AN IMPORTANCE MEASUREMENT FOR VIDEO AND ITS APPLICATION TO TV NEWS ITEMS DISTILLATION

Jin-Hau Kuo, Chin-Wei Fang, Jen-Hao Yeh and Ja-Ling Wu, ***Senior Member, IEEE***

Communication and Multimedia Laboratory, Department of Computer Science and Information Engineering, National Taiwan University, Taipei, Taiwan, R.O.C.

ABSTRACT

This paper presents the method to distill the important frames and shots from long duration video clips. It is based on a voluntary choice by the subject (i.e. a top down approach). First, we introduce the concept of importance measurement. We use some predefined detectable events to represent the importance inferred from the semantics of the syntax. Since the event characteristics depend largely on the video sources, we choose news program videos as our analyzing focus. And then we calculate the importance measurement of frames and shots, respectively. Experiments show that a good subjective test result for news items distillation can be obtained, and the average distillation ratios of frames and shots are 13.18% and 41.23%, respectively.

1. INTRODUCTION

Recently, Google adds image searching to its repertoire of search capabilities. The database indexed over 390 million images from the web. The phenomenon shows that the investigation of multimedia, especially for content analysis and retrieval, is becoming important day by day. In [12], Zhu Liu et al. present a new approach for automatically generating a list of major casts for video. In addition, driven by the trends of mobile communication and personal media applications, *content distillation* (c.f. Fig.1) issues have got more and more attention. Zhang et al. [1] presented a generic framework for video summarization based on the modeling of viewer's attention. Without fully understanding the semantic of the video content, this approach takes advantages of computational attention models so as to eliminate the needs of complex heuristic rules in video summarization. In [2], Shih-Fu Chang presented a new conceptual framework to model content, adaptation processes, utility, resources, and relations among them. The key to the framework is the definition of the notion of utility and the formulation of the problem as one of the constrained utility optimization problem. They used the framework to form a unified view towards problems and solutions they have developed in several media adaptation projects, ranging from video skimming, transcoding, to adaptive streaming.

Seeking for help of psychology, we proposed an importance measurement of video to calculate the human's attention or importance of each frame (c.f. Fig. 1). The approach is based on a voluntary choice by the subject (top down), which is different from Zhang's bottom up approach (based on properties of stimulus). The rest of the paper is organized as follows. The concept of importance measurement (IMM) is addressed in Section 2. Sections 3 and 4 investigate the visual IMM and audio IMM, respectively. Subjective experimental results are reported in Section 5, and a conclusion is given in Section 6.

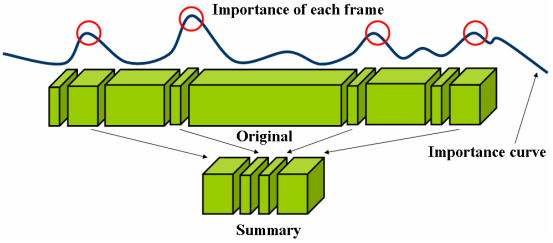


Fig. 1. The interpretation of content distillation and the concept of frame importance.

2. THE IMPORTANCE MEASUREMENT OF VIDEO

The desirable criteria for an ideal content distillation are the following four properties [3]:

1. *Conciseness*: Any segment of the content that is selected for the distillation should contain only necessary information. Thus, each segment should be as short as possible.
2. *Coverage*: The set of segments selected for the distillation should cover all the "key" points of the content.
3. *Context*: The segments selected and their sequencing should be such that prior segments establish appropriate context.
4. *Coherence*: The flow between the segments in the distillation should be natural and fluid.

The first step to make good distillation is *Coverage*. If we know what the key points for a viewer are, we can reserve those to compose the distillation. In [4], Yantis described that the visual system receives vast quantities of information. Attention allows the visual system to select a subset of that information for further processing based either on properties of stimulus (bottom up) or on a voluntary choice by the subject (top down).

The IMM tries to calculate the human's attention or importance of each frame in a video sequence based on a voluntary choice by the subject. Since the subjective of human is semantic and hard to simulate, it is preferable to understand the semantics of the syntax of the domain first. We express the semantics of the syntax of a video basing on detectable events (c.f. Fig. 2). According to the characteristics of a video, video events are classified as continuous and instant ones. For example, the occurrence of a human face is a continuous event (because it lasts at least 1~2 seconds), and the flashlight is an instant one (because it usually last just 1~2 frames).

0-7803-8554-3/04/$20.00 ©2004 IEEE.

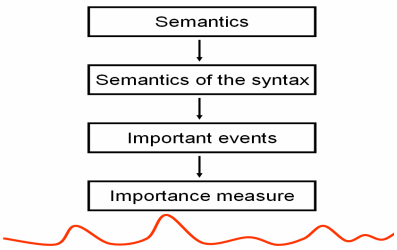


Fig. 2. The hierarchy of importance measurement.

The block diagram of the proposed video summarization (distillation) system is presented in Fig. 3. The video sequence (consists of a set of MPEG compressed news item video clips) is de-multiplexed as video and audio streams. After that, we analyze both of them to get their importance measures, respectively. The block diagram of the visual importance analysis modules is shown in Fig. 4. Note that the relationship between video and audio is also taken into account in this module. Finally, we can measure the importance of each video frame basing on the analytic results and then connect all the importance measures to form the so-called importance curve of the video. And, in this work, key-frames and summarizations are produced according to the produced importance curve. Since the event characteristics depend largely on the video sources, for simplicity and effectiveness, we choose news program videos as our analyzing focus.

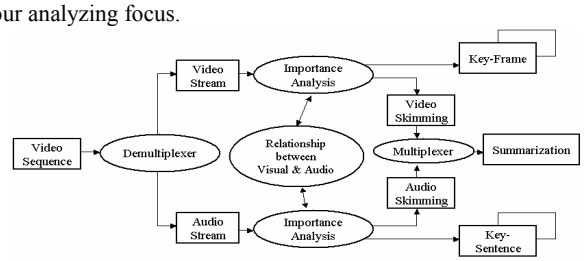


Fig. 3. The block diagram of the proposed video distillation.

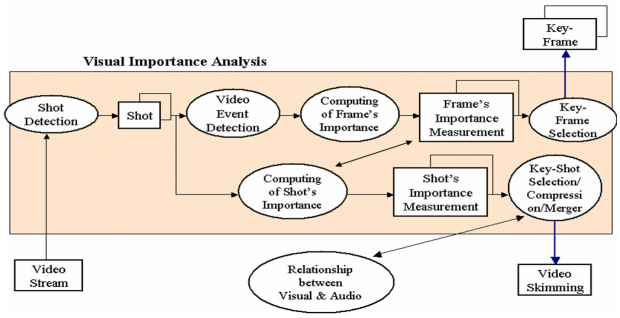


Fig. 4. The block diagram of the visual importance analysis module. The input video stream consists of individual news items videos, which are obtained by using model-free anchorperson detection [5].

2.1. Event definition

Actually, the important events of a TV news are determined according to certain people's behavior (e.g. the behavior of photographers, anchorpersons, post-producers and personages in news). Their actions will influence audiences' or viewers' feeling of one news item. We have classified events as continuous and instant ones and their importance can be further defined as follows.

2.1.1. Continuous events

(a) The duration of face appearance

4W1H (*who, where, when, what and how*) rule often exists in the news science. In TV news, the element *who* is presented as faces. Thus, the appearance and duration of face carries certain degree of importance to audience in news. We use the open source code, Intel® Open Source Computer Vision Library [6], to accomplish the face detection task. It can be replaced by any good face detection approaches.

The IMM of face event can be calculated as:

$$C_{face} = Const. * \frac{MaxFaceArea}{\sum FaceArea} * W, \quad (1)$$

$$W = \begin{cases} 1 & , \sum FaceArea \geq Threshold \\ \dfrac{\sum FaceArea}{Threshold} & , \sum FaceArea \leq Threshold \end{cases}, \quad (2)$$

where $\sum FaceArea$ and $MaxFaceArea$ denote the square measure of all faces and the biggest face areas in one frame, respectively.

(b) The duration of caption text appearance

Undoubtedly, the caption text also carries important information to viewers (*what, when*). The importance measurement of caption event is calculated as

$$C_{text} = Const. * \frac{number\ of\ MB\ with\ Caption}{number\ of\ Total\ MB}, \quad (3)$$

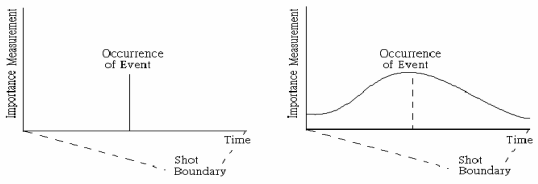
where *MB* denotes macroblock. We modify and extend approach of compressed domain caption detection presented in [7] to check if a block is with caption or not.

2.1.2. Instant events

In fact, an event and its IMM should be continuous in real world. Instant events are resulted from the digital representation of media data (c.f. Fig. 5(a)). Thus, we need to mask their importance measurement by using a continuous curve, such as the one shown in Fig. 5(b). Based on human's perception in [4], we adopt the following masking function:

$$M = \exp(\frac{-\Delta t^2}{2\sigma}) \quad (4) \quad \text{and} \quad \sigma = \frac{Const.}{Dyn} \quad (5)$$

where *Dyn* denotes the speed of camera motion and Δt denotes the distance away from the instant event. *Dyn* can be obtained from parameters of the Global Motion Estimation.



(a) (b)

Fig. 5. The occurrence of (a) instant event and (b) its importance measurement curve.

The instant news video events we defined include:

(a) The occurrence of flashlight (*who*)

In TV news, the photographers' flash is often lightened when important people appears. It usually lasts just for 1~2 frames.

(b) The occurrence of zooming (*who, what*)

The zoom-in and zoom-out operations are used to represent local and global views, respectively. They are both important in modeling the human attention. The cameraman usually uses the zoom-in operation to produce a close-up shot and captures a notable. The corresponding importance measurement and masking curves of zoom-in and zoom-out operations are shown in Fig. 6.

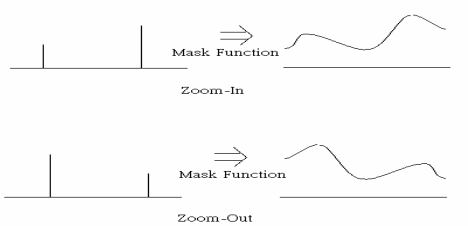


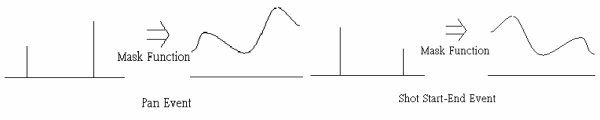
Fig. 6. The masked importance measurement curves of zoom-in and zoom-out operations.

(c) The occurrence of panning (*where*)

In news photograph, the cameraman uses panning to find somebody or look around a place. Thus, the element *where* can be presented as panning operation. There are usually more important scenes at the beginning and the ending of a panning. The importance measurement of panning is shown in Fig. 7(a). Since the beginning of a panning is often out of focus, it is less important than the ending one.

(d) The beginning and ending of a shot

In TV station, post-producer will edit the news items captured by photographer (because of the time for each news item is limited). The accomplished shots of a news item must be the essence. Comparatively, the beginning and ending of a shot carry more important information to audiences (c.f. Fig. 7(b)).



(a) (b)

Fig. 7. (a) The masked importance measurement curve of the panning operation. (b) The masked importance measurement curve of a shot.

We use [8] to characterize global camera motion and the proposed Global Motion Compensated Frame Difference approach [9] to detect accurate shot boundaries (such as hard scene change, dissolve and flashlight).

3. VISUAL IMPORTANCE MEASUREMENT

3.1. Visual IMM of frames

The IMM of frames is defined as:

$$FI = w_c FI_c + (1 - w_c) FI_i, \tag{6}$$

where FI_c and FI_i denote the importance measurements of continuous and instant events, respectively, in one frame. The definitions of FI_c and FI_i are:

$$FI_c = L * \sum_i C_i, \tag{7}$$

where L denotes the length of the current shot and C_i denotes the importance measurement of the continuous event i.

$$FI_i = \sum_i \left(\sum_j b_j L_j I_j \right) * \exp\left(\frac{-(t - t_{cur})^2}{2\sigma}\right), \tag{8}$$

where b_j is a boolean variable to indicate whether instant event j takes place or not; t_{cur} denotes the time instance of current frame; L_j denotes the length of instant event j; I_j denotes the importance measurement of instant event j; t denotes every frame's time instance in the shot. The procedures for computing importance measures of frames are shown in Fig. 8.

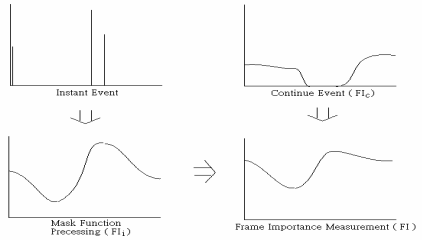


Fig. 8. The procedures for computing importance measures of frames.

3.2. Visual IMM of shots

We extend frames' IMM to shots' one. A long video will be condensed according to the IMM of shots. In [10], S. Uchihashi et al. addressed the importance score of a shot as:

$$I_j = L_j \log \frac{1}{W_k}, \tag{9}$$

where L_j is the length of shot j and W_k is its weight. Given C clusters in a video, a measure of normalized weight W_k for the k_{th} cluster is computed as

$$W_k = \frac{S_k}{\sum_{i=1}^{C} S_i}, \tag{10}$$

where S_k is the total length of all shots in cluster k, found by summing the length of all shots in the cluster. W_k is the proportion of segments from the whole video that are in cluster k.

The importance score becomes larger if the shot is longer, and smaller if the cluster weight is larger (meaning that the shot is not so important). The contributions from the length and the cluster weight can be balanced by weighting the reciprocal of W_k by a factor other than unity.

Based on (6) and the IMM of frames, we define the importance measurement of a shot as:

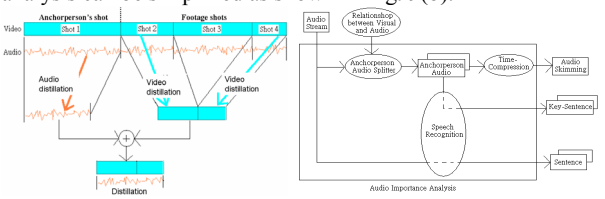
$$I_j = FI^j_{keyframs} \log \frac{1}{W_k}, \tag{11}$$

where $FI^j_{keyframes}$ is the sum of keyframes' FI in shot j.

4. AUDIO IMPORTANCE MEASUREMENT AND THE RELATIONSHIP BETWEEN VIDEO AND AUDIO

The synchronization of audio and video in a news video is loose except for the anchorperson's shots. Furthermore, the anchorperson's speech can be regarded as the essence of audio

in a news item. It is reasonable to take the anchorperson's speech as the distillation of the audio stream. We can, therefore, condense the news footage shots to fit the length of the anchorperson's speech (c.f. Fig. 9(a)). Based on the above assumption and strategy, the diagram of audio importance analysis can be simplified as shown in Fig. 9(b).



(a) (b)

Fig. 9. (a) The relationship between video and audio streams in a news video. (b) The diagram of audio importance analysis.

5. EXPERIMENTAL RESULTS

The goodness of news items distillation has been verified through subjective tests by using one-day life news video program taken from one of the local TV stations in Taiwan (approximately 1 hours in length). The preliminary test result is rather promising (in which the investigators includes both 3 TV station's administrators and 12 members of our Laboratory). The hypothesis is that the importance between a distillation and the original is not far off. Fifteen viewers were randomly presented with the original and the distillation video clips and were asked to report scores of the goodness. The weighted goodness is mapped to the five-grade scale as used in the conventional subjective tests. A condensed video clip scores 100-point means it reserves most of the importance (Excellent) while scoring 20-point denotes the distillation is chaotic (Terrible). The resulting average score (73.82) corroborates with our hypothesis (c.f. Fig. 11). Fig. 10 shows the resultant distillation ratios of frames (13.18%) and shots (41.23%), respectively.

News items	Shot(Number of Shot)			Time(Number of Frame)		
	Original Video (A)	Summarization Video (B)	Ratio (B/A*100%)	Original Video(C)	Summarization Video(D)	Ratio (D/C*100%)
2002-09-19-2	43	9	20.93%	5550	380	6.85%
2002-09-19-3	12	7	58.33%	1950	298	15.28%
2002-09-19-4	15	5	33.33%	1995	212	10.63%
2002-09-19-5	18	10	55.56%	2610	413	15.82%
2002-09-19-6	12	8	66.67%	2100	330	15.71%
2002-09-19-7	20	8	40.00%	2670	341	12.77%
2002-09-19-8	15	8	53.33%	2250	335	14.89%
2002-09-19-9	24	11	45.83%	2445	445	18.20%
2002-09-19-10	18	8	44.44%	2520	335	13.29%
2002-09-19-11	29	12	41.38%	2865	490	17.10%
2002-09-19-12	22	8	36.36%	2745	336	12.24%
Sum	228	94	41.23%	29700	3915	13.18%

Fig. 10. The distillation ratios of frames and shots, respectively.

News items	Excellent (100)	Good (80)	Fair (60)	Bad (40)	Terrible (20)	Score
2002-09-19-2	0	5	5	0	0	70
2002-09-19-3	4	3	2	1	0	80
2002-09-19-4	3	3	2	1	1	72
2002-09-19-5	3	3	4	0	0	78
2002-09-19-6	0	5	3	2	0	66
2002-09-19-7	3	2	3	2	0	72
2002-09-19-8	2	4	4	0	0	76
2002-09-19-9	1	4	4	1	0	70
2002-09-19-10	4	1	4	1	0	76
2002-09-19-11	3	3	3	1	0	76
2002-09-19-12	1	6	3	0	0	76
Sum	24	39	37	9	1	73.82

Fig. 11. The subjective test of news item distillation.

6. CONCLUSION AND FUTURE WORK

Importance measurement for video and its application to TV news items distillation are presented in this paper. The experimental results show quite satisfactory performance of the proposed method of distillation, in terms of distillation ratios and subjective test. In the near future, we will extend our studies to other videos (e.g. drama, sport, commercial and music video) and investigate their domain related knowledge to develop more appropriate processing tools.

7. ACKNOWLEDGE

This work was partially supported by the National Science Council and the Ministry of Education of ROC under the contract No. NSC92-2622-E-002-002, NSC92-2213-E-002-023 and 89E-FA06-2-4-8.

8. REFERENCES

[1] Liwei He, Elizabeth Sanocki, Anoop Gupta, and Jonathan Grudin, "Auto-Summarization of Audio-Video Presentations," *In Proceedings ACM Multimedia*, Orlando FL, Nov. 1999.

[2] H.J Zhang et. al. "A User Attention Model for Video Summarization," *In Proceedings ACM Multimedia*, France, 2002.

[3] S-F Chang, "Optimal Video Adaptation and Skimming Using a Utility-Based Framework," *Tyrrhenian International Workshop on Digital Communications (IWDC-2002)*, Capri Island, Italy, Sept. 2002.

[4] Steven Yantis, *Visual Perception- Essential readings*, Psychology Press, 10 November, 2000.

[5] Xinbo Gao and Xiaoou Tang, "Unsupervised video-shot segmentation and model-free anchorperson detection for news video story parsing," *in IEEE Transactions on Circuits and Systems for Video Technology*, Volume: 12 Issue: 9, Sept. 2002.

[6] http://www.intel.com/research/mrl/research/opencv/, Intel® Open Source Computer Vision Library

[7] Yu Zhong, Hongjiang Zhang, and Anil K. Jain, "Automatic Caption Localization in Compressed Video," *in IEEE Transactions on Pattern Analysis and Machine Inteligence,* VOL. 22, NO. 4, APRIL 2000.

[8] J.-H. Kuo and J.-L. Wu, "An Efficient Algorithm for Scene Change Detection and Camera Motion Characterization Using the approach of Heterogeneous Video Transcoding on MPEG Compressed Videos," *IEEE 3rd Int. Conf. On Information, Communications & Signal Processing,* Oct 2001.

[9] Chin-Wei Fang, Jin-Hau Kuo and Ja-Ling Wu, "An Effective Summarization Tool for News Videos," in *National Computer Symposium*, 2003.

[10] S. Uchihashi, J. Foote, A. Girgensohn, and J. Boreczky, "Video Manga: Generating Semantically Meaningful Video Summaries," *In Proceedings ACM Multimedia*, Orlando FL, Nov. 1999.

[11] Berlin Chen, Hsin-min Wang, and Lin-shan Lee, "Discriminating Capabilities of Syllable-Based Features and Approaches of Utilizing Them for Voice Retrieval of Speech Information In Mandarin Chinese," *in IEEE Trans. on Speech and Audio Processing*, Vol. 10, No. 5, pp. 303-314, July 2002.

[12] Z. Liu and Y. Wang, "Major Cast Detection in Video Using Both Audio and Visual Information," *in ICASSP-2001*, Salt Lake City, Utah, May 2001.