

Semantic-Event Based Analysis and Segmentation of Wedding Ceremony Videos

Wen-Huang Cheng, Yung-Yu Chuang, Bing-Yu Chen, Ja-Ling Wu, Shao-Yen Fang
Yin-Tzu Lin, Chi-Chang Hsieh, Chen-Ming Pan, Wei-Ta Chu, and Min-Chun Tien
Communications and Multimedia Lab

National Taiwan University

{wisley, cyy, robin, wjl, strawinsky, known, nonrat, pan, wtchu, trimy}@cmlab.csie.ntu.edu.tw

ABSTRACT

Wedding is one of the most important ceremonies in our lives. It symbolizes the birth and creation of a new family. In this paper, we present a system for automatically segmenting a wedding ceremony video into a sequence of recognized wedding events, e.g., the couple's wedding kiss. Our goal is to develop an automatic tool for users to efficiently organize, search, and retrieve his/her treasured wedding memories. Furthermore, the event descriptions could benefit and complement the current research in semantic video understanding. Technically, three kinds of event features, i.e., the speech/music discriminator, flashlight detector, and bride indicator, are exploited to build statistical models for each wedding event. Events are then recognized by a hidden Markov model, which takes into account both the fitness of observed features and the temporal rationality of event ordering to improve the segmentation accuracy. We conducted experiments on a rich set of wedding videos, and the results demonstrate the effectiveness of our approach.

Categories and Subject Descriptors

H.3.1 [Content Analysis and Indexing]: Indexing methods; H.4 [Information Systems Applications]: Miscellaneous

General Terms

Algorithms

Keywords

Wedding analysis, event detection, video segmentation

1. INTRODUCTION

A wedding ceremony is an occasion that a couple's families and friends gather together to celebrate, witness, and usher the beginning of their marriage. It is a public announcement of the couple's transition from two separate lives to a family

unit. Often, the couples invite some videographers, whether professional or amateur, to document the wedding as their treasured memento of the ceremony. In this paper, wedding videos are defined as the raw, unedited footage recorded for wedding. Since a wedding video usually spans hours, the development of automatic tools for efficient content classification, indexing, and retrieval becomes crucial.

In this paper, we focus on the recognition of a wedding's group actions, namely wedding events, whereby a wedding is interpreted as a series of meaningful interactions among the participants. Based on the knowledge of wedding customs [1, 2], we define thirteen wedding events, such as the couple's wedding vows, ring exchange, etc. Our goal is to automatically segment a wedding video into a sequence of recognized wedding events. Without loss of generality, we focus on one of the most popular wedding styles, namely the western wedding, that follows the basic *western tradition* [1, 2] and takes place in a church-style venue. Based on our observations, a wedding video typically consists of four parts: preparation, guest seating, main ceremony, and reception. For simplicity, we deal with the third part alone because of its relative significance. In the rest of this paper the term wedding refers to the main ceremony.

In the literature, the study of wedding video analysis has long been ignored. The wedding video is simply to be treated as one of various content sources in home video researches [3, 4, 5]. However, several characteristics make the wedding ceremony videos much more challenging to be processed and analyzed as indicated in the following:

- Restricted spatial information: Since most of the wedding events occur in a single place (e.g., the front of a church altar) and participants basically stay motionless during the ceremony, the conventional scene, color, and motion based techniques [3, 4, 6] are not applicable for pre-partitioning a wedding video or for grouping "similar" shots as the basic unit for further event recognition. Likewise, most of the other content-generic visual features such as texture and edge are not reliable to be utilized.
- Temporally continuous capture: The extraction of broken time stamps is a widely used technique for generating shot candidates or event units of home videos [7, 8]. However, to avoid missing anything important, videographers usually capture a wedding, especially the main ceremony, in a temporally continuous manner without any interruption. As a result, the information of temporal logs is not so useful for wedding segmentation.

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, to republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee.

MIR '07, September 28–29, 2007, Augsburg, Bavaria, Germany.
Copyright 2007 ACM 978-1-59593-778-0/07/0009 ...\$5.00.

- **Implicit event boundary:** Although a wedding ceremony follows a definite schedule to proceed, the boundaries between wedding events are often implicit and unclear. For example, a groom’s entering to the venue is sometimes overlapped with the start of the bride’s entering. It is uneasy to determine an accurate change point for separating two events. This phenomenon not only increases the difficulty of accurate video segmentation but also adds uncertainties in annotating the event ground truth.

To recognize the thirteen wedding events, we adopt three kinds of audiovisual features, i.e., speech/music discriminator, flashlight detector, and bride indicator, as the basic modules to build our wedding video segmentation framework. Each wedding event is represented with a set of statistical models in terms of the extracted features. Since these features are selected based on the understanding of wedding customs [1, 2], they are more discriminative in distinguishing a wedding’s group actions than the aforementioned ones, such as motion and textures. To segment a wedding video, we develop a hidden Markov model (HMM) [9], in which every hidden state is associated with a wedding event and a state transition is governed by how likely the two corresponding wedding events take place in succession. The event sequence is, therefore, automatically determined by finding the most probable path. In summary, our event recognition framework not only uses the model similarity of extracted features, but simultaneously takes the temporal rationality of event ordering into account.

The main contributions of our work are twofold. First, an automatic system is proposed and realized for event-based wedding segmentation. To the best of our knowledge, this work is the first one to analyze and structure wedding videos at the semantic-event level. Even on the general domain of home videos, there is no similar work. The methodology could be extensively applied to the other kinds of home videos that possess similar characteristics as wedding, such as the birthday party and school ceremonies. Second, a taxonomy is developed to categorize the wedding events, whereby we adopted three kinds of discriminative features for robust event modeling and recognition. A superiority of these features is that they can be easily extracted from videos like the conventional ones, e.g., motion. Furthermore, the obtained high-level descriptions could benefit and complement the current research in semantic video understanding.

The rest of this paper is organized as follows. After a discussion of related work, Section 3 presents the taxonomy of wedding events. The extraction of event features and the modeling and segmentation of wedding videos are described in Section 4 and Section 5, respectively. Section 6 shows experimental results, and Section 7 presents our concluding remarks and the directions of future work.

2. RELATED WORK

In this section, we review previous studies on home video analysis. According to their applications, they are classified into four major categories: scene-based segmentation, capture-intent detection, photo-assisted summarization, and highlight extraction. Meanwhile, their pros and cons as compared with our work will be briefly discussed as well.

Scene-based segmentation. A basic segmentation pro-

cess is to cluster relevant shots into groups called scenes. A scene is defined as a subdivision of a video in which either the physical setting is fixed, or when it presents a continuous action in one place [4, 6]. Since the home video content tends to be close in time, the clustering can be simply confined to adjacent shots. Gatica-Perez *et al.* [3] proposed a greedy algorithm that initially treats each shot as a cluster and successively merges a pair of adjacent ones until a Bayesian criterion is violated. The merging order is determined through both the visual and temporal similarities, such as color, edge, and shot duration. Zhai *et al.* [4] located the scene boundaries using the optimization technique – Markov chain Monte Carlo (MCMC). A color-based similarity matrix is constructed for video shots, from which the clusters with high intra- and low inter-similarities are detected as the desired scenes.

Capture-intent detection. A capture-intent refers to an idea, feeling, theme, or message that makes us to capture certain video segments [5, 10], e.g., a sentimental sunset or baby laughing. Since the user’s capture-intent is often expressed through the use of cinematic principles, some researchers exploit the theory of computational media aesthetics [11]. Achanta *et al.* [5] proposed a framework for modeling the capture-intents of four basic emotions, i.e., cheer, serenity, gloom, and excitement. An emotion delivery system is also developed for helping users to enhance the original or convey a new emotion to a given home video. Mei *et al.* [10] further integrated the knowledge of psychology to classify the capture-intents into seven categories, such as close-up view, beautiful scenery, just record, etc. A learning-based mechanism for classifying the capture-intents is then presented using two kinds of feature sets: attention-specific and content-generic features.

Photo-assisted summarization. Personal photo albums can be viewed as an excellent abstract of the corresponding home videos. They share most of the important moments but the photo albums are relatively concise in presenting the contents. Since a still image can be applied to search videos, the summarization task is transformed into the problem of template matching between the two media. Aner-Wolf *et al.* [12] targeted on wedding videos. They represented each shot with one or several mosaics that are used to be aligned with the wedding photos. All shots with successful alignments are collected to generate a summarized video. Similar ideas are adopted by Takeuchi *et al.* [13], but they instead estimated the user’s general preferences on the summarization.

Highlight extraction. Highlights are the video segments with relatively higher semantic or perceptual attractions to users. Since the true understanding of video semantics cannot be achieved by the current computing technologies, the study of human attention models provides an alternative way for detecting perceptual highlights [14, 15]. Hua *et al.* [16] proposed a home video editing system, in which attention-based highlight segments are selected to be aligned with a given piece of incidental music to generate an edited highlight video. Meanwhile, a set of professional editing rules is utilized to optimize the editing quality, e.g., motion activity should match with music tempo. Abowd *et al.* [17] presented a semi-automatic approach for highlight browsing. Home videos need to be manually annotated with a predefined tag hierarchy that helps to group the highlight segments with similar semantic meanings, e.g., clips of all the child’s birthday wishing.

Table 1: Taxonomy of wedding events

Code	Event	Definition
<i>ME</i>	Main Group Entering [†]	Members of the main group walking down the aisle.
<i>GE</i>	Groom Entering	Groom (with the best man) walking down the aisle.
<i>BE</i>	Bride Entering	Bride (with her father) walking down the aisle.
<i>CS</i>	Choir Singing	Choir (with participants) singing hymns.
<i>OP</i>	Officiant Presenting	Officiants giving presentations, e.g., invocation, benediction, and homily.
<i>WV</i>	Wedding Vows	Couple exchanging wedding vows.
<i>RE</i>	Ring Exchange	Couple exchanging wedding rings.
<i>BU</i>	Bridal Unveiling	Groom unveiling his bride’s veil.
<i>MS</i>	Marriage License Signing	Couple (with officiants) signing the marriage license.
<i>WK</i>	Wedding Kiss	Groom kissing his bride.
<i>AP</i>	Appreciation	Couple thanking to certain people, e.g., their parents or all participants.
<i>ED</i>	Ending	Couple (followed by the main group) walking back down the aisle.
<i>OT</i>	Others	Any events not belonging to the above, e.g., lighting a unity candle.

[†] The main group indicates all persons, except the ones in *GE* and *BE*, who are invited to walk down the aisle, e.g., flower girls, ring bearers, groomsmen, bridesmaids, honorary attendants, officiants, etc.



Figure 1: Sample key-frames of the wedding events.

Overall, some observations can be made from the above discussions. First, the so-called event is a more semantic unit for video segmentation as compared with the conventional ones such as frame, subshot, shot, and scene [18, 19]. It represents a stand-alone human activity during a period of time. However, the relevant studies on home media are extremely rare as compared with the other kinds of content sources like sports [18]. Second, the analysis of home media are mostly from the perspective of a viewer or a videographer but not the actual owner or participants. Helping them to explicitly identify what had happened in a video often seems more crucial than simply indicating where would be more significant. These observed phenomena motivate our development of a comprehensive scheme for event-based video analysis and segmentation.

3. WEDDING EVENT TAXONOMY

According to the western tradition [1, 2], a wedding ceremony, whether religious or secular, begins when an assigned attendant (such as the officiant or the bride’s mother) is entering down the aisle and ends while the couple is walking out of the wedding venue. The mid-process may vary depending on the country, religion, local customs, and the wishes of the couple, but the basic elements that constitute the western weddings are almost the same [1, 2]. Therefore, we define thirteen wedding events as listed in Table 1. They are carefully specified to be mutually exclusive and collectively exhaustive. Figure 1 shows sample key-frames.

In addition to the traditions, the common perception of the relative event importance is also taken into account in our taxonomy for further applications such as highlight extraction or video summarization. For example, the three entering events (*ME*, *GE*, *BE*) are traditionally to be viewed as a unity called a processional [1, 2], but they should be explicitly separated because the couple’s arriving is generally much more exciting than others. By contrast, we classify all of the officiants’ formal presentations like invocation and benediction into a single wedding event (*OP*), because they are often invariable in form and the verbal expressions are basically predictable, often not beyond the scope of invoking the God’s blessing upon the marriage or inspiring the attendants’ religious spirits. It is evident that they are not as important as compared to the other ones.

Furthermore, as shown in Table 1, we can find that the taxonomy roughly follows the procession of a wedding ceremony, i.e., from the *ME* event to the *ED* event. However, it should be noted that the actual event ordering is based on each couple’s own wedding program and certain events could be repeated or removed in the ceremony. For example, the *OP* and the *CS* events are often interweaved with other ones. In addition, a simplified ceremony could only contain four events of *WV*, *RE*, *MS*, and *WK*.

4. EVENT FEATURE EXTRACTION

Three kinds of audiovisual features are exploited for event modeling: speech/music discriminator, flashlight detector,

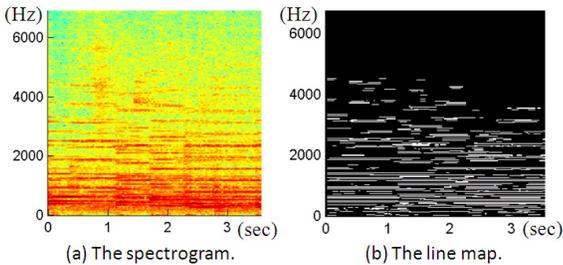


Figure 2: Example of a music signal with (a) its spectrogram using short-time Fourier transform and (b) its corresponding line map.

and bride indicator. In this section, we give mathematical definition for each of them and address the reasons why these features are adopted.

4.1 Speech/Music Discriminator

Traditionally, some of the wedding events contain purely speech and others are always accompanied with music [2]. For example, in the *OP* and the *WV* events, all participants keep quiet to listen to an officiant or the couple speaking. In the *CS* and the *BE* events, choir is singing with piano accompaniment or selected background music (e.g., Mozart’s Wedding March) is playing during the event. Obviously, discrimination between speech and music types from recorded audio plays a key role in wedding event recognition. Because the recorded audio quality is generally poor and often interfered with environmental sound and background noise, the selected speech/music discriminator has to be robust enough to handle such a low-SNR audio input.

Based on some previous studies [20, 21], we choose to use three audio features to build our discriminator: one-third energy crossing (OEC), silent interval frequency (SIF), and music component ratio (MCR) for their empirically stable performances under various noise types. Note that, in our approach, the audio track of wedding videos is converted to the 44 100 Hz mono-channel format first. For simplicity, let $x(n)$ be a discrete-time audio signal with time index n and N denotes the total number of samples in the interval from which features are extracted.

One-third Energy Crossing (OEC). One characteristic of a speech signal is that the corresponding amplitude has obvious variations. Given a fixed threshold δ , the number of audio energy waveform’s crossings over δ is often higher in a speech than that in a music. In this work, for each audio track, we empirically set δ to one-third of its whole range average amplitude. Therefore, OEC is defined as a measurement of the audio’s energy-spectral content as follows:

$$\text{OEC} \triangleq \frac{1}{2} \cdot \sum_{n=2}^N |\text{sign}_{\delta}(x^2(n)) - \text{sign}_{\delta}(x^2(n-1))| \quad (1)$$

where

$$\text{sign}_{\delta}(a) = \begin{cases} 1, & a > \delta \\ 0, & a = \delta \\ -1, & a < \delta \end{cases} \quad (2)$$

As suggested by the previous work [20, 22], the audio track is uniformly segmented into non-overlapping 1-second audio frames. For each audio frame, one feature value is computed

in every 20-ms interval and these 50 short-time feature values are averaged to generate the representative OEC feature for that 1-second frame. The same mechanism is used in SIF extraction, as described in the following paragraph.

Silent Interval Frequency (SIF). Since a speech signal is a concatenation of a series of syllables, it contains more pronouncing pauses than a music signal. Therefore, SIF is defined to measure the silent intervals of an audio signal as follows [20]:

$$\text{SIF} \triangleq I((E < \theta_l) \text{ or } (E < 0.1E_{max} \text{ and } E < \theta_h) \text{ or } (ZC = 0)) \quad (3)$$

where $I(\cdot)$ is the indicator function, E is the root mean square (RMS) of the signal amplitude, E_{max} is the maximum RMS value of the whole audio track, and ZC is the signal zero crossing. To be precise,

$$\text{RMS} \triangleq \sqrt{\sum_{n=1}^N x^2(n)} \quad (4)$$

and

$$\text{ZC} \triangleq \frac{1}{2} \cdot \sum_{n=2}^N |\text{sign}_0(x(n)) - \text{sign}_0(x(n-1))|. \quad (5)$$

In addition, the two thresholds θ_l and θ_h are empirically set to 0.5 and 2, respectively. As described in OEC extraction, we compute a representative SIF feature for each 1-second audio frame by taking average of 50 short-time SIF values.

Music Component Ratio (MCR). Harmonicity is the most prominent characteristic of a music signal. A music signal often contains spectral peaks at certain frequency levels and the peaks last for a period of time. This can be observed from the “horizontal lines” in the spectrogram of music, as shown in Figure 2. MCR is then defined as the average horizontal line number of an audio spectrogram within a second, and its extraction algorithm can be described as follows:

1. Segment the given audio track into 40-ms audio frames with a 10-ms overlap between two successive frames.
2. Compute the spectrogram (Figure 2(a)) of the audio frames using short-time Fourier transform.
3. Convert the spectrogram to a corresponding gray-level image by taking the absolute values of the Fourier coefficients.
4. Construct a line map (Figure 2(b)) from the image using Sobel operation [23], and a 7-order median filter is applied to remove outliers along each of the map rows.
5. Identify all horizontal lines in the line map using Hough transform [23].
6. For each second, calculate the line number from every 4-pixel-wide windows with 2-pixel advance in the line map, and take the average of the line numbers as the final MCR value.

4.2 Flashlight Detector

Wedding attendants, especially the couple’s family members and close friends, often take pictures during the ceremony, and the number of pictures taken roughly represents

Table 2: Examples of flashlight distributions of four successive wedding events in a ceremony.*

1. OP	2. WV	3. RE	4. WK
			
674 (sec)	234 (sec)	142 (sec)	12 (sec)
19 (times)	55 (times)	8 (times)	73 (times)
0.0282 (Hz)	0.2350 (Hz)	0.0563 (Hz)	6.0833 (Hz)

* The third to the fifth rows are the durations, flashlight numbers (manually counted) and flashlight densities of the corresponding wedding events, respectively.

the relative importance of a wedding event. Since the occurrence of camera flashlights correlates closely with the activity of picture-taking [24], the estimation of flashlight density could be an effective visual cue for wedding event discrimination. Table 2 shows an example of flashlight distributions of four successive wedding events in a ceremony. We observe high variance of flashlight distributions among events. For example, the *WK* event is merely 12 seconds long, but there are 73 flashlights. Its density reaches on average 6 times per second. By contrast, the *OP* event is of relatively less importance to the audiences as described in Section 3 and it contains a small number of flashlights even if it has a much longer duration.

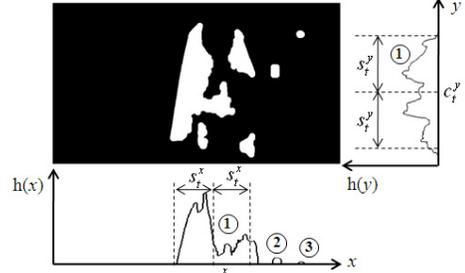
Specifically, flashlights can be detected from abrupt and short global frame intensity increases. In home videos, the durations of flashlights are seldom longer than two video frames. Therefore, in every 1-second interval, we compute a feature value of the flashlight density (FLD) as follows:

$$\text{FLD} \triangleq \sum_{t=2}^{M-1} I((\hat{f}_t^I - \hat{f}_{t-1}^I \geq \epsilon) \text{ and } (\hat{f}_t^I - \hat{f}_{t+1}^I \geq \epsilon)) \quad (6)$$

where M , \hat{f}_t^I are respectively the total number of video frames and the value of average intensity of the frame f_t , and the threshold $\epsilon = 5$ was suggested by the previous work [24] for flashlight detection.

4.3 Bride Indicator

As shown in Table 1, the main figures involved in various wedding events are not the same. For example, groom and the best man are the main characters in the *GE* event; the groom and his bride are the main figures in the *RE* event. The main figures' occurrence pattern gives a visual hint for the event category. A naïve solution would be to recognize all figures in videos. That is, however, not an easy task with today's technology. Fortunately, there are some simple trick to detect the bride, arguably the most important figure in the wedding. According to the western tradition [1, 2], the bride invariably wears white gown and veil as a symbol of purity but the others could have some flexibility in the dress color. Therefore, it is more reliable to indicate the bride's appearance using the truth of her wearing white. However, due to various lighting conditions, the determination of an accurate "bridal white" is extremely difficult and often needs a laborious training process as that of skin color detection [25]. In our current implementation, we compute an approximate bridal white map for a video frame using the following procedure:



(c) The bridal white map with projection histograms.

Figure 3: Example of (a) a video frame with (b) the thresholded image and (c) the bridal white map with projection histograms.

1. Convert a video frame f_t to the HSI color space [23], in which the values are within the range of $[0, 255]$.
2. Set empirically two thresholds ϕ_t^I and ϕ_t^S for the intensity and the saturation respectively for the bridal white:

$$\phi_t^I = \min(240, \hat{f}_t^I + 80) \text{ and } \phi_t^S = 75. \quad (7)$$

3. Construct a thresholded image $\bar{\Gamma}_t$ from the video frame using the above two thresholds, e.g., Figure 3(b). The thresholded image is defined as

$$\bar{\Gamma}_t(\mathbf{p}) = \begin{cases} 1, & \text{if } f_t^I(\mathbf{p}) \geq \phi_t^I \text{ and } f_t^S(\mathbf{p}) < \phi_t^S \\ 0, & \text{otherwise} \end{cases} \quad (8)$$

where \mathbf{p} is the pixel coordinate, and $f_t^I(\mathbf{p})$ and $f_t^S(\mathbf{p})$ denote \mathbf{p} 's intensity and saturation values, respectively.

4. Obtain a bridal white map Γ_t (cf. Figure 3(c)) by removing outliers of $\bar{\Gamma}_t$ using a morphological closing (i.e., erosion followed by dilation) [23]. That is

$$\Gamma_t = \bar{\Gamma}_t \circ Se \quad (9)$$

where Se is a disk structuring element whose radius is 5-pixel wide and \circ denotes the closing operation.

Further, the technique of projection histograms [23] is applied to improve reliability. Specifically, based on the observation that the bride roughly appears in the shape of a white vertical bar (Figure 3(a)), we add a spatial constraint that the white distribution in the vertical direction should be wider than that in the horizontal one. Therefore, we project the bridal white map along the x and y directions to construct two 1-D histograms (Figure 3(c)), from which the isolated component with the maximum white ratio is individually selected. For example, in Figure 3(c), there are three isolated components in the horizontal histogram but only one in the vertical one. We compute standard deviations, s_t^x and s_t^y , of the white distributions for the maximum

components of both axes. In every 1-second interval, a feature value of the bridal white ratio (BWR) is defined as

$$\text{BWR} \triangleq \frac{1}{M} \sum_{t=1}^M \Phi(\Gamma_t) \cdot I(s_t^x < s_t^y) \quad (10)$$

where $\Phi(\Gamma_t)$ returns the white ratio of Γ_t in terms of white pixel number with respect to the map size. Note that we use the average white percentage to avoid making the hard-decision on whether or not the bride does exist in frames.

5. WEDDING MODELING

The objective of wedding modeling is to estimate the event sequence of a wedding video. At each time instance, extracted event features are exploited to recognize the wedding events. On the other hand, a wedding video is a kind of sequential data. Thus, in wedding modeling, it needs not only to consider how likely the acquired features match an event candidate but also the temporal rationality whether the candidate is appropriate to follow the existing sequence immediately. Therefore, we use an effective learning tool, i.e., HMM, to describe the spatio-temporal relations within a wedding video [9]. In Sections 5.1 and 5.2, we first build statistical models of the feature similarity and the temporal ordering for each of the wedding events. Section 5.3 then devises an integrated HMM framework for both the event-based analysis and wedding segmentation.

Before we proceed, note that we divide the video uniformly into a sequence of 1-second units. The main reason for uniform pre-segmentation is that we can not use the conventional video units like shots as the basic units. It is because shots can't be reliably obtained using conventional techniques because of the reasons listed in Section 1. In addition, simplicity of uniform segmentation makes online processing possible in the future. For convenience, let E and F respectively denote the sets of integer enumerations of the wedding events and the event features, i.e., $E = \{i | i = ME, \dots, OT\}$ and $F = \{j | j = \text{OEC, SIF, MCR, FLD, BWR}\}$. Given the t -th video unit, let $\mathbf{e}_t \in E$ be a state variable that indicates the occurrence of a specific wedding event, and $\mathbf{x}_t = (x_t^1, \dots, x_t^{|F|})$ be the feature vector associated with the adopted event features x_t^j .

5.1 Wedding Event Modeling

For each of the wedding events, a statistical feature model is constructed for each of the adopted event features. Specifically, a feature model is a probability distribution describing the likeness of feature values. The use of statistical histograms [23] is a straightforward approach, but their discrete nature often causes unwanted discontinuity in results, especially when a feature value locates near the boundaries of histogram bins. Instead, we accumulate the probability by regarding each feature sample as a Gaussian centered at its own feature value. Assume that, for the i -th event, we have N value samples of the j -th feature $\{x_1^j, \dots, x_N^j\}$ extracted from training clips. The distribution $p_{i,j}$ of the j -th feature for the i -th event can then be obtained as follows:

$$p_{i,j}(\mathbf{x}) = \frac{1}{N} \sum_{n=1}^N \frac{1}{\lambda_j \sqrt{2\pi}} e^{-\frac{(\mathbf{x} - x_n^j)^2}{2(\lambda_j)^2}}, \quad \forall i \in E, \quad \forall j \in F \quad (11)$$

where $\int_{\mathbf{x}=-\infty}^{\infty} p_{i,j}(\mathbf{x}) d\mathbf{x} = 1$ and λ_j is a confidence parameter specifying how we trust the extracted values of the j -th

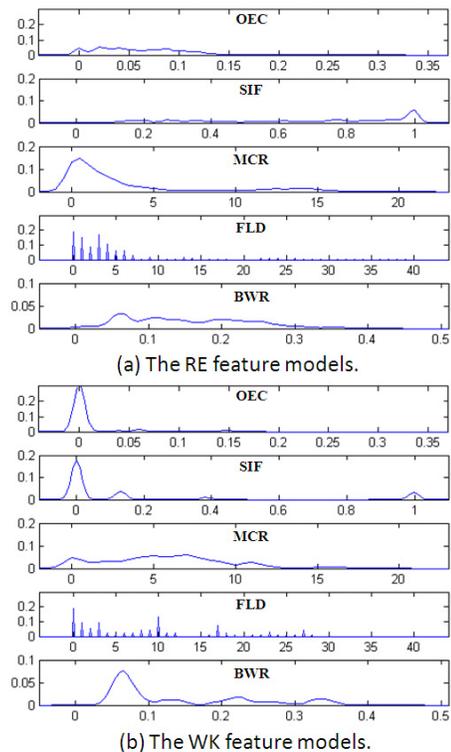


Figure 4: Examples of wedding event models of (a) the RE event and (b) the WK event.

feature. If the extracted feature samples are more accurate and reliable, we can set the parameter to a smaller value.

Since the feature models are used for discriminating the wedding events, the divergence among feature models of different events should be as large as possible. Quantitatively, the divergence of two probability distributions is defined by the symmetric Kullback-Leibler (SKL) distance [26]:

$$D_{SKL}(\mathbf{p}, \mathbf{q}) = \frac{1}{2} \int_y \left[\mathbf{p}(y) \log \frac{\mathbf{p}(y)}{\mathbf{q}(y)} + \mathbf{q}(y) \log \frac{\mathbf{q}(y)}{\mathbf{p}(y)} \right] dy \quad (12)$$

For the j -th feature, the confidence parameter λ_j is chosen to maximize the sum of divergences among the same kind of feature models as follows:

$$\lambda_j = \arg \max_{\lambda} \sum_{i < k, i, k \in E} D_{SKL}(p_{i,j}, p_{k,j}), \quad \forall j \in F \quad (13)$$

To find the optimal λ_j , we use exhausted search and empirically set a search range (e.g., $[0, 1]$) with a desired precision (e.g., 0.05). The optimal confidence parameters we found are $\lambda_{OEC} = 0.005$, $\lambda_{SIF} = 0.015$, $\lambda_{MCR} = 0.5$, and $\lambda_{BWR} = 0.01$. It is worthy to notice that the FLD feature is an exception because its values are discrete. As a result, we manually set $\lambda_{FLD} = 0$ and apply a 9-point normalized filter to sample sequences of the FLD feature as an alternative to the Gaussian-based smoothing.

Therefore, given a video unit (e.g., the t -th one), we can compute the probability that we observe \mathbf{x}_t given this video unit belongs to the i -th wedding event:

$$p(\mathbf{x}_t | \mathbf{e}_t = i) = \prod_{j=1}^{|F|} p_{i,j}(x_t^j), \quad \forall i \in E. \quad (14)$$

Table 3: An adjacency matrix of the wedding events.

	ME	GE	BE	CS	OP	WV	RE	BU	MS	WK	AP	ED	OT
ME	1	1	1										
GE	1	1	1										
BE			1	1	1							1	1
CS				1	1								
OP				1	1	1	1			1	1	1	1
WV						1	1	1					1
RE						1	1	1					1
BU								1					
MS					1				1		1	1	
WK				1	1				1	1			1
AP				1					1		1	1	
ED												1	
OT				1	1				1				

Note that in practice we compute the log-likelihood by taking logarithm of the expression, and thus can give a contributive weight κ_j to the j -th feature model where $\sum_j \kappa_j = 1$. Overall, the proposed event modeling has several advantages. First, it has more tolerance to inaccuracy and uncertainty of the extracted event features. The Gaussian component helps to reduce and diversify the influence of an inaccurate feature value. Second, it avoids the artifacts due to quantization errors in the constructed feature models. The distribution of feature values is faithfully presented without approximation. Figure 4 gives examples of feature statistical models for two wedding events, *RE* and *WK*.

5.2 Event Transition Modeling

The event transition model (ETM) is constructed to describe the probability that a wedding event is immediately followed by another in a wedding ceremony. In other words, it evaluates whether a temporal transition is to be allowed between each pair of the wedding events. Therefore, the ETM can be defined by an $|E| \times |E|$ matrix A as follows:

$$A_{i,k} = Pr(\mathbf{e}_t = k | \mathbf{e}_{t-1} = i), \forall i, k \in E \quad (15)$$

where $A_{i,k}$ is the entry of the i -th row and the k -th column of A , and $t-1, t$ are two successive time instances in seconds. Since all possible transitions are enumerated in A , the marginal probability along each row is unity:

$$\sum_{k=1}^{|E|} A_{i,k} = 1, \forall i \in E. \quad (16)$$

In fact, given a training set of wedding videos with the event ground truth, we can tabulate an approximation of the ETM, namely \tilde{A} . However, the obtained probability distributions are often extremely biased. That is, most of the probabilities are prone to centralize on the diagonal entries, i.e., $\tilde{A}_{i,i}$. This phenomenon is due to the transitions are counted in seconds. For example, assuming that we have two successive events which are both 100 seconds long, only one event transition will be accounted during this 200-second period. Therefore, for each row of \tilde{A} (e.g., the i -th one), we exploit a regularization to balance the probabilities:

$$A_{i,k} = \begin{cases} \omega_i \tilde{A}_{i,k} & , i = k \\ (1 - \omega_i \tilde{A}_{i,i} / 1 - \tilde{A}_{i,i}) \cdot \tilde{A}_{i,k} & , i \neq k \end{cases}, \forall k \in E \quad (17)$$

where ω_i is the regularization factor in the range of $[0, 1]$. To be precise, we shift some of the diagonal probabilities to the off-diagonal ones but keep their relative ratios unchanged. Empirically, all of the diagonal entries are regularized to take approximate 80% probabilities along each row, i.e., $A_{i,i} \approx 0.8$, after regularization.

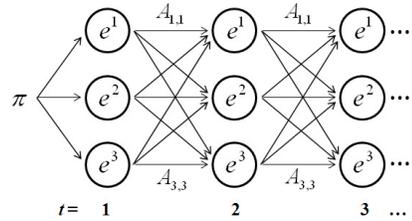


Figure 5: A simplified example of the HMM for wedding segmentation. (See Subsection 5.3 for details.)

Table 3 shows a simplified version for the real ETM we learnt from training videos, called an adjacency matrix, in which the entries with nonzero probabilities in the original ETM are marked as “1” in A . Sparsity of the adjacency matrix shows that few kinds of event transitions are allowed. It also demonstrates the occurrence of wedding events has a strong temporal correlation. This fact reduces the computation cost and increases the reliability of the determined event sequence.

5.3 Wedding Segmentation Using HMM

HMM is a specific instance of state space models, in which the concept of hidden states is introduced to recognize the temporal pattern of a Markov process [9]. Since the sequence of wedding events can be viewed as a first-order Markov data as seen in Section 5.2, we exploit an HMM framework for segmenting wedding videos, in which the wedding event statistical models (Section 5.1) and the event transition model (Section 5.2) are integrated together.

Specifically, given an input wedding video V , it is first partitioned into N 1-second video units, $V = \{\mathbf{v}_1, \dots, \mathbf{v}_N\}$. For each video unit \mathbf{v}_t , $t \in \{1, \dots, N\}$, we have a set of $|F|$ event features associated with it, i.e., $\mathbf{x}_t = (x_t^1, \dots, x_t^{|F|})$. Collecting all the observations $X = \{\mathbf{x}_1, \dots, \mathbf{x}_N\}$, our goal is to find the most probable event sequence S for V , where $S = \{\mathbf{e}_1, \dots, \mathbf{e}_N\}$. Therefore, we develop a left-to-right HMM of $|E|$ states $\{e^i | i \in E\}$, in which each state corresponds to one of the adopted event categories. The HMM is governed by a set of parameters, $\theta = \{\pi, A, \phi\}$, where π , A , and ϕ define the initial state probabilities, the state transition probabilities, and the emission probabilities, respectively [9]. Figure 5 illustrates a trellis representation of an simplified HMM with only three states. Clearly, ϕ and A have been explicitly described by the wedding event models and the event transition model, respectively. Without loss of generality, π is given by a uniform distribution, i.e., $p(\mathbf{e}_1 = i | \pi) = 1/|E|, \forall i \in E$. Accordingly, our goal to find the optimal sequence S formulated as

$$\begin{aligned} S &= \arg \max_s Pr(X, S | \theta) \\ &= \arg \max_s p(\mathbf{e}_1 | \pi) \left[\prod_{t=2}^N p(\mathbf{e}_t | \mathbf{e}_{t-1}, A) \right] \prod_{t=2}^N p(\mathbf{x}_t | \mathbf{e}_t, \phi) \\ &= \arg \max_s p(\mathbf{e}_1 | \pi) \left[\prod_{t=2}^N A_{\mathbf{e}_{t-1}, \mathbf{e}_t} \right] \prod_{t=2}^N \prod_{j=1}^{|F|} p_{\mathbf{e}_t, j}(x_t^j) \end{aligned} \quad (18)$$

where the second and the third terms are derived from Equations 14 and 15, respectively. Because the HMM trellis is equivalent to a directed tree (cf. Figure 5), the solution of S can be efficiently obtained using the Viterbi algorithm [9].

Table 4: The collection of six wedding videos used in the experiments.

Clip	A	B	C	D	E	F
						
Duration	2215 (sec)	410 (sec)	4122 (sec)	3790 (sec)	1062 (sec)	1350 (sec)
Event #	17	8	35	23	15	14

Table 5: Gaussian distributions $N(\mu, \sigma^2)$ of the event durations for all wedding events in the video collection.

Event	ME	GE	BE	CS	OP	WV	RE	BU	MS	WK	AP	ED	OT
(a) from all event samples													
μ_i	92.00	42.33	114.00	139.90	130.91	163.33	135.50	47.33	166.00	11.60	68.33	75.20	149.08
σ_i	38.11	36.25	67.73	104.62	182.28	61.71	13.20	6.66	62.60	1.14	6.66	13.48	67.13
(b) from half of the event samples with shorter durations													
$\tilde{\mu}_i$	45.33	19.00	37.00	56.64	54.24	88.50	111.67	38.67	132.50	10.00	61.33	51.33	97.63
$\tilde{\sigma}_i$	15.95	5.57	1.41	32.08	32.16	26.16	23.63	8.39	33.23	1.00	5.51	24.01	40.17

Table 6: Recognition results of the wedding events where each number is in unit of seconds.

Events	ME	GE	BE	CS	OP	WV	RE	BU	MS	WK	AP	ED	OT	RR(%)
ME	547	0	32	0	0	0	0	0	0	0	0	0	0	94.47
GE	25	99	18	0	0	0	0	0	0	0	0	0	0	69.72
BE	91	0	339	0	0	0	0	0	0	0	0	0	0	78.84
CS	0	0	0	2279	58	0	70	71	170	0	16	9	0	85.26
OP	4	0	0	75	3697	203	484	0	35	12	3	2	127	79.64
WV	0	0	0	0	0	773	22	0	0	0	0	0	0	97.23
RE	0	0	0	0	59	63	553	0	0	0	0	0	0	81.93
BU	0	0	0	23	6	0	0	156	0	0	0	0	0	84.32
MS	0	0	0	0	76	0	0	33	156	0	0	0	0	58.87
WK	0	0	0	11	0	3	0	0	0	76	0	0	0	84.44
AP	52	0	0	0	0	0	3	0	0	0	166	0	0	75.11
ED	0	0	0	0	0	0	0	0	0	0	0	430	0	100.00
OT	0	0	0	509	345	113	113	15	0	0	123	0	604	33.15
RP(%)	76.08	100.00	87.15	78.67	87.17	66.93	44.42	56.73	43.21	86.36	53.90	97.51	82.63	

Therefore, in the input video V , the temporal extent of a detected wedding event, or called an event segment, is defined by collecting successive video units with the same event labeling. Finally, a smoothing scheme is applied to reduce possible labeling noises. Since, in general, a wedding event lasts for at least tens of seconds, we remove the short ones (less than 10 seconds in duration) by merging it into its neighbors. If its both neighbors belong to different event categories, it is merged into the left one; otherwise, all the three events are merged into one event.

6. EXPERIMENTAL RESULTS

This section presents experimental results for the evaluation of the proposed framework in wedding event recognition (Section 6.1) and wedding ceremony video segmentation (Section 6.2). Table 4 summarizes the statistics of the videos used in the experiments.

Currently, we have a total of six wedding video clips, each of them contains a complete recording of a ceremony. Three observers (none of the clip owners) collaboratively annotate the event ground truth. Table 4 also reports durations and numbers of the annotated events for all six videos. The following experiments were performed using a leave-one-out cross-validation strategy, in which models were trained from five clips and tested on the remaining one, and the whole training-testing procedure was iterated six times.

6.1 Event Recognition Analysis

Table 6 summarizes the event recognition results in unit of seconds, presented in the form of confusion matrix [22], where the leftmost column represents the actual event categories while the top-most row indicates the resultant ones recognized by the HMM framework. The confusion matrix is accumulated from results of all clips in the collection. The recognition precision (RP) and the recognition recall (RR) for each of the event categories are reported in Table 6. As described in Section 1, since the actual boundaries between wedding events are not always precise, the recognition result of a video unit is claimed to be correct if it hits the ground truth within a tolerant range. Instead of setting a universal range value, we adopt a dynamic setting scheme based on the recognized event categories because the event durations vary greatly as shown in Table 5(a). For each event category, we compute a truncated mean $\tilde{\mu}_i$ of the event samples by discarding half of the samples that are longer in duration (Table 5(b)), and then the range value is set to $\min(0.2\tilde{\mu}_i, \xi)$, where we set $\xi = 10$ so that the overlaps between events vary according to event categories but do not exceed 10 seconds. Here, we use a truncated mean but not the standard mean since durations of the shorter samples are more consistent and their average value would be more reliable, which can be observed from that the truncated variance is much smaller than the standard one as shown in Table 5.

Overall, as shown in Table 6, a large amount of the detected wedding events reach over 70% in both RP and RR

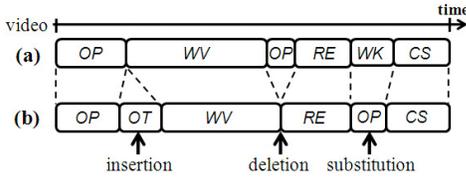


Figure 6: Edit operations for transforming (a) a reference event string to (b) the comparing one.

values. Some of them even achieve the level of 85%, such as *WK* and *ED* events. Several observations were made from this table: 1) A few recognition errors are associated with *CS* and *OP* events, especially the later one. This phenomenon is usually unavoidable because a wedding event, such as *OP* or *MS* event, is sometimes arranged to accompany with choirs singing and the whole ceremony is generally hosted by wedding officiants who would like to give some short presentations within a wedding event. They also cause severe degradations in RP values for both *RE* and *BU* events. 2) The confusion matrix is sparse and the recognition errors show grouping effects. That is, the wedding events of a similar group are prone to be mis-classified to each other, e.g., the set of the entering events (*ME*, *GE*, *BE*) and the set of the couple’s committing events (*WV*, *RE*). From Table 3, we can find that the events of each event set correspond to the ones that are more probable to occur in succession. Thus the recognition errors partially come from the implicit event boundaries. 3) The RR value of the *OT* event is relatively low. This is due to the fact that *OT* event is inherently varied in form. For example, it could be reading of poetry or lighting of the unity candle. Compared with the other kinds of wedding events, *OT* event is the most difficult one to be modeled. Moreover, it severely influences the overall recognition performance by spreading out the recognition errors over various event categories.

6.2 Video Segmentation Analysis

In this section, we further evaluate the segmentation performance of our approach. Since in practice the temporal extent of a wedding event is perceived as a whole by users, the segmentation results are compared at the event level but not the second level. We follow a similar idea exploited in the longest common substring problem [27]. That is, we represent a wedding video as a symbol string where the alphabet consists of the event codes in Table 1. Note that the symbol string is generated in unit of detected events, and each symbol corresponds to an event segment of the wedding video. Therefore, for each of the testing wedding clips, the segmentation performance is measured by the number of the required edit operations (substitution, insertion, and deletion) for transforming the reference string corresponding to the ground truth into the string corresponding to the recognition result. Figure 6 shows an example of transforming strings. The less the edit operations are needed, the better the segmented videos match the ground truth.

Table 7 shows the statistics. We claim an event segment as correct if it hits the ground truth in more than 80% of its duration. The resultant segmentation precision (SP) and the segmentation recall (SR) are then defined as follows:

$$SP = \frac{\text{Corrects}}{\text{Corrects} + \text{Substitutions} + \text{Insertions}} \cdot 100\%, \quad (19)$$

Table 7: Segmentation results in which each number is in unit of event segments (without duration-based filtering).

Clip	Corr.	Sub.	Ins.	Del.	SP(%)	SR(%)	SF(%)
A	16	1	12	0	55.17	94.12	69.57
B	5	1	0	2	83.33	62.50	71.43
C	28	1	23	6	53.85	80.00	64.37
D	22	1	19	0	52.38	95.65	67.69
E	12	0	6	3	66.67	80.00	72.73
F	12	1	10	2	52.17	85.71	64.86
Avg.					60.60	83.00	70.05

Table 8: Segmentation results in which each number is in unit of event segments (with duration-based filtering).

Clip	Corr.	Sub.	Ins.	Del.	SP(%)	SR(%)	SF(%)
A	16	1	5	0	72.73	94.12	82.05
B	5	1	0	2	83.33	62.50	71.43
C	26	1	12	8	66.67	74.29	70.27
D	21	1	13	1	60.00	91.30	72.41
E	12	0	3	3	80.00	80.00	80.00
F	11	0	6	3	64.71	78.57	70.97
Avg.					71.24	80.13	75.42

$$SR = \frac{\text{Corrects}}{\text{Corrects} + \text{Substitutions} + \text{Deletions}} \cdot 100\%. \quad (20)$$

In addition, the F-measure, $SF = 2 \cdot SP \cdot SR / (SP + SR)$, is provided as a metric for evaluating the integral performance.

From Table 7, we can see that SR values generally achieve 80% high, i.e., most of the event segments are correctly identified. A low value of Clip-B comes mostly from its small event number in the ground truth as shown in Table 4. By contrast, the overall SP values are relatively low, at the level of 60%. Compared with the ground truth, a large amount of redundant events are erroneously “inserted” in the segmentation results by our approach. These are mainly caused by the following two reasons. First, the erroneous events are generated in a one-to-many pattern. A single event that has been deleted from the ground truth usually turns into a series of successive erroneous ones in the resultant event sequence. For example, a deleted *OT* event would result in a catenation of *CS* and *OP* events. Second, the erroneous events are prone to exist around an event boundary of the ground truth. The same phenomenon has been observed from the recognition errors as reported in Section 6.1.

Since the erroneous events are “mutated” from parts of the original event segments, in general, they have a shorter duration as compared with the same kind of wedding events. Therefore, we use a duration-based filtering scheme to identify and possibly correct the abnormal ones. Specifically, for each of the event categories, we exploit the truncated models (Section 6.1 and Table 5(b)) to determine a lower bound of the reasonable event duration, i.e., $\Omega_i = \tilde{\mu}_i - \alpha_i \tilde{\sigma}_i$, where a rational scalar α_i is empirically set within the range of [1.5, 2]. If an event segment is recognized as the i -th event category and its duration is less than Ω_i , we merge it into its left neighbor in our current implementation. Table 8 summarizes the segmentation results after applying the duration-based filtering. Compared with Table 7, the number of inserted erroneous events is effectively reduced

and on average a 10% improvement is obtained for SP values. This improvement is accompanied by a slight decrease in SR values because some correct events would be filtered out at the same time.

Overall, as shown in Table 8, the integrated performance of our system is satisfactory. It achieves the level of 70% in terms of the SF metrics. Furthermore, with the assist of the duration-based filter, the tendencies of both SP and SR behaviors are much more balanced and consistent. The statistical results may not be comprehensive but it is encouraging. It gives us support and confidence that, as long as we capture well the content characteristics, we are able to conduct high-level semantic analysis of home videos through the use of generic and easily extracted audiovisual features. That is also an advantage of the proposed framework to be plausible for real applications.

7. CONCLUSIONS

In this paper, we proposed and realized a system for event-based wedding analysis and segmentation. According to the wedding customs, we developed a taxonomy for classifying the wedding events, whereby three kinds of discriminative high-level features are exploited for robust event modeling and recognition. To the best of our knowledge, this work is the first one to analyze and structure wedding videos on the basis of semantic events. Therefore, it can help users to access, organize, and retrieve his/her treasured contents in an automatic and more efficient way. Many aspects of our approach can be improved. First, it is possible to explore more semantic features for event recognition. For example, speaker recognition would be helpful for discriminating the events of dense speech, such as *WV* and *RE* events. Next, more extensive and complete evaluation of our system is a must. Meanwhile, it is crucial to have a common evaluation benchmark for wedding videos. In the future, we will continue our investigation in these directions.

8. ACKNOWLEDGMENTS

This work was partially supported by the National Science Council of R.O.C. under grants NSC 95-2622-E-002-018, NSC 95-2752-E-002-006-PAE, and NSC 95-2221-E-002-332. It was also supported by National Taiwan University under grant 95R0062-AE00-02.

9. REFERENCES

- [1] L. M. Spangenberg. *Timeless Traditions: A Couple's Guide To Wedding Customs Around The World*. Universe Publishing, 2001.
- [2] D. Warner. *Diane Warner's Contemporary Guide to Wedding Ceremonies*. New Page Books, 2006.
- [3] D. Gatica-Perez, A. Loui, and M.-T. Sun. Finding structure in home videos by probabilistic hierarchical clustering. *IEEE TCSVT*, 13(6):539–548, June 2003.
- [4] Y. Zhai and M. Shah. Automatic segmentation of home videos. *Proc. IEEE ICME'05*, pages 9–12, 2005.
- [5] R. S.V. Achanta, W.-Q. Yan, and M. S. Kankanhalli. Modeling intent for home video repurposing. *IEEE MM*, 13(1):46–55, Jan.-Mar. 2006.
- [6] Z. Rasheed and M. Shah. Detection and representation of scenes in videos. *IEEE TMM*, 7(6):1097–1105, Dec. 2005.
- [7] P. Yin, X.-S. Hua, and H.-J. Zhang. Automatic time stamp extraction system for home videos. *Proc. IEEE ISCAS'02*, pages 73–76, 2002.
- [8] M. Cooper, J. Foote, A. Girgensohn, and L. Wilcox. Temporal event clustering for digital photo collections. *Proc. ACM MM'03*, pages 364–373, 2003.
- [9] C. M. Bishop. *Pattern Recognition and Machine Learning*. Springer, 2006.
- [10] T. Mei, X.-S. Hua, H.-Q. Zhou, and S. Li. Modeling and mining of users' capture intention for home videos. *IEEE TMM*, 9(1):66–77, Jan. 2007.
- [11] C. Dorai and S. Venkatesh. Computational media aesthetics: Finding meaning beautiful. *IEEE MM*, 8(4):10–12, Oct.-Dec. 2001.
- [12] A. Aner-Wolf and L. Wolf. Video de-abstraction or how to save money on your wedding video. *Proc. IEEE WACV'02*, pages 264–268, 2002.
- [13] Y. Takeuchi and M. Sugimoto. Video summarization using personal photo libraries. *Proc. ACM MIR'06*.
- [14] Y.-F. Ma, L. Lu, H.-J. Zhang, and M. Li. A user attention model for video summarization. *Proc. ACM MM'02*, pages 533–542, 2002.
- [15] W.-H. Cheng, C.-W. Wang, and J.-L. Wu. Video adaptation for small display based on content recomposition. *IEEE TCSVT*, 17(1):43–58, Jan. 2007.
- [16] X.-S. Hua, L. Lu, and H.-J. Zhang. Optimization-based automated home video editing system. *IEEE TCSVT*, 14(5):572–583, May 2004.
- [17] G. D. Abowd, M. Gauger, and A. Lachenmann. The family video archive: An annotation and browsing environment for home movies. *Proc. ACM MIR'03*.
- [18] M. Chen, S.-C. Chen, M.-L. Shyu, and K. Wickramaratna. Semantic event detection via multimodal data mining. *IEEE SPM*, Mar. 2006.
- [19] J.-H. Lim, Q. Tian, and P. Mulhem. Home photo content modeling for personalized event-based retrieval. *IEEE MM*, 10(4):28–37, Oct.-Dec. 2003.
- [20] C. Panagiotakis and G. Tziritas. A speech/music discriminator based on rms and zero-crossings. *IEEE TMM*, 7(1):155–166, Feb. 2005.
- [21] T. Zhang and C.-C. J. Kuo. *Content-based Audio Classification and Retrieval for Audiovisual Data Parsing*. Kluwer, 2001.
- [22] Y. Li and C. Dorai. Instructional video content analysis using audio information. *IEEE Trans. Audio, Speech, Lang. Process.*, 14(6):2264–2274, Nov. 2006.
- [23] R. C. Gonzalez and R. E. Woods. *Digital Image Processing, 2ed*. Prentice-Hall, 2001.
- [24] B. T. Truong and S. Venkatesh. Determining dramatic intensification via flashing lights in movies. *Proc. IEEE ICME'01*, pages 61–64, 2001.
- [25] S. L. Phung *et al.* Skin segmentation using color pixel classification: Analysis and comparison. *IEEE TPAMI*, 27(1):148–154, Jan. 2005.
- [26] T. M. Cover and J. A. Thomas. *Elements of Information Theory, 2ed*. Wiley, 2006.
- [27] T. G. Cormen, C. E. Leiserson, and R. L. Rivest. *Introduction to algorithms, 2ed*. MIT Press, 2001.