Signal Processing 110 (2015) 143–154

Contents lists available at ScienceDirect

Signal Processing

journal homepage: www.elsevier.com/locate/sigpro

# An audio–visual human attention analysis approach to abrupt change detection in videos

Yanxiang Chen, Minglong Song, Lixia Xue, Xiaoxue Chen, Meng Wang*

*School of Computer Science and Information Engineering, Hefei University of Technology, Hefei, China*

ARTICLE INFO

*Article history:*
Received 5 April 2014
Received in revised form
25 July 2014
Accepted 3 August 2014
Available online 19 August 2014

*Keywords:*
Bayesian surprise
Abrupt changes detection
Unsupervised
Audio–visual
Time-synchrony

ABSTRACT

Abrupt changes in videos, such as sudden running, usually indicate abnormal events and play a significant role in attracting human attention. We propose an approach to detect abrupt changes in videos based on Bayesian surprise theory, which considers both visual and audio modalities. Specifically, after generating surprise curves from the audio and visual modalities, we obtain a synchronized sequence based on the time-synchrony between audio–visual series in videos. The approach is fully automated and does not require any prior information. Experimental results from tests on human behavior and natural scene video datasets demonstrate that the proposed method is able to detect abrupt changes like sudden running or the collapse of an object. The proposed approach is further evaluated on the entire dataset we collected.



## 1. Introduction

Currently, many video surveillance systems to ensure public safety are widely used in locations such as supermarkets, underground parking garages, banks, and so on. Additionally, many researchers have focused on studies that rely on supervised or prior information [1–5]. However, it is a profound and meaningful task for researchers to analyze these videos without any priori information; that is, the analysis must be done in an unsupervised mode, a requirement that can be satisfied by surprise.

Surprise is an interesting and amazing concept. Each day is full of surprises, ranging from car crashes, and sudden loud noises, amongst others. Thus, how to detect surprising events in videos has become an increasingly hot topic. As we know, the human brain can quickly react to surprises and select a suitable response. However, proposing methods and criteria to identify or recognize surprising events in the field of computer vision remains a key question. In fact, the detection of surprising events in video scenes is a stimulus-driven, bottom-up rather than task-dependent, top-down process [6–9].

Several approaches have been proposed to characterize the potential behavioral importance or surprise of visual stimuli computationally [10,11]. Itti introduced the saliency model [12] in 1998, and improved the method in 2001 and 2005 [13,14]. In this model, a saliency map is computed by combining different features' channels including intensity, color, orientation, flicker, and motion. That is, we choose these five multi features as the representation of the picture, furthermore there are many other multiple features or modalities for pictures or information representation [15,16]. However, a saliency map only reflects the most interesting parts in pictures or videos [17], which are not always the surprising parts.

To complement this spatial definition of importance using a saliency model, the novelty model [18,19] emphasizes the temporal dimension. Novelty of a stimulus is defined as the degree to which its visual appearance does not fit the statistics of previously observed image samples

* Corresponding author. Tel.: +86 551 62904883.
*E-mail address:* eric.mengwang@gmail.com (M. Wang).

http://dx.doi.org/10.1016/j.sigpro.2014.08.006
0165-1684/

at a given location [20,21]. In more detail, novelty detection starts by assuming a model for the data, and then new data samples are evaluated against the current model. If the probability of the observed data is low given the model, the pixel is labeled as containing a novel stimulus and the same data samples are then used to adapt the models parameters.

The novelty model resembles saliency in time, while the saliency model is somewhat similar to novelty over space. Itti developed a theory that resolves this duality and incorporated it into a principled Bayesian notion of surprise model [22–24], which achieved good results from video experiments. In the Bayesian surprise model, surprise is defined in general Bayesian terms as the difference between a prior belief distribution and a posterior distribution that is generated when new data is observed. Surprise is then measured as the average of the log-odd ratio or KullbackLeibler (KL) divergence. For each input frame [25,26], the surprise model produces a master surprise map, which is generated by combining local temporal and spatial surprise maps.

As we know, abrupt changes in aural signals can also be viewed as salient events and often play a significant role in attracting human attention [27,28]. However, most current methods consider only visual information, ignoring the aural information. To make better use of the audio information in video, we produce a unified framework to detect abrupt changes by taking both modalities into account. Traditionally, audio features such as MFCC [29] have been widely used in the field of audio attention processing. In [30] Maragos proposed the following three components, maximum average Teager energy (MTE), mean instant amplitude (MIA), and mean instant frequency (MIF), to account for time-varying amplitude and frequency of audio signals. However, how to apply the above parameters to locate abrupt changes in audio signals precisely still needs to be further studied. Besides, integrating audio signals and visual sequences also can be viewed as

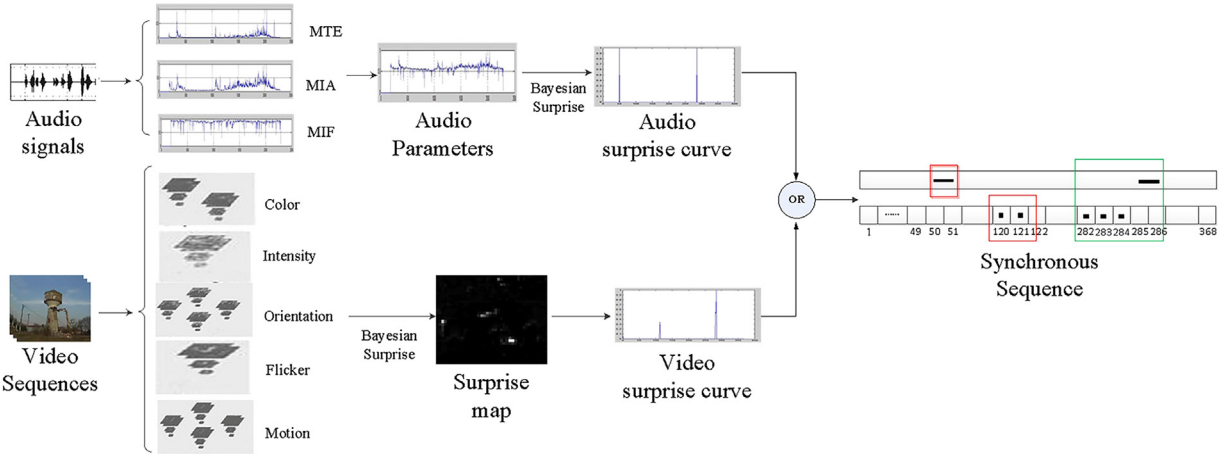


**Fig. 1.** Schematic of the proposed audio–visual fusion criterion. Starting with the audio and visual sequences, we extract three audio features (MTE, MIA, and MIF) and five low-level visual features (intensity, color, orientation, flicker, and motion). Bayesian surprise theory is then applied to the selected audio and visual features and the two surprise curves are obtained. Finally, the time-synchrony sequence is created by computing the logical *OR* of the audio and video surprise curves.

**Table 1**
List of important notations.

| Notation | Description |
|---|---|
| $\mathcal{M}$ | The model space |
| $D$ | The new data observation |
| $P(M)_{M\in\mathcal{M}}$ | The prior probability distribution of models $M$ |
| $P(M\|D)_{M\in\mathcal{M}}$ | The posterior probability distribution of models $M$ |
| $\lambda$ | The rate parameter, which is estimated as the detector's output value |
| $M(\lambda)$ | The given data are modeled as Poisson distributions |
| $\alpha,\alpha'$ | The shape parameters of gamma distribution |
| $\beta,\beta'$ | The scale parameters of gamma distribution |
| $\gamma(\lambda;\alpha,\beta)$ | The gamma distribution |
| $\zeta$ | The forgetting factor which ranges from 0 to 1 |
| $\Gamma(\cdot)$ | Euler gamma function |
| $\Psi(\cdot)$ | The digamma function |
| $K$ | The number of audio formants |
| $\Psi_d(\cdot)$ | Teager energy operator |
| $r_k(n)$ | The $k$th formant modulation signal |
| $a_k(n)$ | The amplitude of the $k$th formant modulation signal |
| $f_k(n)$ | The frequency of the $k$th formant modulation signal |
| $[\omega_1,\omega_2,\omega_3]$ | The weighting vector of MTE, MIA and MIF |