattern Analysis and Machine Intelligence*, Vol 2*2,* No 9, pp 970-982, 2000.

[11] L. Stelmach, W.Tam, and P hearty, "Static and dynamic spatial resolution in image coding: an investigation of eye movements", *in proc SPIE 1453*, San Jose , pp 147-152, Feb 1992.

[12] F. W. M. Stentiford, "An estimator for visual attention through competitive novelty with application to image compression," *Picture Coding Symposium*, Seoul, 24-27 April, 2001.

[13] K. N. Walker, J. C. Taylor, "Locating Salient Object Features", British Machine Vision Conference, 1998.

[14] Wolfe J. In: De Valois, "Visual Attention", *KK, editor. Seeing. 2nd ed, CA: Academic Press*, San Diego, pp. 335-386, 2000.

[15] Wolfe, J.M. "Guided Search 2.0: A Revised Model of Visual Search", *Psychonomic Bulletin & Review*, Vol 1, No 2, pp 202-238, 1994.

[16] Wolfe, J.M, "Guided Search 4.0: Current Progress with a model of visual search", to appear.

[17] A Yarbus, "Eye movement and vision", *Plenum Press*, New York, 1967.