A User Attention Model for Video Summarization

Yu-Fei Ma, Lie Lu, Hong-Jiang Zhang and Mingjing Li
Microsoft Research Asia
3F, Beijing Sigma Center, 49 Zhichun Road
Beijing 100080, China
{yfma, llu, hjzhang, mjli}@microsoft.com

ABSTRACT
Automatic generation of video summarization is one of the key techniques in video management and browsing. In this paper, we present a generic framework of video summarization based on the modeling of viewer's attention. Without fully semantic understanding of video content, this framework takes advantage of computational attention models and eliminates the needs of complex heuristic rules in video summarization. A set of methods of audio-visual attention model features are proposed and presented. The experimental evaluations indicate that the computational attention based approach is an effective alternative to video semantic analysis for video summarization.

Keywords
Video summarization, attention model, skimming, video content analysis

1. INTRODUCTION
One of the key technologies required for efficient management of video data is automatic video content summarization. A concise and informative video summary enable a user to quickly figure out the overview contents of a video and decide whether the whole video program is worth watching. To generate a perfect summarization of a given video requires good understanding of video semantic content. However, automatic understanding of the semantic content of general videos is still far beyond the intelligence of today's computing systems, despite the significant advances in computer vision, image processing, pattern recognition, and machine learning algorithms.

Video summarization offers a concise representation of the original video clips by showing the most representative synopsis of a given video sequence. Generally speaking, there are two fundamental types of video summarization: static video abstract and dynamic video skimming. A static abstract, also known as a static storyboard, is a collection of salient images or key-frames extracted from the original video sequence. A dynamic skimming consists of a collection of associated audio-video sub-clips selected from a video sequence, but with much shortened length.

A well known approach to extracting multiple key-frames from shot is based on frame content changes computed by features, such as color histogram [1] or motion activity [2]. Zhuang proposed an unsupervised clustering scheme to adaptively extract key-frames from shots [3]. These methods require a pre-defined threshold or key-frame number to control the density of key-frames in a shot. Recently, some shot-independent approaches are also proposed. For instance, Girgensohn proposed a time-constrained clustering approach to key-frame extraction from an entire video [4]. Other more sophisticated methods include the integration of the motion and spatial activity analysis with face detection technologies [5], a progressive multi-resolution key-frame extraction techniques [6], and object-based approach [7].

Static video summary, although effective, often cannot preserve the time-evolving dynamic nature of video content. Moreover, the audio track is lost, which is an important content channel of video. A skimmed sequence, on the other hand, is able to provide users a more impressive preview of an entire video. Many literatures have addressed that dynamic video skimming is an indispensable tool for video browsing. One of the most straightforward approaches is to compress the original video by speeding up the playback [8]. However, the abstract factor in this approach is limited by the playback speed in order to keep the speech comprehensible. The InforMedia system [9] generates short synopsis of video by integrating audio, video and textual information. By combining language understanding techniques with visual feature analysis, this system gives reasonable results. However, satisfactory results may not be achievable by such a text-driven approach when speech signals are noisy, which is often the case in life video recording. Another approach to generating the semantically meaningful summaries is event-oriented abstraction scheme, such as the one presented in [10]. Recently, more sophisticated techniques have also been proposed. For example, the trajectories of objects are used in [11]. The linear dynamical system theory is applied in [12]. Singular value decomposition is adopted to summarize video content [13].

Despite numerous efforts in generating static or dynamic video summary, the results are still far from satisfactory. The direct sampling or low level feature based approaches are inconsistent with human perception. The semantic oriented methods are, in general, far from meeting human expectations, since precise textual information may not be always available for summarization. Moreover, those systems that totally neglect audio track are not able to generate impressive results. Also, the algorithms involving large number of summarization rules or over-intensive computation are normally impractical to real

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee.
Multimedia'02, December 1-6, 2002, Juan-les-Pins, France.
Copyright 2002 ACM 1-58113-620-X/02/0012…$5.00.

applications.

Aiming at removing the limitation of current algorithms while avoiding fully semantic understanding, we have developed a generic video summarization framework based on modeling of viewer's attention. The main contributions of this work are twofold: a new scheme to model viewer attention in viewing video, and a set of algorithms to apply such modeling in video summarization.

Attention is a neurobiological conception. It implies the concentration of mental powers upon an object by close or careful observing or listening, which is the ability or power to concentrate mentally. Computational attention allows us to break down the problem of understanding a live video sequence into a series of computationally less demanding and localized visual, audio, and linguistic analytical problems.

The computational attention methodologies have been studied and become a powerful tool in active vision systems. Itti *et al*. have reviewed their recent works on computational models of focal visual attention, and presented a bottom-up, saliency- or image-based visual attention system [14]. By combining multiple image features into a single topographical saliency map, the attended locations are detected in the order of decreasing saliency by a dynamical neural network [15, 16]. In [17], Salah *et al*. presented a selective attention-based method for visual pattern recognition. Promising results were obtained when applying this method in handwritten digit recognition and face recognition. In [18], Bollmann *et al.* proposed another attention system to control the gaze shifts of an active vision system by integrating with dynamic features, motion-induced attention. In our previous work, we also proposed a computational motion attention model, which has been used to generate video skimming [19].

In this paper, we extend our work presented in [19] to build up a generic user attention model by integrating a set of attention models extracted from video sequence. Among them, the static attention model is the reasonable extension of the work in [15] according to video characters. With such video attention model, video summarization can be generated based on the user attention curve resulted from the attention modeling. Since each frame is assigned an attention value, it is easy to determine which frame or which segment of frames is more likely to attract viewer's attention. In this way, the number of needed key-frames in a shot is determined by the number of wave crest on attention curve. The dynamic skims are also extracted according to the wave crests of attention curve without the need of sophisticated rules.

The rest of this paper is organized as follows: Section 2 introduces the framework of user attention model. Section 3 and Section 4 discuss the attention modeling methods of visual and audio features respectively. A new video summarization approach, including both static and dynamic summarization, based on the user attention model is described in Section 5. The performance of summarization approaches is evaluated in a user study experiment. Section 6 describes the experiment method and reports the experimental results. Finally, Section 7 presents the concluding remarks and discussions.

2. USER ATTENTION MODEL

Video is a compound of image sequence, audio track, and textural information. The image sequence presents motion (object motion and camera motion), color, texture, shapes and text regions. The audio channels consist of speech, music, and various special sounds. The textual information which is represented in linguistic form can be obtained from three sources: closed caption, automatic speech recognition (ASR), and superimposed text. Since human attention is always attracted by these information elements, a complete user attention model should be an integrated set of visual, audio, and linguistic attentions.

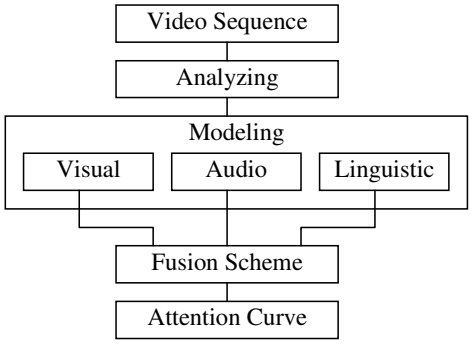

Figure 1. Framework of user attention model

Figure 1 presents the proposed framework of user attention model. To obtain this model, the video sequence is analyzed to extract all kinds of content features first. Then, a set of attention modeling methods are employed to generate the attention models of three information channels. Any extractable video, audio, and linguistic features can be integrated into such user attention framework, if its computational attention model is available. Therefore, this is an extendable framework. Fusion scheme is another key issue of this framework. Linear combination, user interaction, or machine learning based method can all be adopted. In the fusion process, attention models are integrated to generate a user attention curve for a given video sequence.

Currently, we adopt a linear combination to implement fusion scheme due to its effectiveness and simplicity. With such a scheme, each attention model should be normalized to [0~1]. Let *A* denote attention model, it can be computed as

$$A = w_v \cdot \overline{M_v} + w_a \cdot \overline{M_a} + w_l \cdot \overline{M_l} \quad (1)$$

where w_v, w_a, w_l are the weights for linear combination, and $\overline{M_v}$, $\overline{M_a}$, $\overline{M_l}$ are normalized visual, audio, and linguistic attention models, respectively, which are defined as follows:

$$M_v = \left(\sum_{i=1}^{p} w_i \cdot \overline{M_i} \right) \times \left(\overline{M}_{cm} \right)^{S_{cm}} \quad (2)$$

$$M_a = \left(\sum_{j=1}^{q} w_j \cdot \overline{M_j} \right) \times \left(\overline{M}_{as} \right)^{S_{as}} \quad (3)$$

$$M_l = \sum_{k=1}^{r} w_k \cdot \overline{M_k} \quad (4)$$

where w_i, w_j, w_k are weights in visual, audio, and linguistic attention models respectively; and $\overline{M_i}$, $\overline{M_j}$ and $\overline{M_k}$ are the normalized attention model components in each attention model. $\overline{M_{cm}}$ is the normalized camera attention, which is used as visual attention model's magnifier. S_{cm} works as the switch of magnifier.

If S_{cm}>=1, the magnifier is open; while S_{cm}=0, the magnifier is closed. The higher the S_{cm} value is, the more powerful the magnifier is. Similar to camera attention, $\overline{M_{as}}$ is normalized audio saliency attention, which is also used as a magnifier of audio attention. As magnifiers, M_{cm} and M_{as} are all normalized to [0~2]. In the definition of attention models (1~4), all weights are used to control the user's preference to the corresponding channel. They can be adjusted automatically or interactively.

Since the user attention model presented in this paper is an extensible framework, any computational visual, audio, or linguistic attention model can be integrated into this framework. In this paper, we only discuss the modeling methods of some of the most salient audio-visual features to demonstrate the effectiveness of proposed user attention model and their applications in video summarization. The details of modeling methods and summarization scheme will be presented in the following sections. The linguistic attention modeling method is not discussed in this paper.

3. VISUAL ATTENTION MODELING

In this section, we discuss in detail how to model visual attentions. In an image sequence, there are many visual features, including motion, color, texture, shape, text region, etc. However, all these features can be classified into two classes: dynamic and static features. Also, some recognizable objects, such as face, will more likely attract human attention. Besides, camera operations are often used to induce reviewer's attention. Therefore, visual attention models are proposed to model the visual effects due to motion, static, face, and camera attention.

3.1 Motion Attention Model

Motion attention model is built based on motion vector field (MVF). That is, for a given frame in a video sequence, we extract the motion field between the current and the next frame and calculate a set of motion characteristics. In our implementation, all video sequences are stored as MPEG format, so MVFs are extracted from MPEG data directly.

If we treat MVF as the retina of eyes, the motion vectors will be the perceptual response of optic nerves. We assume that a MVF has three inductors: *Intensity inductor*, *Spatial Coherence inductor*, and *Temporal Coherence inductor*. When the motion vectors in a MVF go through such inductors, they will be transformed into three kinds of maps. Then, these normalized outputs of inductors are fused into a saliency map by linear combination. In this way, the attended regions can be detected from saliency map image by image processing methods.

According to our basic assumption, there will be three inductors at each location of macro block $MB_{i,j}$. The *Intensity Inductor* induces motion energy or activity, called motion intensity $\boldsymbol{I}$, and is computed, namely, as the normalized magnitude of motion vector,

$$\boldsymbol{I}(i,j) = \sqrt{dx_{i,j}^2 + dy_{i,j}^2} \Big/ MaxMag \tag{5}$$

where ($dx_{i,j}$, $dy_{i,j}$) denote two components of motion vector, and *MaxMag* is the maximum magnitude in a MVF.

The *Spatial Coherence Inductor* induces the spatial phase consistency of motion vectors. The regions with consistent motion vectors have high probability to be in one moving object. In contrast, the regions with inconsistent motion vectors are more likely located at the boundary of objects or in still background. We measure spatial coherency by using a similar method in [20]. First, we compute a phase histogram in a spatial window with the size of $w \times w$ (pixels) at each location of a macro block. Then, we measure the phase distribution by entropy as follows:

$$\boldsymbol{Cs}(i,j) = -\sum_{t=1}^{n} p_s(t)\mathbf{Log}(p_s(t)) \tag{6}$$

$$p_s(t) = SH_{i,j}^{w}(t) \Big/ \sum_{k=1}^{n} SH_{i,j}^{w}(k) \tag{7}$$

where $SH^{w}_{i,j}(t)$ is the spatial phase histogram whose probability distribution function is $p_s(t)$, and n is the number of histogram bins.

Similar to spatial coherence inductor, we define temporal coherency, the output of *Temporal Coherence Inductor,* in a sliding window of size L (frames) along time axis, as:

$$\boldsymbol{Ct}(i,j) = -\sum_{t=1}^{n} p_t(t)\mathbf{Log}(p_t(t)) \tag{8}$$

$$p_t(t) = TH_{i,j}^{L}(t) \Big/ \sum_{k=1}^{n} TH_{i,j}^{L}(k) \tag{9}$$

where $TH^{L}_{i,j}(t)$ is the temporal phase histogram whose probability distribution function is $p_t(t)$, and n is the number of histogram bins.

In this way, we obtain motion information from three channels $\boldsymbol{I}$, $\boldsymbol{Cs}$, $\boldsymbol{Ct}$, together they compose a motion perception system. Figure 2 (a-c) gives an example of the outputs from the three inductors. Since the outputs from the three inductors, ***I, Cs,*** and ***Ct***, characterize the dynamic spatio-temporal attributes of motion in a particular way, we define motion attention as following:

$$B = \boldsymbol{I} \times \boldsymbol{Ct} \times (1 - \boldsymbol{I} \times \boldsymbol{Cs}) \tag{10}$$

By (10), the outputs from ***I, Cs,*** and ***Ct*** channels are integrated into a motion saliency map, as shown in Figure 2 (d), in which the motion attention areas can be identified precisely.

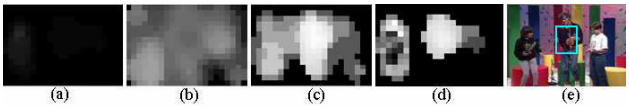

Figure 2. Motion attention detection. (a) ***I***-Map, (b) ***Cs***-Map, (c) ***Ct***-Map, (d) Saliency map, (e) Original image in which the motion attention areas are marked by the blue box.

In order to detect the salient motion regions, we employ following image processing procedures: 1) Histogram balance; 2) Median filtering; 3) Binarization; 4) Region growing; and 5) Region selection. With the results of motion attention detection, we may calculate motion attention model by accumulating the brightness of the detected motion attention regions in saliency map as follows:

$$M_{motion} = \left(\sum_{r \in \Lambda} \sum_{q \in \Omega_r} \boldsymbol{B}_q \right) \Big/ N_{MB} \tag{11}$$

where $\boldsymbol{B}_q$ is the brightness of a macro block in saliency map, Λ is the set of detected areas with motion attention, Ω_r denotes the set of macro blocks in each attention area, and N_{MB} is the number of macro blocks in a MVF which is used for the normalization purpose. The M_{motion} value of each frame in a video then forms a

continuous motion attention curve along the time axis. Figure 8(f) shows an example of such curve.

3.2 Static Attention Model

While motion attention model can reveal most of attentions in video due to the inherent characters of video, it has limitations. For instance, a static background region may attract human attention even through there is no motion. Therefore, there is a need for a static attention model to make up such deficiency.

Itti *et al* defined a saliency-based visual attention model for static scene analysis [14]. We adopt a similar approach. Some changes are made in order to meet the requirements of our attention model framework and the inherent characteristic of videos. The modification mainly focuses on how to generate a time serial attention curve from individual saliency maps.

Similar to the work in [15], a saliency map is generated from each frame by the 3 channel saliency maps computation, *color contrasts*, *intensity contrasts,* and *orientation contrasts*. Then, the final saliency map is built by applying the iterative method proposed in [16]. In stead of locating human's focus of attention orderly, we detect the regions that are most attractive to human attention by binarizing the saliency map. As shown in Figure 3, the size, the position and the brightness of attended regions in gray saliency map decide the degree of human attention they attracted. The blue boxes are marked on original image according to the binarized saliency map. The binarization threshold is estimated in an adaptive manner.

As a result, we define the static attention model according to the number of attended regions and their position, size and brightness in saliency map as follows,

$$M_{static} = \frac{1}{A_{frame}} \sum_{k=1}^{N} \sum_{(i,j) \in R_k} B_{i,j} \cdot w_{pos}^{i,j} \tag{12}$$

where $B_{i,j}$ denotes the brightness of the pixels in saliency regions R_k, N denotes the number of saliency regions, A_{frame} is the area of frame, and $W_{pos}^{i,j}$ is a normalized Gaussian template with the center located at the center of frame. Since human usually pay more attentions to the region near to the center of frame, we use such normalized Gaussian template to assign a weight to the position of the saliency regions. Figure 8(g) shows an example static attention curve of a video segment.

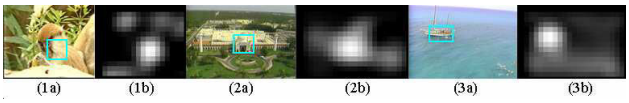

Figure 3. Static attention detection

3.3 Face Attention Model

Face is one of the most salient characters of human beings. The appearance of dominant faces in video frames certainly attracts viewers' attention. Therefore, face attention model is an important part of user attention model, which is able to enhance the power of user attention model significantly.

By employing a real time face detection module presented in [21], we obtain the face information in each frame, including the number of faces, and their poses, sizes, and positions. Figure 4 (a) gives an example of face detection result in a video frame. In current system, total 7 face poses (with out-plane rotation) can be detected, from the frontal to the profile. The size and position of a face usually reflect the importance of the face. Hence, we model face attention as

$$M_{face} = \sum_{k=1}^{N} \frac{A_k}{A_{frame}} \times \frac{w_{pos}^{i}}{8} \tag{13}$$

where A_k denotes the size of k^{th} face in a frame, A_{frame} denotes the area of frame, w_{pos}^{i} is the weight of position defined in Figure 4(b), and $i \in [0,8]$ is the index of position. With this face attention model, we may calculate face attention value at each frame to generate face attention curve. Figure 8(h) shows an example of face attention curve of a video clip in which closed-up faces always have higher values.

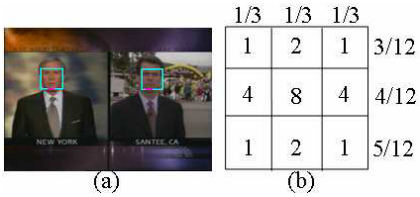

Figure 4. Face detection and position weights

3.4 Camera Attention Model

As we all know, camera motion are always utilized to emphasize or neglect a certain objects or a segment of video, that is, to guide viewers' attentions. Therefore, camera motions are also very useful for formulating user attention model.

Generally speaking, if we let the z- axis go through the axes of lens, and be perpendicular to the image plane x-y, camera motion can be classified into the following types: 1) Panning and tilting, resulted from camera rotations around the x- and y-axis, respectively, both referred as panning in this paper; 2) Rolling, resulted from camera rotations around the z-axis; 3) Tracking and booming, resulted from camera displacement along x- and y-axis, respectively, both referred as tracking in this paper; 4) Dollying, resulted from camera displacement along z-axis; 5) Zooming (In/Out), resulted from lens' focus adjustment; and 6) Still.

By using the affine motion estimation, we can easily determine the camera motion type and speed accurately. However, the challenge is how to map these parameters to the effect they have in attracting the viewer's attention. We derive the camera attention model based on some general camera work rules.

First, the attention factors caused by camera motion are quantified to the range of [0~2], as shown in Figure 5. In the visual attention definition (2), camera motion model is used as a magnifier, which is multiplied with the sum of other visual attention models. A value higher than 1 means emphasis, while a value smaller than 1 means neglect. If the value is equal to 1, the camera does not intend to attract human's attention. If we do not want to consider camera motion in visual attention model, it can be closed by setting the switch coefficient s_{cm} to 0.

Then, we model camera attention based on the following assumptions. 1) Zooming and dollying are always used to emphasize something. The faster the zooming/dollying speed, the more important the content focused is. Usually, zoom-in or dollying forward is used to emphasize the details, while zoom-out or dollying backward is used to emphasize an overview scene. In Figure 5, we consider dollying the same as zooming. 2)

If a video producer wants to neglect something, horizontal panning is applied. The faster the speed is, the less important the content is. On the contrary, unless a video producer wants to emphasize something, vertical panning is not used since it bring viewers unstable feeling. The panning along other direction is more seldom used which is usually caused by mistakes. 3) Other camera motions have no obvious intention. We assign them a value of 1 and leave the attention determination to other visual attention models. 4) If the camera motion changes too frequently, we consider it random or unstable motions. This case is also modeled as 1.

With above assumptions, we model camera motions as shown in Figure 5. First, a zooming process is intended to emphasize the end part of the zooming sequence whereas the frames during the zooming are usually not very important. We assume the importance increases temporally when camera zooms. As shown in Figure 5(a), the attention degree is assigned to 1 when zooming is started, and the attention degree of the end part of the zooming is direct ratio to the speed of zooming V_z. If a camera becomes still after a zooming, the attention degree at the end of the zooming will continue for a certain period of time t_k, and then return to 1, as shown in Figure 5(b).

Also, the attention degree of panning is determined by two aspects: the speed V_p and the direction γ. The attention can be modeled as the product of the inverse of speed and the quantization function of direction as shown in Figure 5(c). Taking the first quadrant as an example in Figure 5(d), we map motion direction $\gamma \in [0 \sim \pi/2]$ to [0~2] by a subsection function. 0 is assigned to direction $\gamma=\pi/4$, 1 is assigned to direction $\gamma=0$, and 2 assigned to direction $\gamma=\pi/2$. The first section is monotonously decreasing while the second section is monotonously increasing. Similar to zooming, if the camera becomes still after a panning, the attention degree will continue for a certain period of time t_k, and the attention degree will be only inverse ratio to the speed of panning V_p as shown in Figure 5(e).

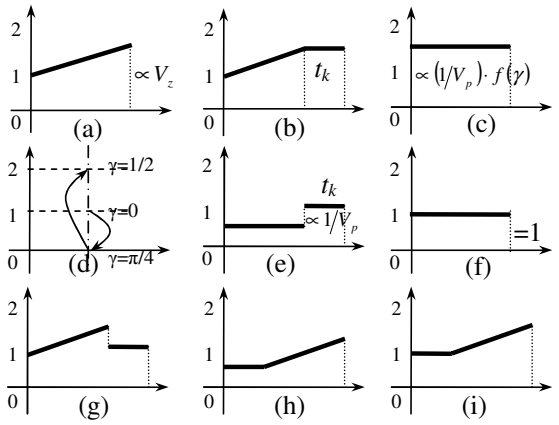

Figure 5. Camera attention modeling (a) Zooming, (b) Zooming followed by still, (c) Panning, (d) Direction mapping function of panning, (e) Panning followed by still, (f) Still and other types of camera motion, (g) Zooming followed by panning, (h) Panning followed by zooming, (i) Still followed by zooming.

Figure 5(f) shows the model of other camera motions, which is a constant value 1, the same as still. In addition, if zooming is followed by a panning, they are modeled independently as shown in Figure 5(g). However, if other types of motion are followed by a zooming, the start attention degree of zooming is determined by the end of these motions. Figure 5(h) and (i) give examples of the zooming followed by panning and still respectively. Figure 8(i) shows an example of camera motion attention curve.

4. AUDIO ATTENTION MODELING

Audio attentions are the important parts of user attention model framework. Speech and music are semantically meaningful for human beings. On the other hand, the loud and sudden sound effects always grab human attention. Therefore, in this section, we define 3 audio attention models: audio saliency attention, speech attention, and music attention.

4.1 Audio Saliency Attention Model

Many characters can be used to represent audio saliency attention model. However, the most fundamental one is loudness. Whether the sound is speech, music, or other special sound (such as whistle, applause, laughing, and explosion), people are always attracted by the louder or sudden sound if they have no subjective intention. Since loudness can be represented by energy, we model audio saliency attention based on audio energy. In general, people may pay attention to an audio segment if one of the following cases occurs. One is the audio segment with absolute loud sound, which can be measured by average energy of an audio segment. The other is the loudness of audio segment being suddenly increased or deceased. We measure such sharp increases or decreases by energy peak. Hence, the audio saliency model is defined as:

$$M_{as} = \overline{E_a} \cdot \overline{E_p} \tag{14}$$

where $\overline{E_a}$ and $\overline{E_p}$ are the two components of audio saliency: normalized average energy and normalized energy peak in an audio segment. They are calculated as follows respectively.

$$\overline{E_a} = E_{avr} / MaxE_{avr} \tag{15}$$

$$\overline{E_p} = E_{peak} / MaxE_{peak} \tag{16}$$

where E_{avr} and E_{peak} denote the average energy and energy peak of an audio segment, respectively. $MaxE_{avr}$ and $MaxE_{peak}$ are the maximum average energy and energy peak of an entire audio segment corps. A sliding window is used to compute audio saliency along an audio segment. Similar to camera attention, audio saliency attention also plays a role of magnifier in audio attention model. Figure 8(j) shows an example of such curve.

4.2 Speech and Music Attention Models

Besides some special sound effects, such as laugh, whistle and explosion, human always pay more attention to the speech or music because speech and music are important cues of a scene in video. In general, music is used to emphasize the atmosphere of scenes in video. Hence a highlight scene is always accompanied with music background. On the other hand, textual semantic information is always conveyed by speech. For example, speech is more attractive to audience in TV news.

Obviously, the audience usually pays more attention to the salient speech or music segments if they are retrieving video clips with speech or music. The saliency of speech or music can be measured by the ratio of speech or music to other sounds in an

audio segment. Based on our previous work on audio classification [22, 23], music and speech ratio can be calculated with the following steps. First, an audio stream is segmented into sub-segments. Then, a set of features are computed from each sub-segment. The features include mel-frequency cepstral coefficients (MFCCs), short time energy (STE), zero crossing rates (ZCR), sub-band powers distribution, brightness, bandwidth, spectrum flux (SF), linear spectrum pair (LSP) divergence distance, band periodicity (BP), and the pitched ratio (ratio between the number of pitched frames and the total number of frames in a sub-clip). Support vector machine is finally used to classify each audio sub-segment into speech, music, silence, and others. With the results of classification, speech ratio and music ratio of a sub-segment are computed as follows.

$$M_{speech} = N^{w}_{speech} \Big/ N^{w}_{total} \quad (17)$$

$$M_{music} = N^{w}_{music} \Big/ N^{w}_{total} \quad (18)$$

where M_{speech} and M_{music} denote speech attention and music attention model, respectively. N^{w}_{speech} is the number of speech sub-segments, and N^{w}_{music} is the number of music sub-segments. The total number of sub-segments in an audio segment is denoted by N^{w}_{total}. Such numbers are all accumulated in a segment w. Figure 8(k) and (l) show the examples of speech and music attention curve respectively.

5. VIDEO SUMMARIZATION SCHEME

A video summarization scheme is developed based on the proposed user attention model. Figure 6 shows a flowchart of the proposed scheme, including both static and dynamic summarization. In this scheme, the component attention models are first computed; then, a user attention curve is generated by the linear combination fusion scheme. An example segment of attention curve is shown in Figure 8(e). Based on this curve, both key-frames and video skims are extracted.

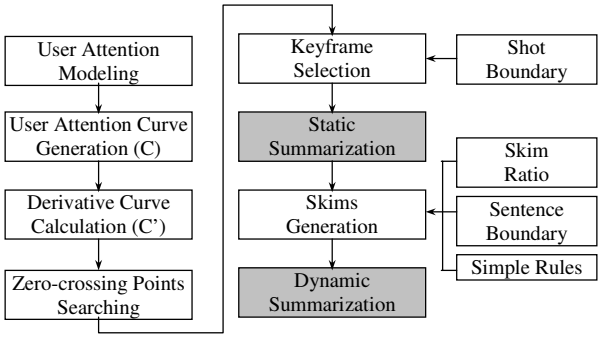

Figure 6. Video summarization flowchart

User attention curve is composed of a time series of the attention values associated with each frame in a video sequence. By smoothing and normalizing, a number of crests can be identified from such a curve. According to the definition of user attention model, the video segments with crests are most likely to attract the viewer's attentions. Therefore, it is reasonable to assume that key-frames and skims should be extracted from these crests. In order to determine the precise position of the peak of a crest, a derivative curve is computed, which is shown in Figure 8(d). The zero-crossing points from positive to negative on derivative curve are the locations of wave crest peaks. In Figure 8(c), a pin with the height equal to the peak attention value is used to represent the key-frame. In this way, all key-frames in a video sequence can be identified without shot boundary detection.

The attention value of a key-frame can be used as the importance measure of the key-frame. Based on such a measure, a multi-scale static abstraction is generated conveniently by ranking the importance of key-frames. For shot-based extraction, key-frames between two shot boundaries can be used as representative frames of a shot. The maximum attention value of the key-frames in a shot is used as its importance indicator. In case that there is no crest in a shot, the middle frame is chosen as the key-frame, and the important value of this shot is assigned to zero. If only one key-frame is required for each shot, the key-frame with the maximum attention is selected. On the other hand, if the total number of key-frames allowed is less than the number of shots in a video, shots with lower importance values will be neglected.

Many approaches can be used to create dynamic video skims based on the user attention curve. We have developed a straightforward shot-based approach to skim generation, which does not require complex heuristic rules. Once a skim ratio is given, skim segments are selected around each key-frame according to the skim ratio within a shot. The process of skim segment selection is illustrated in Figure 7.

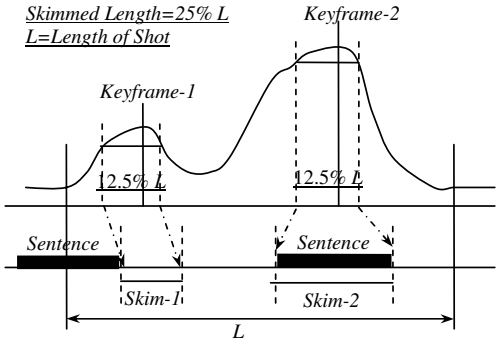

Figure 7. Video summarization scheme

In order to make the sound of a skimmed video smoother, the speech in audio track should not be interrupted within a sentence. So sentence boundary is indispensable information for video skimming. Although it is difficult to fully retrieve each sentence, there are some useful criteria. The most fundamental and important one is that there usually is a pause or silence duration between sentences. However, due to background sound or noise, an audio clip between two sentences may not be a silence. Thus, an adaptive background sound level detection is needed for the purpose of estimating the threshold for pause detection. In our implementation, we segment speech into sentences by the following steps: 1) Adaptive background sound level detection, which is used to set threshold. 2) Pause and non-pause frame identification, using energy and ZCR information. 3) Result smoothing based on the minimum pause length and the minimum speech length, respectively. 4) Sentence boundary detection, which is determined by longer pause duration.

Besides attention curves, shot boundary, sentence boundary and key-frames, only four simple rules are used to create video skims, as shown in Figure 7. 1) Any segment should not be shorter than

the minimum length L_{min}. Usually, L_{min} is set to 30 frames, because the segment shorter than 30 frames is not only too short to convey content, but also creates annoying effects. 2) Given a skim ratio, the length of each skim segment is determined by the length of a shot and the number of key frames in the shot. The skim length of a shot should be distributed to each key frame in this shot evenly. If the average length of skim segment is smaller than L_{min}, the key-frame with minimum attention value is removed. Then, the skim length is re-distributed to the rest of key-frames. This process is carried out iteratively until the average length is higher than minimum length L_{min}. 3) If a skim segment is beyond the shot boundary, it is trimmed at the boundary. 4) The skim segment boundaries must be adjusted according to speech sentence boundaries to avoid splitting speech sentence, either aligning to sentence's boundary like *skim-2* in Figure 7, or evading the sentence's boundary like *skim-1* in Figure 7.

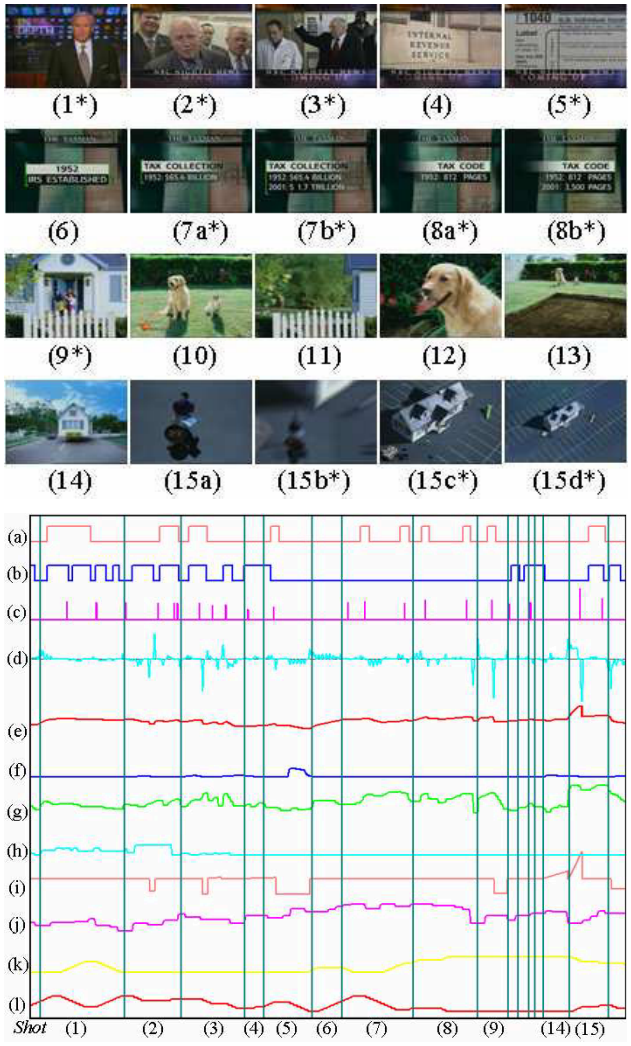

Figure 8. Attention model curves and summarization examples. (a) Skimming curve. (b) Sentence boundary. (c) Zero-crossing curve. (d) Derivative curve. (e) User attention curve. (f) Motion attention curve. (g) Static attention curve. (h) Face attention curve. (i) Camera attention curve. (j) Audio saliency attention curve. (k) Speech attention curve. (l) Music attention curve.

6. EVALUATIONS

Although the issues of key-frame extraction and video summary have been intensively addressed, there is no standard method to evaluate algorithm performance. The assessment of the quality of a video summary is a strong subjective task. It is very difficult to do any programmatic or simulated comparison to obtain accurate evaluations, because such methods are not consistent with human perception. To evaluate our proposed video summarization approach, we have carried out a user study experiment.

Figure 8 gives some examples of key-frames, a segment of user attention model curves and the generated skim selection curve at the ratio of 30%. The top of Figure 8 shows 15 shots by their key-frames. The key-frames marked with asterisk are selected to generate video skims. The bottom of Figure 8 gives corresponding segment of attention model curves, derivative curve, zero-crossing curve, shot boundary, sentence boundary, and skimming curve. The vertical lines are shot boundaries. There are 15 shots in this example segment, which is extracted from nbcnews.mpg, a news video program in our test dataset. The segments of positive pulses in the skim selection curve of Figure 8(a) denote selected skims. Similarly, the segments of positive pulses in Figure 8(b) denote speech sentences. In addition, it is worth noticing that two key-frames are detected in shot-15 (15a and 15d), though four frames (15 a-d) are displayed here to illustrate a process of zooming-out. This zooming-out segment is identified by camera motion detection algorithm and emphasized by camera attention model curve in Figure 8(i). As a result, this segment is selected to be a part of the video skims. Also, in our implementation, all weights for linear combination are assigned to 1, and all magnifiers' switches are also assigned to 1 currently.

To quantitatively evaluate the quality of video summarization, we invited 20 volunteered subjects, including 9 males and 11 females, who gave their subjective scores to the video summaries generated by the proposed approach. Our experiment is composed of three parts: single key-frame evaluation, multiple key-frame evaluation, and video skimming evaluation. In order to keep our subjects innocent of video content before they look through the video skims, they were required to evaluate the dynamic summary first.

Table 1. Test videos

No.	Video	Genre	Shot	Time
I	Animals.mpg	Documentary	83	9:14
II	Bahamas.mpg	Sightseeing	53	5:39
III	Basketball.mpg	Sport game	75	10:02
IV	Nbcnews.mpg	TV news	488	30:10
V	TheBoy.mpg	Home video	43	12:35
Total	——	——	**742**	**67:40**

We have chosen test data from various types of videos with different lengths, since the proposed attention model is aimed to be generic for detecting the important content of general videos. Totally five video sequences are selected as our test dataset. They include Animals.mpg, an educational documentary; Bahamas.mpg (no audio track), a sightseeing video; Basketball.mpg, a sports video; nbcnews.mpg, a half an hour news program; and

TheBoy.mpg, a segment of home video. The total length of test videos is close to 70 minutes. The details of test videos are listed in Table 1. As a fundamental structure of video, shot is automatically detected by employing the algorithm proposed in [24]. The total number of shots segmented from test videos is 742.

6.1 Static Summarization Evaluation

The number of key-frame selected for presenting a shot is tailored to applications. We evaluate static abstraction from two aspects: single key-frame and multiple key-frames. The main concern of single key-frame representation is how informative a key-frame is as a representative icon of shot. Unlike single key-frame, the multiple key-frames can provide user more information, at the meantime, showing the process of evolving events. Thus, its assessment criteria are different from that for single key-frame. Users usually have more concerns if the extracted key frames are redundant or insufficient. In our experiments, the subjects were required to give an assessment of being *good*, *neutral*, or *bad* to single key-frame representation, and give an assessment of being *good*, *too much*, or *too few* to multiple key-frames representation.

During the evaluation, the subjects briefly reviewed the shots by double-clicking their icons first. Then, they were asked to give an assessment to the single and multiple key-frames, respectively. The single key-frame of a shot was used as the icon of the shot. All shot icons are displayed in a planar view. When a subject clicks on an icon, the multiple key-frames will be displayed in another view below. In this way, the subjects can make their judgments on two representations shot by shot conveniently.

Table 2. The evaluation of static video abstraction: Single

No.	100 (Good)	50 (Neutral)	0 (Bad)	Avg.
I	62.90	17.30	2.80	86.2
II	39.85	11.40	1.75	85.94
III	53.2	18.15	3.65	83.03
IV	411.4	58.65	17.95	90.31
V	32.50	8.85	1.65	85.87
Avg.	——	——	——	**86.27**

In order to obtain a quantitative assessment, we quantify the subjects' assessment of a single key-frame into score 100, 50 and 0, corresponding to *good*, *neutral*, and *bad,* respectively. Then the average score of a single key-frame is calculated according to the number of shots counted in each category. The statistical results are listed in Table 2. The average score for single key-frame extraction evaluation is above 86.

Table 3. The evaluation of static video abstraction: Multiple

No.	Good (G)	Too Much (M)	Too Few (F)	G/All
I	62.25	9.65	11.10	0.7500
II	39.90	3.00	10.10	0.7528
III	52.45	6.85	15.70	0.6993
IV	403.40	34.70	49.90	0.8266
V	30.20	8.05	4.75	0.7023
Avg.	——	——	——	**0.7462**

For the evaluation of multiple key-frames, the quantified measure, user satisfaction, is defined as a ratio of the number of shots with *good* assessment to the total number of shots. The details are listed in Table 3. The average user satisfaction score is close to 75%. One of the reasons for such promising results is that the computational attention models are consistent with human perception. The other is the user attention curve embeds the audio information, which provide important hints for key-frame selection, such as speech or some special sound effects.

6.2 Dynamic Summarization Evaluation

Generally speaking, users expect that the skimmed video is as short and as informative as possible. However, it is difficult to achieve the two objectives at the same time. Though for a given video program, the skim ratio could be controlled automatically by setting a threshold in the user attention curve, it is more reasonable to let the user choose a skim ratio in our experiments. In our experiments, 15% and 30% skimmed video are considered.

Two criteria are set up for performance evaluation. Whether the skimmed video is informative is most important for video skimming. So, we set the first criterion as *informativeness.* Besides, whether the skimmed video is still as enjoyable as the original one is also an important criterion. We call it *enjoyablity.* The *enjoyablity* assesses not only the smoothness of an image sequence, but also the influence of speech and music. Consequently, in this part of experiment, the subjects were required to assign two 0~100 scores to the skimmed videos according to these two criteria, *informativeness* and *enjoyablity,* respectively. Since the users may think the original video is not enjoyable or informative enough, we also give the subjects the right to assign the two scores to the original video as well. Finally, we normalize the scores of skimmed videos by the scores assigned to the original ones.

In order to obtain the precise scores, the subjects should not know the content of the test videos before they look through the skimmed video. Therefore, this part of experiment was performed first before the key-frame evaluation. The subjects looked through the skimmed video sequences from high to low skim ratio in turn controlled by the evaluation program. That is, they look through 15%, 30% of the skimming and the original videos one by one without fast forward or backward functions. The subjects were only allowed to give the scores to test sequence when he/she finishes viewing a sequence. After finishing the viewing of the original video, they have the chance to modify the scores assigned to the skimmed sequences, because they would understand the video content in more details by that time. With the scores subjects assigned, the average normalized *enjoyablilty* and *informativeness* scores are calculated, which reflect the user's satisfactory degree to the skimmed videos.

The experimental results are list in Table 4. The scores in the non-highlighted rows are the average real scores from 20 subjects. The highlighted rows show the average scores normalized by the subjects' scores on original video. The overall average results at the bottom of the table are calculated based on the normalized scores. From Table 4, we can see that the average *enjoyablility* and *informativeness* scores at 15% skim ratio are 66.10 and 63.01, respectively, while the ones at 30% are 76.38 and 74.86, respectively. These numbers indicate that the proposed skimming scheme is effective in high skim ratio. When a video is skimmed

by 70% or 85%, i.e., the skim ratio is 30% or 15%, viewers' satisfactory degree drops 23.62%, and 33.9% in *enjoyablility* score, and 25.14% and 36.99% in *informativeness* score, respectively.

Table 4. The evaluation of dynamic video skimming

No.	Enjoyablity			Informativeness		
	15%	30%	100%	15%	30%	100%
I	62.65	73.40	95.40	63.60	75.70	99
	65.67	76.93	100	64.24	76.46	100
II	63.00	70.40	88.4	62.60	71.05	91.85
	71.27	79.64	100	68.15	77.35	100
III	54.15	66.30	93.05	56.40	69.95	97.40
	58.19	71.25	100	57.91	71.82	100
IV	62.25	72.65	96.60	62.55	73.30	99.65
	67.55	75.21	100	62.77	73.56	100
V	60.70	70.60	89.50	59.4	72.00	95.85
	67.82	78.88	100	61.97	75.12	100
Avg.	**66.10**	**76.38**	——	**63.01**	**74.86**	——
Drop (%)	**33.90**	**23.62**	——	**36.99**	**25.14**	——

The *enjoyablility* score is usually higher than the *informativeness* score at the same skim ratio. However, the difference is very limited. This is possibly due to the fact that it is difficult for the subjects to discriminate unenjoyable from uninformative, because when a viewer cannot obtain the information they expect, he/she tends to also feel the skimmed one is not enjoyable. On the other hand, when the original does not have audio track, the result is quite different. As indicated by the scores on video Bahamas.mpg, the absolute *informativeness* score of the original is the lowest among test videos: 91.85 only; But, the normalized *enjoyablity* scores of the skims of this video are the highest. This is a case that *enjoyablity* becomes more important than *informativeness*, since there is less information in this sequence.

7. CONCLUSIONS

Automatic video summarization is a powerful tool for video browsing and accessing. In this paper, we have presented a new approach to video summarization, including both key-frame extraction and video skimming, without fully semantic content understanding and the requirements of complex heuristic rules. This approach constructs video summaries based on the modeling of how viewers' attentions are attracted by motion, object, audio and language when viewing a video program. A generic and extendable user attention model has been designed and a set of modeling methods for audio-visual features have been proposed to support this framework. The user study results show that both static and dynamic video summarizations are able to meet the users' requirements of video browsing. The average satisfaction scores of single key-frame and multiple key-frames extracted by the proposed approach are 86.27 and 74.62, respectively. The satisfaction scores of dynamic summarization are both over 63 at the 15% ratio and about 75 at the 30% ratio. Such promising result not only verifies the effectiveness of the proposed summarization scheme, but also indicates that the computational attention model is an efficient method for extracting the important content from video without fully semantic analysis.

As a generic and extendable framework, the proposed video attention model can also be used in other applications, which will be our future works. On the other hand, the fusion scheme requires further studies, because it is yet to be proved as to what kind of fusion scheme is most effective currently. In addition, the evaluation method of summarization also needs further improvement. A more independent measure will be investigated to evaluate the smoothness of the dynamic summarization so that it is not heavily affected by the *informativeness* score.

8. ACKNOWLEDGEMENTS

We would like to express our appreciation to Xiangyang Lan and Feng Yu, both from Tsinghua University. During their internship with us, they implemented several components presented in this paper. We also would like to express our appreciation to Prof. Chong-Wah Ngo of City University of Hong Kong, who gave us valuable advices during his visit to our lab.

9. REFERENCES

[1] H.J Zhang, *et al,* "An integrated system for content-based video retrieval and browsing," Pattern Recognition, vol.30, no.4, pp.643-658, 1997.

[2] W. Wolf, "Key frame selection by motion analysis," *Proc. of* ICASSP'96, vol.2, pp.1228-1231, 1996.

[3] Y. Zhuang, *et al,* "Adaptive key frame extraction using unsupervised clustering," *Proc. of* ICIP'98, 1998.

[4] A. Girgensohn, and J.Boreczky, "Time-constrained key frame selection technique," *Proc. of* ICMCS, pp. 756-761, 1999.

[5] F. Dufaux, "Key frame selection to represent a video," *Proc. of* ICME 2000.

[6] P. Campisi, A. Longari and A.Neri, "Automatic key frame selection using a wavelet based approach," *Proc. of* SPIE, vol. 3813, pp. 861-872, July 1999.

[7] C. Kim and J. N. Hwang, "An integrated scheme for object-based video abstraction," *Proc. of* ACM Multimedia 2000. Los Angeles, CA, 2000.

[8] N. Omoigui, L. He, A. Gupta, J. Grudin and E. Sanoki, "Time-compression: System concerns, usage, and benefits," *Proc. of* ACM ICH'99, 1999.

[9] M.A. Smith, T. Kanade, "Video skimming and characterization through the combination of image and language understanding techniques," *Proc. of* Computer Vision and Pattern Recognition, 1997.

[10] N. Jeho, H.T. Ahmed, "Dynamic video summarization and visualization," *Proc. of* ACM Multimedia, October 1999.

[11] A. Stefanidis, P. Partsinevelos, P.;Agouris, P.;Doucette, "Summarizing video datasets in the spatiotemporal domain," *Proc. of* 11th Inter. Workshop on Database and Expert Systems Applications, 2000.

[12] X. Orriols, X. Binefa, "An EM algorithm for video summarization, generative model approach," *Proc. of* ICCV, 2001.

[13] Y.H. Gong, X. Liu, "Video Summarization Using Singular Value Decomposition**,"** *Proc. of* CVPR, June, 2000.

[14] L. Itti, C. Koch, "Computational Modeling of Visual Attention," *Nature Reviews Neuroscience*, Vol. 2, No. 3, pp. 194-203, Mar. 2001.

[15] 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.

[16] L. Itti, C. Koch, "A Comparison of Feature Combination Strategies for Saliency-Based Visual Attention Systems," *Proc. of* SPIE Human Vision and Electronic Imaging IV (HVEI'99), San Jose, CA, Vol. 3644, pp. 473-82, Jan 1999.

[17] A.A. Salah, E. Alpaydin, and L. Akarun, "A Selective Attention-based method for Visual Pattern Recognition with Application to Handwritten Digit Recognition and Face recognition," *IEEE Trans. On* Pattern Analysis and Machine Intelligence, Vol.24, No. 3, Mar. 2002.

[18] M. Bollmann, R. Hoischen, B. Mertsching, "Integration of Static and Dynamic Scene Features Guiding Visual Attention," Paulus, E. Wahl, F. M. (eds.). Springer, 1997, pp. 483-490.

[19] Y.F Ma, H.J. Zhang, "A Model of Motion Attention for Video Skimming," *Proc. of* ICIP, 2002.

[20] Y.F. Ma, H.J. Zhang "A New Perceived Motion based Shot Content Representation," *Proc. of* ICIP, 2001.

[21] S. Z. Li, et al., "Statistical Learning of Multi-View Face Detection," *Proc. of* ECCV 2002.

[22] L. Lu, H. Jiang, H. J. Zhang, "A Robust Audio Classification and Segmentation Method," *Proc. of* the 9th ACM International Conference on Multimedia, pp. 203-211, 2001

[23] L. Lu, Stan Li, H. J. Zhang, "Content-based Audio Segmentation Using Support Vector Machines". *Proc. of* ICME 2001, pp956-959, Tokyo, Japan, 2001.

[24] H.J. Zhang, A. Kankanhalli, and S.W. Smoliar, "Automatic partitioning of full-motion video," *Multimedia Systems*, 1, pp.10-28, 1993.