AN INFORMATION THEORETIC MODEL OF SPATIOTEMPORAL VISUAL SALIENCY

Guoping Qiu[1+], *Xiaodong Gu*[2], *Zhibo Chen*[2], *Quqing Chen*[2] *and Charles Wang*[2]

[1]The University of Nottingham and Hong Kong Baptist University
[2]Thomson Corporate Research, Beijing

ABSTRACT

This paper presents a principled and practical method for the computation of visual saliency of spatiotemporal events in full motion videos. Based on the assumption that uniqueness or informative-ness correlates with saliency, our model predicts the saliency of a spatiotemporal event based on the information it contains. To compute the uniqueness of the spatiotemporal events, we model the joint spatial and temporal conditional probability distributions of the spatiotemporal events and compute their spatiotemporal saliencies in a natural and integrated framework. To make the information theoretic model practical, we have developed methods to simplify the model and computational process. Testing results on several video sequences demonstrate that our model is effective in predicting visually salient spatiotemporal events and is comparable to state of the art. It is expected that our principled and practical model will find widespread applications in multimedia content analysis and processing.

1. INTRODUCTION

Salient visual features and motions that attract human attention can be important and powerful cues for visual information analysis and processing, including content-based coding, compression, transmission/rate control, indexing, browsing, display and presentation. Whilst there have been a fair amount of previous work investigating the detection and extraction of visually salient features in still images [1-3], not much has been done to address the same problem in video sequences. Although several authors have made some initial attempts, e.g., [4, 5], most of these methods are rather ad hoc. A common approach of these previous attempts is that they first compute the spatial and temporal saliencies independently and then fuse them in some rather arbitrary manners. For spatial saliency, these methods manipulate the contrasts of various visual features (intensity, colour, texture, etc.) in some heuristic ways to generate the numerical scores of saliency. For temporal saliency, these methods assume that it is somehow related to visual motion. They first detect motions based on some known motion detection methods and then compute the temporal saliency scores as some heuristically chosen functions of motion vectors.

Ad hoc and heuristic spatial and temporal visual saliency computational methods can have some serious weaknesses. First, motion estimation is difficult. Second, even with accurate true motion estimation, it is not clear how exactly motion relates to visual saliency and a quantitative measure that relates motion with visual saliency is difficult to obtain. Third, assuming spatial and temporal saliencies can be correctly computed, it is not clear how they should be fused together. In most cases previous methods gave arbitrary weightings to temporal and spatial saliencies to obtain a final score of spatiotemporal saliency.

Whilst most of the spatial visual saliency models are based on the calculation of visual contrasts, e.g, [1], Topper [6] introduced the idea of using Shannon's self-information to measure the perceptual saliency in still images. Recently, Bruce [2] picked up Topper's original suggestion and studied the relationship between visual saliency and local statistics, again, in still images.

In this paper, we extend the information theoretic approach to visual saliency computation to full motion video and have developed a principled and integrated practical method for computing spatiotemporal saliency. The organization of the paper is as follows. In section 2, we describe the integrated spatiotemporal visual saliency model. In section 3, we present a practicable computational method for the implementation of the model. In section 4, we present experimental results and section 5 concludes the paper.

2. AN INFORMATION THEORETIC MODEL OF SPATIOTEMPORAL VISUAL SALIENCY

We wish to develop a computational model for predicting the visual saliency of the pixels in each frame of a video. In this work, we adopt an information theoretic approach and assume that the "uniqueness" or "informative-ness" correlates with visual saliency. According to Shannon's information theory, an event contains high information if it is unique; on the other hand, it contains low information if the event occurs frequently. The information of an event x,

[+]This work was done when the first author was visiting Thomson Corporate Research Beijing

$I(x)$, is inversely proportional to the likelihood of observing x, Shannon defined this quantity as

$$I(x) = -\log(p(x)) \quad (1)$$

Clearly, according to (1), the computational task now becomes that of finding the probability distributions of the pixels. Before formally presenting our model, we first analyze what factors are likely to affect the uniqueness of the pixels and how we can go about computing the probability distributions of pixels in the context of a full motion video.

Instead of trying to compute the visual saliency of each individual pixel, we divide each frame into small patches and we call a patch in a particular frame a *spatiotemporal event*. The reason any given patch in any given frame is a spatial and temporal event is because it has a spatial as well as a temporal context, as illustrated in Figure 1. For a given spatiotemporal event, $B(x, y, t)$, its spatial context is its current frame $F(t)$; its temporal context is defined as the spatiotemporal events with the same spatial location in the previous N-1 frames, $V(x, y, t\text{-}1)=\{B(x, y, t\text{-}1), B(x, y, t\text{-}2), \ldots, B(x, y, t\text{-}N+1)\}$.

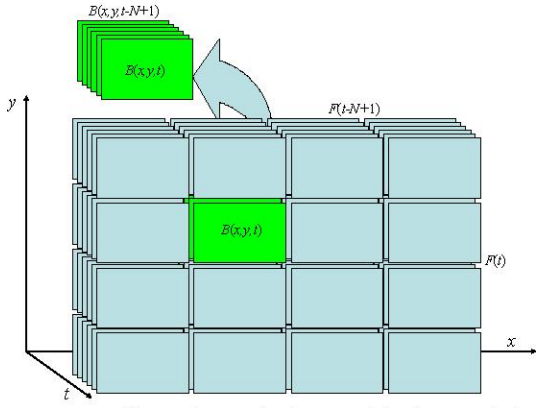

Figure 1: An illustration of the spatial time relations of spatiotemporal events. A spatiotemporal event is an $m \times n$ block (patch) of pixels, $B(x, y, t)$, located at spatial co-ordinate (x, y) in the frame at temporal (time) location t, $F(t)$. N patches at spatial locations (x, y) from frame $F(t)$, $F(t\text{-}1)$, …, $F(t\text{-}N+1)$ form a space-time 3D patch volume $V(x, y, t)=\{B(x, y, t), B(x, y, t\text{-}1), \ldots, B(x, y, t\text{-}N+1)\}$, which records how the patch located at (x, y) evolves over time.

The uniqueness of a spatiotemporal event $B(x, y, t)$ is clearly affected by its spatial and temporal contexts. If an event is unique in the spatial context, it is likely that it is salient; similarly, if it is unique in the temporal context it is also likely to be salient. Obviously, the spatial and temporal contexts jointly influence the uniqueness of a spatiotemporal event. The task now becomes that of modeling these spatial and temporal contexts to predict the uniqueness or saliency of the spatiotemporal events.

2.1 The Model

Based on the assumption that saliency is directly related to uniqueness or informative-ness, the spatiotemporal saliency score of the spatiotemporal block $B(x, y, t)$, $SSS(x, y, t)$ can be modeled by the amount of information contained in $B(x, y, t)$. According to Shannon information theory (1), the information of the spatiotemporal event $B(x, y, t)$ given its spatial context $F(t)$ and temporal context $V(x,y,t\text{-}1)$ can be computed as

$$SSS(x, y, t) = -\log(p(B(x, y, t) \mid V(x, y, t-1), F(t))) \quad (2)$$

Equation (2) is our new information theoretic model of spatiotemporal visual saliency which integrates spatial and temporal information for predicting the visual saliency of spatiotemporal events in a principled, concise and elegant manner.

The model is concise and elegant, however, it is important to point out that its practical realization is extremely challenging. One of the reasons that information theory has not been widely used in the analysis of video and other very high dimensional data is that it involves some extremely difficult computational tasks. The difficulty primarily stems from the "curse of dimensionality". From (2), our task of computing the saliency is to estimate the conditional probability. The dimensionality of the variables involved in the computation is usually high and the space these variables reside grows exponentially with each new dimension added. As the dimensionality increases, more and more samples are needed to estimate the probability. For example, for distributions in R^6, even 2 million samples are not sufficient to approximate the entropy correctly [7]. Furthermore, we not only have to face the "curse of dimensionality" problem, but also have to make the computational process reasonably fast in order to make the model practically useful for applications such as real time video analysis and processing. In the next section, we present a practicable computational method for the implementation of our new spatiotemporal visual saliency model.

3. COMPUTATIONAL METHODS

3.1 Model Simplification

The conditional probability of the spatiotemporal event $B(x, y, t)$ has two conditions, spatial condition $F(t)$ and temporal condition $V(x, y, t\text{-}1)$. It is not difficult to see that it would be very difficult to estimate this probability. In order to develop practical solutions, we simplify the model by assuming that the spatial and temporal conditions are independent. Based on this assumption, we can write the joint conditional probability as

$$p(B(x, y, t) \mid V(x, y, t-1), F(t)) = p(B(x, y, t) \mid V(x, y, t-1)) p(B(x, y, t) \mid F(t)) \quad (3)$$

With (3), we can now compute two conditional probabilities of the spatiotemporal event, one for the spatial condition and the other for the temporal condition. Even though (3) has simplified the model somewhat, because the variables involved are still of very high dimension, the task of estimating these two conditional probabilities still poses some serious computational difficulties.

3.2 Temporal Conditional Probability

We can write

$$p(B(x,y,t)|V(x,y,t-1)) = \frac{p(B(x,y,t),V(x,y,t-1))}{p(V(x,y,t-1))} = \frac{p(V(x,y,t))}{p(V(x,y,t-1))} \quad (4)$$

From (4), we see that the temporal conditional probability of the spatiotemporal event $B(x,y,t)$ can be computed by estimating the probability of the 3D space-time patch volumes $V(x,y,t)$ and $V(x,y,t-1)$. Consider a 4x4 patch and a temporal context of 2 frames (N=2), then $V(x,y,t) \in R^{32}$, estimating the probability in the 32D space will be impracticable requiring prohibitive large amount of data and computational effort.

To get round this computational difficulty and to make our computational task tractable, we can try and transform our high dimensional spatiotemporal event vectors into a space where each dimension is independent. The independence in each dimension allows us to decompose the multidimensional probability estimation problem into that of estimating the distributions of multiple independent 1D random variables.

Computationally, this can be done by performing independent component analysis (ICA) as in [2]. However, ICA is an extremely under constrained problem and computationally too expensive for real time video analysis. Another way to overcome the computational difficulty is by simplifying the assumption that the distributions of the spatiotemporal events is approximately normally distributed, in which case, we can transform the events into a space where each dimension is uncorrelated. Given the normal assumption, uncorrelated implies independence. In the signal processing literature, there exist a number of orthogonal transforms which can be used to project high dimensional data into uncorrelated spaces. Principal Component Analysis (PCA) is one possibility. Although PCA is optimal in the mean square error sense, its transform bases are data dependent and it can be computationally expensive. Instead, we use a well-known data independent orthogonal transform, the discrete cosine transform (DCT), extensively used in image and video coding, to transform our spatiotemporal events and volumes into uncorrelated space for the purpose of estimating the spatiotemporal events distributions.

Let ϕ_k, k = 1, 2, … K, be the K orthogonal transform bases, the procedure for computing $p(V(x, y, t)$ follows these steps:

Step 1: $c_k(x,y,t) = \phi_k V(x,y,t) \quad \forall x, y$

Step 2: Compute the probabilities $p_k = p(c_k(x, y, t)), \forall k$

Step 3: Compute $p(V(x,y,t)) = \prod_k p_k$

With this procedure, we can compute $p(V(x, y, t)$ and $p(V(x, y, t\text{-}1)$, which in turn enables us to compute the temporal conditional likelihood of the spatiotemporal event $B(x, y, t)$ according to equation (4).

3.3 Spatial Conditional Probability

It is important to note that the spatial context that will influence the uniqueness of a spatiotemporal event $B(x,y,t)$ is the current frame $F(t)$ because only the current frame and $B(x,y,t)$ will be simultaneously visible to the viewers. The probability $p(B(x,y,t)|F(t))$ is therefore equivalent to $p(B(x,y,t))$. With this, we only need to estimate the distribution of the spatiotemporal event $B(x,y,t)$ against all the events in $F(t)$. We can again resort to the simplification assumption and technique of sub-section 3.2 to perform the estimation. The procedure can therefore proceed in the following steps

Step 1: $c_k(x,y,t) = \phi_k B(x,y,t) \quad \forall x, y$

Step 2: Compute the probabilities $p_k = p(c_k(x, y, t)), \forall k$

Step 3: Compute $p(B(x,y,t)) = \prod_k p_k$

3.4 The Spatiotemporal Saliency Score

Follow (2) and (3) the spatiotemporal saliency score of the spatiotemporal event $B(x,y,t)$ can be computed as

$$\begin{aligned} SSS(x,y,t) &= -\log(p(B(x,y,t)|V(x,y,t-1),F(t))) \\ &= -\log(p(B(x,y,t)|V(x,y,t-1))) - \log(p(B(x,y,t)|F(t))) \\ &= S_t(x,y,t) + S_s(x,y,t) \end{aligned} \quad (5)$$

where $S_t(x,y,t)$ and $S_s(x,y,t)$ are temporal and spatial saliency scores respectively.

Note in (5) that the spatiotemporal saliency in our model can also be written as the sum of spatial and temporal components. However, unlike previous methods, where these components are computed independently first and then fused together in some arbitrary ways, in our model, this decomposition is natural and derived from the joint spatiotemporal saliency in a principled and naturally way.

4. EXPERIMENTAL RESULTS

To evaluate the effectiveness of our model, we have implemented the model using C++ program and tested on

several video sequence. The video data format used in our experiment is CIF (352x288 pixels per frame). For spatiotemporal event patch size = 4 x 4, and a temporal context N = 2, using all DCT coefficients except the DC component for the estimation of the probability distributions, and without any code optimization, our implementation is real time on a Pentium PC. This shows that from a computational perspective our method is practical.

A quantitative evaluation of results is difficult. Here we present some visual examples. We have also conducted experiments using data from an eye tracking system and results are competitive with other state of the art methods.

Figure 2 shows the saliency maps of some particular frames of two sequences used in the experiments, also shown are the spatial and temporal components of the saliency.

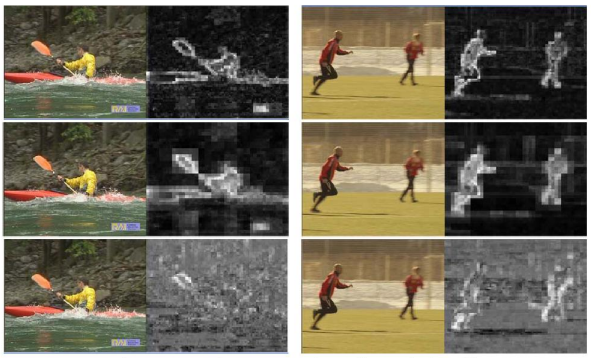

Figure 2: Spatiotemporal, temporal and spatial saliency scores of two example sequences. Left: 118th frame of the Canoe sequence. Right: 125th frame of the Soccer sequence. Top row: The spatiotemporal saliency score (SSS). Middle: The spatial component of the SSS (S_t). Bottom: The temporal component of the SSS (S_s). Note that in both sequences, the spatial saliency scores are rather noisy while the temporal saliency scores are relatively clean.

Figure 3 shows a comparison of our model with the method in [1]. It is seen that our method performs similarly or better than that in [1]. More example result of our method is shown in Figure 4.

5. CONCLUDING REMARKS AND FUTURE WORK

In this work, we have developed an information theoretic based spatiotemporal visual saliency model for predicting the visual saliency of spatiotemporal events in full motion video. Contrast to previous ad hoc approaches, we derive our concise and elegant model in a principled way guided by information theory. To overcome the inherent computational difficulties associated with the use of information theory, we have developed fast and practical computational solutions for the implementation of our model. Experimental results further demonstrate the effectiveness and practicability of our new method.

Our model is currently being applied to various multimedia processing tasks including video summarization and object tracking.

Figure 3: Comparison with the method in [1] (based on our own implementation of the method in [1] for video). Left: original frame, Middle: top 20 salient regions detected by the method of [1], Right: top 20 salient regions detected by our method. Note in the figure, for each salient patch, pixels within a radius of 69 pixels are shown.

Figure 4: More examples. From left to right: original frame image, top 20 salient regions (for each salient patch, pixels within a radius of 69 pixels are shown), another original frame, top 20 salient regions.

6. REFERENCES

[1] L. Itti, C. Koch, E. Niebur, "A Model of Saliency-Based Visual Attention for Rapid Scene Analysis," IEEE Trans. On Pattern Analysis and Machine Intelligence, 1998

[2] Neil D.B. Bruce, "Features that draw visual attention: an information theoretic perspective", Neurocomputing, vol.65-66, pp.125–133, 2005

[3] T Kadir and M Brady. "Scale, saliency and image description", IJCV, 45(2):83–105, November 2001

[4] Y-F Ma, et al, "A User Attention Model for Video Summarization", Proceedings of ACM Multimedia, 2003

[5] Y. Zhai and M. Shah, "Visual Attention Detection in Video Sequences Using Spatiotemporal Cues", Proceedings of ACM Multimedia, 2006

[6] T.N. Topper, "Selection mechanisms in human and machine vision", Ph.D. Thesis, University of Waterloo, 1991

[7] D.B.Russakoff, C.Tomasi, T.Rohlfing and C.R.Maurer, Jr., "Image Similarity Using Mutual Information of Regions", ECCV (3) 2004:596-607