Semantic Analysis for Automatic Event Recognition and Segmentation of Wedding Ceremony Videos

Wen-Huang Cheng, *Student Member, IEEE,* Yung-Yu Chuang, *Member, IEEE,* Yin-Tzu Lin, Chi-Chang Hsieh, Shao-Yen Fang, Bing-Yu Chen, *Member, IEEE,* and Ja-Ling Wu, *Fellow, IEEE*

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 recognizable wedding events, e.g. the couple's wedding kiss. Our goal is to develop an automatic tool that helps users to efficiently organize, search, and retrieve his/her treasured wedding memories. Furthermore, the obtained event descriptions could benefit and complement the current research in semantic video understanding. Based on the knowledge of wedding customs, a set of audiovisual features, relating to the wedding contexts of speech/music types, applause activities, picture-taking activities, and leading roles, are exploited to build statistical models for each wedding event. Thirteen wedding 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 collection of wedding videos and the promising results demonstrate the effectiveness of our approach. Comparisons with conditional random fields show that the proposed approach is more effective in this application domain.**

Index Terms— **Home videos, wedding ceremonies, semantic content analysis, event detection, video segmentation.**

I. 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 new 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 refer to the raw, unedited footage recorded for wedding. Since a wedding video usually spans hours, the development of automatic tools for efficient content classification, indexing, searching, 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 participants. Based on the knowledge of wedding customs [1], [2], we define thirteen wedding events, such as the couple's wedding vows. Our goal is to automatically segment a wedding video into a sequence of recognizable wedding events. Without loss of generality, we focus on one of the most popular wedding styles, 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 research on home videos [3], [4], [5]. Although it shares some common properties with other kinds of home videos, such as frequent poor-quality contents and unintentional camera operations [3], [4], several characteristics make it much more challenging to be analyzed:

- 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 techniques based on scene, color, and motion [3], [4], [6] are not applicable to pre-partition a wedding video or to group "similar" shots into basic units 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 temporal logs are not useful for wedding segmentation.
- Implicit event boundary: Although a wedding ceremony proceeds following a definite schedule, 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 not easy to determine an accurate change point to separate two events. This phenomenon not only increases the difficulty of accurate video segmentation but also adds uncertainties to annotate the event ground truth.

Manuscript received February 29, 2008; revised July 4, 2008. This work was partially published in the ACM Workshop on Multimedia Information Retrieval (MIR), 2007 [29]. 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.

Wen-Huang Cheng, Yung-Yu Chuang, Bing-Yu Chen, and Ja-Ling Wu are with the Graduate Institute of Networking and Multimedia, National Taiwan University, Taipei 10617, Taiwan, R.O.C. (e-mail:{wisley, cyy, robin, wjl}@cmlab.csie.ntu.edu.tw).

Yin-Tzu Lin, Chi-Chang Hsieh, and Shao-Yen Fang are with the Department of Computer Science and Information Engineering, National Taiwan University, Taipei 10617, Taiwan, R.O.C. (e-mail:{known, nonrat, strawinsky}@cmlab.csie.ntu.edu.tw)

TABLE I
TAXONOMY OF WEDDING EVENTS

Code	Event	Definition
ME	Main Group Entering[†]	Members of the main group walking down the aisle.
GE	Groom Entering	Groom (with the best man) walking down the aisle.
BE	Bride Entering	Bride (with her father) walking down the aisle.
CS	Choir Singing	Choir (with participants) singing hymns.
OP	Officiant Presenting	Officiants giving presentations, e.g. invocation, benediction, and homily.
WV	Wedding Vows	Couple exchanging wedding vows.
RE	Ring Exchange	Couple exchanging wedding rings.
BU	Bridal Unveiling	Groom unveiling his bride's veil.
MS	Marriage License Signing	Couple (with officiants) signing the marriage license.
WK	Wedding Kiss	Groom kissing his bride.
AP	Appreciation	Couple thanking to certain people, e.g. their parents or all participants.
ED	Ending	Couple (followed by the main group) walking back down the aisle.
OT	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.

To recognize the thirteen wedding events, we adopt a set of audiovisual features, relating to the wedding contexts of speech/music types, applause activities, picture-taking activities, and leading roles, as the basic event features to build our wedding video segmentation framework. Each wedding event is represented by 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 wedding events than the aforecited features, such as motion and textures. To effectively 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 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. More information about this work can be found at our website [10].

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 wedding videos at the semantic-event level. For any type of home videos, our work might also be the first one to achieve the semantic event analysis. The proposed 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 a set of carefully selected audiovisual features for robust event modeling and recognition. The true power of these features is that they are effective in discriminating various wedding events but their extractions from videos are as easy as the conventional ones. Furthermore, the obtained high-level descriptions could benefit the current research in semantic video understanding.

The rest of this paper is organized as follows. Section II presents the taxonomy of wedding events. The extraction of event features and the modeling and segmentation of wedding videos are described in Section III and Section IV, respectively. Section V depicts the experimental results, and Section VI presents our concluding remarks and the directions of future work.

II. Wedding Event Taxonomy

According to the western tradition [1], [2], a wedding ceremony, whether religious or secular, begins when an assigned attendant (such as an officiant) is entering down the aisle and ends while the couple is walking out of the wedding venue. The mid-process may vary depending on countries, religions, 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 I. See our prior work [29] for sample key frames and more details.

III. Event Features Development and Extraction

Effective event modeling is built on top of reliable event features. The understanding of wedding customs [1], [2] gives valuable insights to the process of feature exploration. Several key observations, which are found to be useful in discriminating the wedding events, are first presented in Section III-A. In Section III-B, we develop corresponding audiovisual features, including four audio features and two visual features. They are collected together as event features for later event modeling.

A. Key Observations

According to the western traditions [1], [2], wedding events are observed to behave differently in four main aspects: speech/music types, applause activities, picture-taking activities, and leading roles. In the following, we explain in detail for each of the key observations and then give corresponding guidance on the development of relevant event features.

TABLE II

THE TENDENCY OF WEDDING EVENTS IN THEIR BEHAVIOR OF SPEECH/MUSIC TYPES, APPLAUSE ACTIVITIES, PICTURE-TAKING ACTIVITIES, AND LEADING ROLES.*

	S/M[a]	App.[b]	Pic.[c]	Leading Roles[d]
ME	–	N	L^+	main group
GE	–	N	–	groom, (best man)
BE	M	–	H^+	bride, (bride's father)
CS	M	–	L^-	choir, (wedding participants)
OP	S	N	–	officiants
WV	S	N	H^-	bride, groom, officiants
RE	S	N	H^-	bride, groom, officiants
BU	S	–	H^-	bride, groom
MS	–	N	–	bride, groom, (officiants)
WK	–	Y	H^+	bride, groom
AP	–	Y	–	bride, groom, (wedding participants)
ED	M	Y	H^-	bride, groom, (main group)
OT	–	–	–	–

* "–" in the blanks means no obvious tendency.

[a] S: speech events, M: music events.

[b] Y: applause events, N: non-applause events.

[c] L^-, L^+, H^-, H^+: events with the activity of picture-taking from low to high.

[d] People in parentheses are optional.

1) *Speech/Music Types:* Traditionally, some wedding events contain purely speech and others are 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, a choir is singing with piano accompaniment or the selected background music (e.g. Mozart's Wedding March) is played during the event. The tendency of wedding events in speech/music types is shown in Table II. Obviously, the discrimination between speech and music types from recorded audio plays a key role in wedding event recognition. However, because the quality of the recorded audio is generally poor and often interfered with environmental sound and background noise, the selected audio features related to the speech/music discrimination have to be robust enough to survive such a low-SNR audio input.

2) *Applause Activities:* Applause is usually expected from wedding attendants as the expression of approval or admiration at certain moments during the ceremony. For example, in the *WK* and the *ED* events, the couple routinely receives a burst of applause at the moments when they are kissing or walking back down the aisle. By contrast, in the *OP* and the *WV* events, wedding attendants rarely applaud in order to keep the solemnity and avoid interfering with the ongoing wedding speech. Thus, effective applause detection is beneficial to the recognition of wedding events, cf. Table II. Note that, for our applications, the applause especially refers to the ones created by a group of people rather than by an individual. Specifically, the applause is generated by the group act of hands clapping and naturally the group members tend to clap at slightly different rates. This phenomenon makes the sound of applause difficult to be analyzed without the use of prior knowledge [11], [12]. Therefore, a common technique is to exploit the physical properties of applause [12], [13] to identify its appearance in the audio track of wedding videos.

TABLE III

EXAMPLES OF FLASH DISTRIBUTIONS OF FOUR SUCCESSIVE EVENTS.*

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, flash numbers (manually counted), and flash densities of the corresponding events, respectively.

3) *Picture-taking Activities:* 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 the relative importance of a wedding event. Table II illustrates the generally observed frequency of taking pictures during various wedding events. Since the occurrence of camera flashes correlates closely with the activity of picture-taking [14], the estimation of flash density could be an effective visual cue for wedding event discrimination. Table III shows an example of flash distributions for four successive wedding events in a ceremony. We observed high variations in flash distributions among events. For example, the *WK* event is merely 12 seconds long, but there are 73 flashes. Its density reaches six times per second, on average. By contrast, the *OP* event is of relatively less importance to the audiences, and it contains a small number of flashes even if it lasts for a much longer duration.

4) *Leading Roles:* As shown in Table II, the leading roles involved in various wedding events are different. For example, groom and the best man are the main characters in the *GE* event; the groom, his bride, and officiants are the main focuses in the *RE* event. The main characters' occurrence pattern gives a visual hint for the event category. A naïve solution would be to recognize all roles in videos. This is, however, not a trivial task with today's technology. Fortunately, there are some simple tricks to detect the bride, inarguably the most important focus of a wedding. According to the western tradition [1], [2], the bride invariably wears a white gown and veil as a symbol of purity but the other female roles have flexibility in their dress color. Therefore, it is more reliable to represent the bride's appearance assuming she wears white.

B. Selected Features for Event Modeling

Based on the observations of Section III-A, four kinds of audiovisual features, related to the scopes of speech/music discrimination, applause detection, flash detection, and bride indication, are developed as basic features for event modeling.

1) *Event Features Related to Speech/Music Discrimination:* As mentioned in Section III-A.1, the audio recordings of weddings are often with poor quality. The selected audio features have to be discriminative enough between speech/music types for the given low-SNR inputs. However, in the literature, most studies address the discrimination problem only for clean data or with the assumption of known noise types [11], [15]. To identify the audio features that are resistant to noises, we first

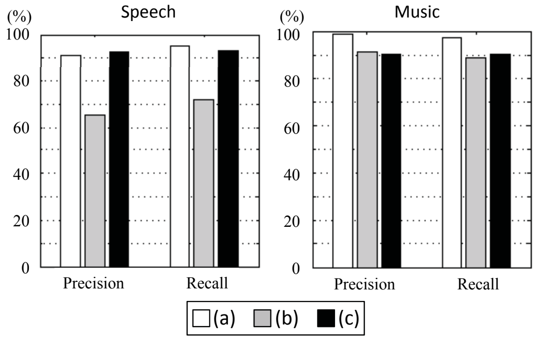


Fig. 1. Classification results of the audio types of speech (the left subplot) and music (the right subplot) on three audio datasets of (a) Internet radio, (b) Internet radio with added white noises (5 dB), and (c) audio tracks from home videos. (See Section III-B.1 for details.)

collect a comprehensive set of candidate features from the previous work [11], [15], [16] and determine the more reliable ones using feature selection algorithms [17], [18].

Initially, tens of audio features are collected to form a candidate set, including the short-time energy, energy crossing, band energy ratio, root mean square (RMS), normalized RMS variance, zero crossing (ZC), joint RMS/ZC, bandwidth, silent interval frequency, mel-frequency cepstral coefficients (MFCCs), frequency centroid, maximal mean frequency, harmonic degree, music component ratio, etc. [11], [15], [16]. Each of the collected audio features is assessed by information theoretical measures [17], [18], so as to estimate its discriminability between the speech and music types. At the end, three of them are chosen for their stable performances under various noise types. They are the one-third energy crossing (OEC), the silent interval frequency (SIF), and the music component ratio (MCR), as detailed in our prior work [29].

As a result, we use OEC, SIF, and MCR to practically realize a multi-class SVM classifier for speech/music discrimination [21]. The classifier has been evaluated on three small audio datasets, each containing approximately three-hour sources. The first dataset is collected from Internet radio and the second is obtained by adding 5 dB white noises to the first one. In addition, we constitute the third one from audio tracks of two kinds of home videos, i.e. the wedding and the birthday party. Here, sound of birthday party is included because its audio contents have higher variations and contain more diversified sound effects. For example, some are taken place at a quiet home, and others are in a noisy restaurant with crowd laughing, talking, and cheering. Then, a fivefold cross-validation experiment [9] is conducted for the classifier on each dataset and the results measured by average precisions and recalls are illustrated in Figure 1. The performance shows that the proposed audio features work quite well even for the audio with a substantial amount of noises.

2) Event Features Related to Applause Detection: The same feature selection mechanisms, as described in the previous section, are applied to identify the noise-resistant audio features for detecting the presence of applause in low-SNR audio recordings. However, based on our experiments, the audio features in the previous section generally do not perform very well. Instead, a specific audio feature is developed for applause detection. This feature exploits the physical properties of applause, indicated in Section III-A.2: when applause is coming up in the audio signal, a significant increase in magnitude can be observed over the whole power spectrum [12], [13]. An example is illustrated in Figure 2(a). For comparison, two power spectrums taken from consecutive time instances of a wedding audio are depicted in the same figure. The spectrum with applause (the top solid curve) is around 20 dB larger in magnitude than the one without applause (the bottom dotted curve) for almost all frequencies. To capture the global variations of audio magnitudes, an audio feature of the weighted short-time energy (WSE) is employed.

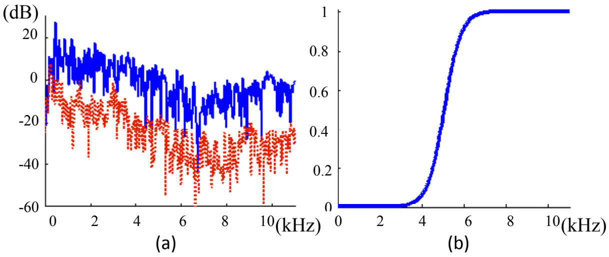


Fig. 2. Examples of (a) two power spectrums of a wedding audio from consecutive time instances, one with applause (the top solid curve) and another without applause (the bottom dotted curve), and (b) a sigmoidal filter function.

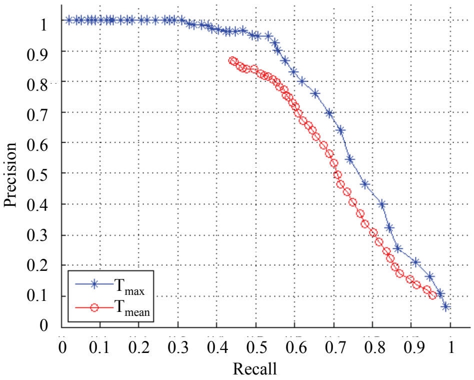


Fig. 3. Precision-recall curves of the applause detection results using two different thresholds. (See Section III-B.2 for details.)

- ***Weighted Short-time Energy (WSE).*** The feature value of weighted short-time energy is defined as the weighted sum over the spectrum power (in decibels) of an audio signal at a given time as follows:

$$\text{WSE} \triangleq \frac{1}{\text{WSE}_{max}} \int_0^{\omega_s} W(\omega) \cdot 10 \log(|SF(\omega)|^2 + 1) d\omega \tag{1}$$

where $SF(\omega)$ is the short-time Fourier transform coefficient of the frequency component ω, and $W(\omega)$ is the corresponding weighting function. In addition, ω_s denotes the sampling frequency and WSE_{max} is the maximum WSE in the audio track as a normalization factor. The calculation of WSE is special in that the spectrum power is in a logarithmic unit of decibels. Summation in the decibel domain is the same as multiplication in the energy domain. The logarithmic nature leads that a large WSE

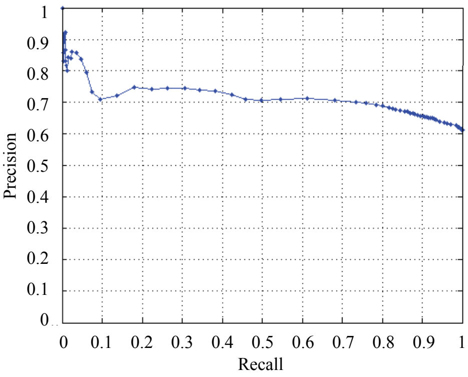


Fig. 4. Precision-recall curves of the bride indication results. (See Section III-B.4 for details.)

value comes from a global trend of high power over the whole spectrum but not few dominant frequencies. In a wedding, since human speech is commonly observed and the signals are bandlimited to around 3.2 kHz [15], $W(\omega)$ is chosen to be a sigmoidal function (cf. Figure 2(b)) in order to suppress the contributions from low frequencies.

$$W(\omega) = \frac{1}{1 + e^{-\omega_1(\omega - \omega_2)}}, \tag{2}$$

where ω_1 and ω_2 are control parameters and are respectively set to 2.5 (kHz) and 5.0 (kHz). The input audio track is first segmented into non-overlapping 1-second audio frames. For each audio frame, one feature value is computed for every 50-ms interval with a 10-ms overlap. A median filter is then applied to diminish possible noises. Instead of aggregation, based on our experiments, the maximum of these 25 feature values is selected as the representative WSE feature for that 1-second frame.

To verify the capability of WSE, a simple trial is conducted to detect the applause presented in audio recordings using two different thresholds: T_{max} and T_{mean}. That is, given a series of WSE values, we compute two thresholds by individually multiplying the maximum value and their mean to a numerical factor between [0,1]. Then applause can be located at the positions with higher WSE values than the chosen threshold. Figure 3 illustrates the average detection results on 15 audio tracks from a set of collected home videos, including wedding and birthday parties. The inclusion of birthday parties is for the same reason as described in Section III-B.1. Overall, the performance is acceptable and shows that WSE can capture applause effectively even for noisy home video recordings.

3) Event Features Related to Flash Detection: Flashes of picture-taking can be detected from abrupt and short increases of the global intensity in a video frame. As suggested in Section III-A.3, a visual feature of the flash density (FLD) is defined. See our prior work [29] for details.

To get more insight into the feature of FLD, we apply the flash detection algorithm to one wedding video used in later experiments, i.e. the Clip-A in Table V. In terms of flash numbers, 457 flashes are correctly detected among the 482 true ones, and there are 17 false positives. The detecting precision and recall are 94.81% and 96.41%, respectively. The results show that flashes can be robustly captured with our feature.

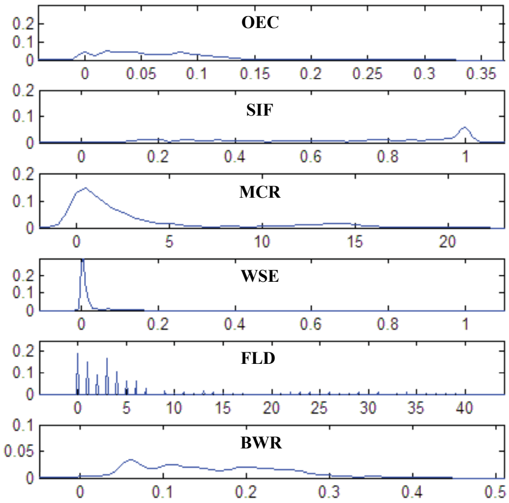


Fig. 5. Example of the *RE* wedding event model.

4) Event Features Related to Bride Indication: As mentioned in Section III-A.4, the bride is an important leading role in wedding events and her appearance can be detected by the color of "bridal white". However, due to various lighting conditions, the determination of real bridal white is extremely difficult and often needs a laborious training process similar to that of the skin color detection [22]. Instead, we approximate bridal white map for each video frame in our prior work [29], whereby a corresponding visual feature, bridal white ratio (BWR), can then be defined [29].

For understanding its performance, a simple trial is carried out for the bride indication by making binary decisions (i.e. presence or absence) on the basis of the obtained BWR values. Given a predefined threshold, a higher BWR value corresponds to the bride's presence, otherwise her absence. Figure 4 illustrates precision-recall curves of the detecting results for the Clip-A in Table V. The "hard-decision" performance is promising and we believe that the resulted "soft-decision" BWR is helpful for our modeling task.

IV. Wedding Modeling

The objective of wedding modeling is to estimate the event sequencing of a wedding video. At each time instance, extracted event features are exploited to recognize the wedding events. In addition, a wedding video is a kind of sequential data. The occurrence of a wedding event highly depends on the category of its preceding neighbors. 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, HMM, to describe the spatio-temporal relations of events within a wedding video [9]. In Sections IV-A and IV-B, we first build statistical models for feature similarity and temporal ordering for each of the wedding events. Section IV-C then devises an integrated HMM framework for both the event-based analysis and the wedding segmentation.

TABLE IV

AN EVEN TRANSITION MODEL OF THE WEDDING EVENTS.

	ME	*GE*	*BE*	*CS*	*OP*	*WV*	*RE*	*BU*	*MS*	*WK*	*AP*	*ED*	*OT*
ME	0.80	0.11	0.09										
GE	0.12	0.80	0.08										
BE			0.80	0.04	0.16								
CS				0.80	0.16							0.01	0.03
OP				0.07	0.80	0.03	0.01			0.01	0.02	0.02	0.04
WV						0.80	0.13	0.03					0.03
RE						0.03	0.80	0.13					0.03
BU								0.80		0.20			
MS					0.07				0.80		0.07	0.07	
WK				0.03	0.11				0.03	0.80			0.03
AP				0.12					0.04		0.80	0.04	
ED												1.00	
OT				0.05	0.14				0.02				0.80

Note that we uniformly divide the wedding video into a sequence of 1-second units. The main reason is that we can not use conventional video units, such as shots, as the basic analysis units because they cannot be reliably obtained using conventional techniques as mentioned in Section I. In addition, uniform segmentation makes online processing possible. For convenience, let E denotes an index set [23] of the wedding events, where the indexing consists of a *bijective* mapping from the event set $E_S = \{ME, GE, \ldots, OT\}$ to a set of natural numbers, i.e. $E = \{1, 2, \ldots, |E_S|\}$. Similarly, F is an index set corresponding to the collection of event features $F_S = \{\text{OEC}, \text{SIF}, \text{MCR}, \text{WSE}, \text{FLD}, \text{BWR}\}$. For the t-th video unit, let $\mathbf{e}_t \in E$ be the corresponding state variable that indicates the occurrence of a specific wedding event, and let $\mathbf{x}_t = (x_t^1, \ldots, x_t^{|F|})$ be the feature vector associated with the specific event features x_t^j, $j \in F$.

A. 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 likelihood of feature values. The use of statistical histograms [20] is a naïve approach, but their discrete nature often causes unwanted discontinuity in results, especially when a feature value locates near the bin boundaries. Instead, we accumulate the probability by regarding each feature sample as a Gaussian centered at the sample. Assume that, for the i-th event, we have N samples for the j-th feature $\{x_1^j, \ldots, x_N^j\}$ extracted from the training clips. The distribution $p_{i,j}$ of the j-th feature for the i-th event can then be obtained as

$$p_{i,j}(\mathbf{x}) = \frac{1}{N} \sum_{n=1}^{N} \frac{1}{\lambda_j \sqrt{2\pi}} e^{-(\mathbf{x} - x_n^j)^2 / 2(\lambda_j)^2}, \ \forall i \in E, \ \forall j \in F, \tag{3}$$

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 feature. That is, if the extracted feature samples are more accurate and reliable, we can set λ_j to a smaller value.

Since the feature models are used for discriminating the wedding events, the divergence among feature models of different wedding events should be as large as possible. Quantitatively, the divergence of two probability distributions $\mathbf{p}$ and $\mathbf{q}$ can be defined by the symmetric Kullback-Leibler (SKL) distance, $D_{SKL}(\mathbf{p}, \mathbf{q})$ [17]. 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. That is,

$$\lambda_j = \arg\max_{\lambda} \sum_{i,k \in E, \ i<k} D_{SKL}(p_{i,j}, p_{k,j}) \tag{4}$$

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 $\lambda_{OEC} = 0.005$, $\lambda_{SIF} = 0.015$, $\lambda_{MCR} = 0.5$, $\lambda_{WSE} = 0.0025$, and $\lambda_{BWR} = 0.01$. It is worthy to notice that FLD 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 the sample sequences of FLD feature values 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 that this video unit belongs to the i-th wedding event:

$$p(\mathbf{x}_t | \mathbf{e}_t = i) = \prod_{j=1}^{|F|} p_{i,j}(x_t^j) \tag{5}$$

Note that, in practice, we compute the log-likelihood by taking logarithm of the expression, and thus obtain a contributive weight κ_j to the j-th feature model, where $\sum_j \kappa_j = 1$. In our experiments, we used a fixed set of weights, i.e. $\kappa_{OEC} = 0.25$, $\kappa_{SIF} = 0.2$, $\kappa_{MCR} = 0.1$, $\kappa_{WSE} = 0.1$, $\kappa_{FLD} = 0.1$, and $\kappa_{BWR} = 0.25$. They are automatically specified by optimizing the recognition accuracy of wedding events through a cross-validation process (cf. Section V). An interesting phenomenon is that the audio-based event features take as high as two-thirds of the weights. This implies that audio information seems more crucial for the wedding analysis.

Overall, the proposed event modeling has the following advantages. First, it has good 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 can be faithfully represented without approximation. Figure 5 gives an example of feature statistical models.

B. 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, ETM can be defined by an $|E| \times |E|$ matrix A as follows:

$$A_{i,k} = Pr(\mathbf{e}_t = k | \mathbf{e}_{t-1} = i), \ \forall i, k \in E \tag{6}$$

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

TABLE V

THE COLLECTION OF SIX WEDDING VIDEOS USED IN OUR 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 VI

THE STATISTICS OF MEANS μ AND VARIANCES σ^2 OF EVENT DURATION FOR EACH EVENT CATEGORY IN OUR VIDEO COLLECTION (UNIT: SECONDS).

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 VII

THE RECOGNITION RESULTS OF ALL WEDDING EVENTS (UNIT: 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	80	0	**350**	0	0	0	0	0	0	0	0	0	0	**81.40**
CS	0	0	0	**2320**	93	0	0	0	0	42	64	0	154	**86.79**
OP	1	0	5	212	**3622**	145	459	4	0	2	28	8	156	**78.03**
WV	0	0	0	43	77	**602**	73	0	0	0	0	0	0	**75.72**
RE	0	0	0	0	55	152	**442**	6	0	0	0	0	0	**67.48**
BU	0	0	0	0	0	0	0	**183**	0	2	0	0	0	**98.92**
MS	0	0	0	9	113	0	0	0	**143**	0	0	0	0	**53.96**
WK	0	0	0	0	0	0	0	0	0	**87**	0	0	0	**100.00**
AP	30	0	0	23	2	0	0	0	0	0	**164**	0	2	**74.21**
ED	0	0	0	0	3	0	0	0	0	0	0	**427**	0	**99.30**
OT	0	0	0	586	509	130	96	17	0	0	48	0	**436**	**23.93**
RP(%)	**80.09**	**100.00**	**86.42**	**72.66**	**80.96**	**58.50**	**41.31**	**87.14**	**100.00**	**65.41**	**53.95**	**98.16**	**58.29**	

In fact, given a training set of wedding videos with the event ground truth, we can tabulate an approximation of 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 fact that 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 as follows:

$$A_{i,k} = \begin{cases} \gamma_i \tilde{A}_{i,k} & , \ i = k \\ (1-\gamma_i \tilde{A}_{i,i})/(1-\tilde{A}_{i,i}) \cdot \tilde{A}_{i,k} & , \ i \neq k \end{cases}, \ \forall k \in E \tag{7}$$

where γ_i is the regularization factor in the range of $[0, 1]$. 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 approximately 80% probabilities along each row, i.e. $A_{i,i} \approx 0.8$.

Table IV shows the ETM we learnt from training videos, in which the blank entries represent zero probabilities. Sparsity of the ETM shows that few types of event transitions are allowed. It also demonstrates the occurrence of wedding events has a strong temporal correlation. This fact helps to reduce the computation cost and to increase the reliability of the determined event sequencing.

C. 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, we exploit an HMM framework for segmenting wedding videos, which integrates the wedding event statistical models (Section IV-A) and the event transition model (Section IV-B).

Specifically, given an input wedding video V, it is first partitioned into N 1-second video units, $V = \{\mathbf{v}_1, \ldots, \mathbf{v}_N\}$. For each video unit $\mathbf{v}_t$, $t \in \{1, \ldots, N\}$, we have a set of $|F|$ event features associated with it, i.e. $\mathbf{x}_t = (x_t^1, \ldots, x_t^{|F|})$. Collecting all the observations $X = \{\mathbf{x}_1, \ldots, \mathbf{x}_N\}$, our goal is to find the most probable event sequencing S for V, where $S = \{\mathbf{e}_1, \ldots, \mathbf{e}_N\}$. Therefore, we develop a left-to-right HMM with $|E|$ states $\{e^i | i \in E\}$, in which each state corresponds to one of the adopted event categories. The HMM is governed by

TABLE VIII

THE RECOGNITION RESULTS SOLELY BASED ON THE FEATURE SIMILARITY WITHOUT EXPLOITING THE EVENT TRANSITION MODELING.

Events	ME	GE	BE	CS	OP	WV	RE	BU	MS	WK	AP	ED	OT
(a) using audio features only													
RP(%)	34.54	30.14	42.57	78.39	69.81	0	0	71.55	0	12.09	0	34.80	69.32
RR(%)	87.39	59.86	87.21	64.46	88.49	0	0	44.86	0	96.55	0	80.93	13.89
(b) using visual features only													
RP(%)	20.30	12.58	30.64	47.10	62.68	36.61	0	20.40	0	4.32	0	16.59	45.49
RR(%)	87.74	66.20	29.07	45.23	14.81	8.43	0	44.32	0	87.36	0	74.65	11.91
(c) using audiovisual features all													
RP(%)	44.12	21.62	55.13	75.04	76.01	81.48	0	28.72	0	14.38	0	38.52	71.51
RR(%)	93.96	69.72	73.72	73.55	91.29	5.53	0	45.95	0	100.00	0	83.49	21.08

a set of parameters, $\theta = \{\pi, A, \phi\}$, where π, A, and ϕ are the initial state probabilities, the state transition probabilities, and the emission probabilities, respectively [9]. 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 presumed to be a uniform distribution, i.e. $p(\mathbf{e}_1 = i|\pi) = 1/|E|, \forall i \in E$. Accordingly, our goal for finding the optimal sequencing S can be formulated as

$$\begin{aligned} S &= \arg\max_{s} \; Pr(X, S|\theta) \\ &= \arg\max_{s} \; p(\mathbf{e}_1|\pi) \left[\prod_{t=2}^{N} p(\mathbf{e}_t|\mathbf{e}_{t-1}, A) \right] \prod_{t=2}^{N} p(\mathbf{x}_t|\mathbf{e}_t, \phi) \\ &= \arg\max_{s} \; p(\mathbf{e}_1|\pi) \left[\prod_{t=2}^{N} A_{\mathbf{e}_{t-1}, \mathbf{e}_t} \right] \prod_{t=2}^{N} \prod_{j=1}^{|F|} p_{\mathbf{e}_t, j}(x_t^j) \qquad (8) \end{aligned}$$

where the last two terms are derived from Eqns. (5) and (6), respectively. Because the HMM trellis is equivalent to a directed tree, the solution of S can be efficiently obtained using the Viterbi algorithm [9].

After labeling each 1-second unit of the input video, 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 errors. 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 proceeding and succeeding neighbors belong to different event categories, it is merged into the left one; otherwise, all the three events are merged into one event.

V. EXPERIMENTAL RESULTS

This section first presents experimental results for the evaluation of the proposed framework in wedding event recognition (Section V-A) and wedding ceremony video segmentation (Section V-B). This, we show comparisons with another well-known algorithm, linear-chain conditional random fields (LCRF), and an extension of our system to a practical scenario in Sections V-C and V-D, respectively.

In our experiments, we used a total of six wedding video clips, as shown in Table V. Each of them contains a complete recording of a wedding ceremony. Three observers (none of the clip owners) collaboratively annotated the event ground truth. Our 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. In addition, our current system is programmed using Matlab without code optimization, and running on a machine with Intel P4 3.0 GHz CPU. Based on Section V-A, the average testing time for a clip is about 15 times longer than its original video length, and the feature extraction accounts for around 96% of the time.

A. Event Recognition Analysis

Table VII summarizes the event recognition results in unit of seconds, presented in the form of a confusion matrix [19], 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) are also reported. As described in Section I, since the actual event boundaries 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 among different wedding events as shown in Table VI(a). Initially, for each event category, all of the event samples are sorted by duration in descending order. We then compute a truncated mean $\tilde{\mu}_i$ of the event duration (Table VI(b)) by ignoring the samples of the first half (i.e. the longer ones in the top half), and the range value is set to $\min(0.2\tilde{\mu}_i, \xi)$, where we set $\xi = 10$ so that the tolerant ranges vary according to event categories but do not exceed 10 seconds. Here, we use a truncated mean but not the standard mean because of its better statistical reliability. That is, as shown in Table VI(b), the truncated variances are generally much smaller than the standard ones in Table VI(a), which implies that durations of the shorter samples would be more consistent. More importantly, by ignoring the longer samples, a smaller tolerant range can be naturally obtained to enforce a stricter standard for recognition hits.

Overall, as shown in Table VII, large amounts of the detected wedding events reach over 70% in both RP and RR values. Some of them even achieve the level of 85%, such as *BU* and *ED* events. Several observations could be 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

TABLE IX

THE SEGMENTATION RESULTS WITHOUT DURATION-BASED FILTERING (UNIT: EVENT SEGMENTS).

Clip	Corr.	Sub.	Ins.	Del.	SP(%)	SR(%)	SF(%)
A	16	1	10	0	59.26	94.12	72.73
B	5	1	0	2	83.33	62.50	71.43
C	28	2	19	5	57.14	80.00	66.67
D	22	1	18	0	53.66	95.65	68.75
E	12	0	6	3	66.67	80.00	72.73
F	12	1	9	1	54.55	85.71	66.67
Avg.					**62.44**	**83.00**	**71.27**

TABLE X

THE SEGMENTATION RESULTS WITH DURATION-BASED FILTERING (UNIT: EVENT SEGMENTS).

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	27	1	10	7	71.05	77.14	73.97
D	21	1	12	1	61.76	91.30	73.68
E	12	0	3	3	80.00	80.00	80.00
F	11	0	6	3	64.71	78.57	70.97
Avg.					**72.26**	**80.61**	**76.21**

MS, is sometimes arranged to be accompanied with choirs singing and the whole ceremony is generally hosted by wedding officiants who often give some short presentations within a wedding event. They also cause severe degradations in RP values for both *RE* and *AP* 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 misclassified 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 IV, 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 *OT* event is relatively low. This is due to the fact that *OT* event is inherently varied in forms. For example, it could be 'reading of poetry' or 'lighting of the unity candle'. Moreover, it severely influences the overall recognition performance by spreading out the recognition errors over various event categories.

As a comparison with the HMM-based modeling, we perform event recognition solely based on the maximum similarity of audiovisual features among the events without exploiting the temporal relation of event transitions. Table VIII shows the results when using (a) the four audio features only, (b) the two visual features only, and (c) all six audiovisual features. Generally, the use of both audio and visual features together outperforms the use of either unimodal features alone, but the results are still not as good as those in Table VII, in which event transition modeling is augmented. This shows evidences to support the effectiveness of the HMM framework.

B. 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 at the 'second' level. We follow a similar idea exploited in the longest common substring problems [24]. That is, we represent a wedding video as a symbol string where the alphabet consists of the event codes given in Table I. 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, the segmentation performance is measured by 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 [29]. The less the edit operations are needed, the better the segmented videos match with the ground truth.

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

$$\text{SP} = \text{Corrects}/(\text{Corrects} + \text{Substitutions} + \text{Insertions}), \quad (9)$$

$$\text{SR} = \text{Corrects}/(\text{Corrects} + \text{Substitutions} + \text{Deletions}). \quad (10)$$

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

From Table IX, 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 number of events as shown in Table V. 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, consider an event subsequence of the ground truth, *WK-CS-ED*. If *CS* is not detected and the direct transition from *WK* to *ED* is not allowed (i.e. the transition probability equals zero), the HMM framework would be forced to go through a longer path of erroneous events to connect *WK* and *ED*, such as *WK-OT-MS-ED*. Also, when a succession of two events has never been observed in the training data, its zero transitive probability could cause the same problem. 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 V-A.

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 correct the abnormal ones. Specifically, for each of the event categories, we exploit the truncated models (Section V-A and Table VI(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 X summarizes the segmentation results

TABLE XI
THE PERCENTAGE OF TOTAL EVENT DURATION FOR EACH OF THE EVENT CATEGORIES IN OUR VIDEO COLLECTION

Event	ME	GE	BE	CS	OP	WV	RE	BU	MS	WK	AP	ED	OT	
	4.45%	1.09%	3.30%	20.53%	35.87%	6.11%	5.03%	1.42%	2.04%	0.67%	2.19%	3.30%	14.00%	100%

TABLE XII
LCRF RECOGNITION RESULTS OF ALL WEDDING EVENTS (UNIT: SECONDS).

Events	*ME*	*GE*	*BE*	*CS*	*OP*	*WV*	*RE*	*BU*	*MS*	*WK*	*AP*	*ED*	*OT*	**RR(%)**
ME	**394**	0	30	155	0	0	0	0	0	0	0	0	0	**68.05**
GE	0	**99**	18	25	0	0	0	0	0	0	0	0	0	**69.72**
BE	51	0	**339**	16	0	0	0	0	0	0	0	0	24	**78.84**
CS	0	0	0	**2221**	164	0	0	36	0	12	0	4	236	**83.09**
OP	0	0	0	403	**3967**	0	0	0	0	6	0	2	261	**85.51**
WV	0	0	0	0	356	**283**	37	0	0	0	95	0	24	**35.60**
RE	0	0	0	0	270	0	**300**	0	0	0	85	0	0	**45.80**
BU	0	0	0	0	11	0	20	**58**	0	0	9	18	69	**31.35**
MS	0	0	0	0	17	0	0	0	**137**	0	0	0	111	**51.70**
WK	0	0	0	0	39	0	0	5	0	**19**	0	11	13	**21.84**
AP	17	0	0	66	0	0	0	0	0	0	**48**	0	90	**21.72**
ED	0	0	0	0	0	0	0	0	83	26	0	**289**	32	**67.21**
OT	0	0	0	545	749	0	0	0	0	8	56	118	**346**	**18.99**
RP(%)	**85.28**	**100.00**	**87.60**	**64.73**	**71.18**	**100.00**	**84.03**	**58.59**	**62.27**	**26.76**	**16.38**	**65.38**	**28.69**	

after applying the duration-based filtering. Compared with Table IX, 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 X, the 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 give us support and confidence that, as long as we capture well the content characteristics, it is possible to conduct high-level semantic analysis of home videos through the use of generic and easily extracted audiovisual features.

C. Performance Comparisons with LCRF Models

To further evaluate the validity of HMM approach, we compare the performance of HMM with that of the linear-chain conditional random fields (LCRF) [25], [26]. LCRF is a well-known probabilistic framework for labeling and segmenting sequence data. In statistical relational learning, HMM and LCRF are known as a *generative-discriminative pair* [26], in the sense that HMM measures the joint probability of sequential observations and the corresponding label sequences but LCRF is to estimate the conditional probability of associated label sequences given the observations. Therefore, LCRF is also believed to be superior to HMM in representing long-range dependencies of the observations [25], [26].

As a comparison to HMM framework (cf. Section IV-C), in LCRF modeling the goal to find the optimal sequence S given observations X is differently formulated as

$$S = \arg\max_{s} \; Pr(S|X, \theta^{'}) \qquad (11)$$

where $\theta^{'}$ denotes the model parameters. A quasi-Newton method (i.e. BFGS [26]) is then adopted to optimize the estimation of $\theta^{'}$ from the training data. Following the same experimental procedures described in Sections V-A and V-B, both the event recognition and the video segmentation results are summarized in Tables XII, XIII, and XIV.

In Table XII, some observations can be made: 1) As compared with the results of HMM in Table VII, LCRF performs much worse in RR values than RP values. A half of the RR values is below the 50% level and some are even down to the level of 20%, such as *WK*, *AP*, and *OT* events. Also, the events with low RR values often have relatively lower RP values. This phenomenon might partly come from two reasons. One is the inherent event properties and the other is the unbalanced amount of training samples. For example, as shown in Table XI, the summed percentage of total event duration for these low-RR events is less than that of a single *OP* event in our video collection. By contrast, HMM's performance (in both the RP and RR values) is more consistent and stable, cf. Section V-A. It seems that HMM approach could be more robust for unbalanced classification. 2) The low RR values are inappropriate for real applications. For example, the events of *WV*, *RE*, *BU*, and *WK* are arguably the most important moments in a wedding and also the most frequent pieces users would like to review in wedding videos [1]. However, a large number of those events are not detected by the LCRF model, cf. Table XII. By contrast, HMM approach performs better in the RR values, e.g. both *BU* and *WK* events are higher than 95%, although the corresponding RP values are comparably lower. For users, they would more like to see "fakes" rather than totally miss anything important. 3) Similar to HMM results, *OT* event is still a main culprit for bad recognition performance and the performance is even worse for LCRF model. In Table XII, the effects can be observed from the widespread errors associated with the *OT* event.

Tables XIII and XIV give the video segmentation results of LCRF model, with and without duration-based filtering. From Table XIII, we can see that most SR values are only

TABLE XIII
LCRF SEGMENTATION RESULTS WITHOUT DURATION-BASED FILTERING
(UNIT: EVENT SEGMENTS).

Clip	Corr.	Sub.	Ins.	Del.	SP(%)	SR(%)	SF(%)
A	16	1	11	0	57.14	94.12	71.11
B	5	1	0	2	83.33	62.50	71.43
C	26	3	17	6	56.52	74.29	64.20
D	20	0	16	3	55.56	86.96	67.80
E	10	1	2	4	76.92	66.67	71.43
F	11	0	15	3	42.31	78.57	55.00
Avg.					**61.93**	**77.19**	**68.72**

TABLE XIV
LCRF SEGMENTATION RESULTS WITH DURATION-BASED FILTERING
(UNIT: EVENT SEGMENTS).

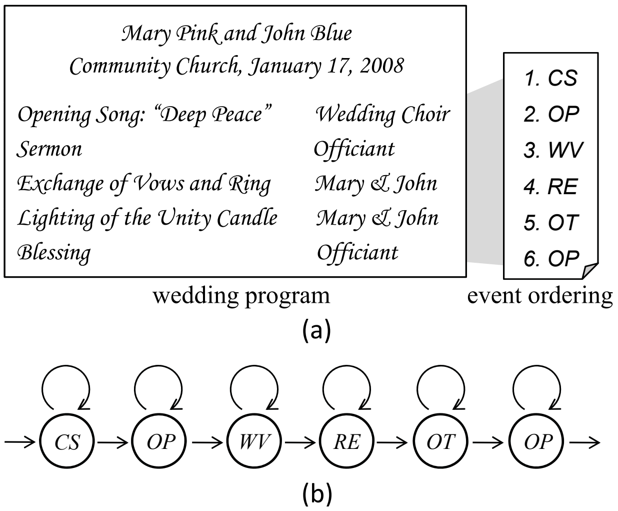
Clip	Corr.	Sub.	Ins.	Del.	SP(%)	SR(%)	SF(%)
A	13	0	2	4	86.67	76.47	81.25
B	3	1	0	4	75.00	37.50	50.00
C	17	2	5	16	70.83	48.57	57.63
D	16	0	4	7	80.00	69.57	74.42
E	10	0	1	5	90.91	66.67	76.93
F	9	0	4	5	69.23	64.29	66.67
Avg.					**78.77**	**60.51**	**68.44**

around 60% and 70% levels. In comparison with HMM results (cf. Table IX), the degradation is due to that more ground-truth events were not detected. On the other hand, an interesting thing is the burst increase in number of deletions after duration-based filtering, as shown in Table XIV. For example, the deletions for clip C rise to near three times as the original. Based on our observations, it is also caused by the low RRs as described above. This fact makes the detected duration of events tend to be shorter than their actual lengths in the ground truth and easy to be removed during the filtering. In terms of duration, our HMM approach would be more accurate to include complete contents in the detected events.

D. Extension to the Scenario with Known Event Ordering

In this section, we investigate an extension of our work to the scenario when the actual event ordering of a wedding video is available. The investigation is conducted for two purposes. First, by reducing the temporal uncertainty, it is more reliable for us to examine the true capability of the proposed audiovisual features in discriminating various wedding events. Second, the scenario creates an opportunity for users to interact with our system so as to possibly improve the segmentation accuracy. For example, the ordering information can be obtained by manual input or semi-automatic transcription from the couple's wedding programs [2], such as Figure 6(a).

Under the assumption of known event ordering, our original task is in some sense converted into the type of *change-point problem* [4], [24]. That is, the problem is to determine the set of boundaries where event transitions happen. Therefore, instead of using the proposed HMM framework, a modified state space model is built for each wedding video, in which each state corresponds to one of the known events and the states are arranged in the form of a Markov chain according to the given event ordering, as illustrated in Figure 6. Note that the directed edges are simply used to indicate allowable transitions but not assigned with any transition weights. The most probable event sequence is then computed by exploiting dynamic programming [9], [24], and the event boundaries can be automatically located at the points of transition among different states.



Mary Pink and John Blue
Community Church, January 17, 2008

Opening Song: "Deep Peace"	*Wedding Choir*
Sermon	*Officiant*
Exchange of Vows and Ring	*Mary & John*
Lighting of the Unity Candle	*Mary & John*
Blessing	*Officiant*

Fig. 6. (a) A sample wedding program accompanied with the transcribed event ordering, and (b) the state diagram in form of a Markov chain built according to the above event ordering.

Table XV summarizes the segmentation results, in which "Detects" are defined as the number of detected event segments and "Corrects" (cf. Section V-B) indicate the number of correct ones among "Detects". The overall performance is satisfactory. Both the precisions and recalls reach a high level of more than 80%. The results are very encouraging. It not only demonstrates the effectiveness of our audiovisual features but also implies that the minor requirement of user intervention could greatly improve performance for practical applications.

VI. Conclusions and Future Work

In this paper, we have presented an HMM-based system for event-based wedding video analysis and segmentation. According to the wedding customs, we developed a taxonomy for classifying wedding events, whereby a set of discriminative audiovisual event features are exploited for robust event modeling. It can help users to access, organize, and retrieve his/her treasured contents in an automatic and more efficient way. To the best of our knowledge, this work is the first one to analyze and structure wedding videos on the basis of semantic events. Actually, it might also be the first one for semantic event analysis on any domain of home videos.

Many aspects of our approach can be improved, as detailed below.

1) It is possible to explore more semantic features for event recognition. For example, speaker change detection or identification would be helpful in discriminating the events with dense speech, such as *WV* and *RE* events.
2) The modeling mechanisms could be improved. For example, on one hand, advanced fusion schemes of the feature models can be adopted. One example is hierarchical

TABLE XV

SEGMENTATION RESULTS IN THE CASE WHEN EVENT ORDERINGS ARE AVAILABLE (UNIT: EVENT SEGMENTS).

Clip	**Det.**	**Corr.**	**P(%)**	**R(%)**	**F(%)**
A	17	16	94.12	94.12	94.12
B	6	5	83.33	62.50	71.43
C	30	28	93.33	80.00	86.15
D	23	23	100.00	100.00	100.00
E	12	12	100.00	80.00	88.89
F	14	11	78.57	78.57	78.57
Avg.			**91.56**	**82.53**	**86.81**

classification that combines homogeneous features as mid-level concepts and then builds event models on top of these concepts [27], [28]. In addition, the cultural differences would be taken into account, such as the variety of speech characteristics (e.g. phonies) between eastern and western languages in the selection of audio features. On the other hand, the development of a time-variant event transition model could produce more reasonable event sequences. Moreover, there are other modeling tools worthy of further study for the wedding analysis, such as higher-order HMM, Bayesian network, finite state machine, their combinations, and so forth.

3) Other extensions of our work are to be investigated. For example, sometimes, multiple recordings of the same wedding ceremony are available from participants. The joint analysis would benefit the detection task of semantic events and enable more creative applications.
4) More extensive and thorough evaluation of our system is a must. Moreover, since home videos are private data and usually hard to be acquired, it is beneficial to have a common database and relevant evaluation benchmarks for wedding videos.

In the future, we will continue our investigation in these directions.

REFERENCES

[1] L.M. Spangenberg. *Timeless traditions: a couple's guide to wedding customs around the world*. Universe Publishing, New York, NY, 2001.

[2] D. Warner. *Diane warner's contemporary guide to wedding ceremonies*. New Page Books, Franklin Lakes, NJ, 2006.

[3] D. Gatica-Perez, A. Loui, and M.-T. Sun, "Finding structure in home videos by probabilistic hierarchical clustering," *IEEE Trans. Circuits and Syst. for Video Technol.*, vol. 13, no. 6, pp. 539–548, June 2003.

[4] Y. Zhai and M. Shah, "Automatic segmentation of home videos," in *Proc. 2005 IEEE Intl. Conf. Multimedia and Expo (ICME'05)*, pp. 9–12, 2005.

[5] R.S.V. Achanta, W.-Q. Yan, and M.S. Kankanhalli, "Modeling intent for home video repurposing," *IEEE Multimedia*, vol. 13, no. 1, pp. 46–55, Jan.-Mar. 2006.

[6] Z. Rasheed and M. Shah, "Detection and representation of scenes in videos," *IEEE Trans. Multimedia*, vol. 7, no. 6, pp. 1097–1105, Dec. 2005.

[7] P. Yin, X.-S. Hua, and H.-J. Zhang, "Automatic time stamp extraction system for home videos," in *Proc. 2002 IEEE Intl. Symp. Circuits and Syst. (ISCAS'02)*, pp. 73–76, 2002.

[8] M. Cooper, J. Foote, A. Girgensohn, and L. Wilcox, "Temporal event clustering for digital photo collections," in *Proc. 11th ACM Intl. Conf. Multimedia (MM'03)*, pp. 364–373, 2003.

[9] C.M. Bishop. *Pattern Recognition and Machine Learning*. Springer, New York, NY, 2006.

[10] `http://www.cmlab.csie.ntu.edu.tw/~wisley/`

[11] W.-H. Cheng, W.-T. Chu, and J.-L. Wu, "Semantic context detection based on hierarchical audio models," in *Proc. 5th ACM Intl. Workshop on Multimedia Information Retrieval (MIR'03)*, pp. 109–115, 2003.

[12] B.H. Repp, "The sound of two hands clapping: an exploratory study," *J. Acoust. Soc. Amer.*, vol. 81, no. 4, pp. 1100–1109, Apr. 1987.

[13] L. Peltola, C. Erkut, P.R. Cook, and V. Valimaki, "Synthesis of hand clapping sounds," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 15, no. 3, pp. 1021–1029, Mar. 2007.

[14] B.T. Truong and S. Venkatesh, "Determining dramatic intensification via flashing lights in movies," in *Proc. 2001 IEEE Intl. Conf. Multimedia and Expo (ICME'01)*, pp. 61–64, 2001.

[15] T. Zhang and C.-C. J. Kuo. *Content-based audio classification and retrieval for audiovisual data parsing*. Kluwer, Norwell, MA, 2001.

[16] C. Panagiotakis and G. Tziritas, "A speech/music discriminator based on rms and zero-crossings," *IEEE Trans. Multimedia*, vol. 7, no. 1, pp. 155–166, Feb. 2005.

[17] T.M. Cover and J.A. Thomas. *Elements of information theory*, 2nd ed. Wiley, Hoboken, NJ, 2006.

[18] H. Pen, F. Long, and C. Ding, "Feature selection based on mutual information: criteria of max-dependency, max-relevance, and min-redundacy," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 27, no. 8, pp. 1226–1238, Aug. 2005.

[19] Y. Li and C. Dorai, "Instructional video content analysis using audio information," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 14, no. 6, pp. 2264–2274, Nov. 2006.

[20] R.C. Gonzalez and R.E. Woods. *Digital image processing*, 2nd ed. Prentice-Hall, Upper Saddle River, NJ, 2001.

[21] R.-E. Fan, P.-H. Chen, and C.-J. Lin, "Working set selection using the second order information for training SVM," *J. Mach. Learning Research*, vol. 6, pp. 1889–1918, 2005.

[22] S.L. Phung, Sr. A. Bouzerdoum, and Sr. D. Chai, "Skin segmentation using color pixel classification: analysis and comparison," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 27, no. 1, pp. 148–154, Jan. 2005.

[23] R.L. Graham, D.E. Knuth, and O. Patashnik. *Concrete mathematics: a foundation for computer science*, 2nd ed. Addison-Wesley, Indianapolis, IN, 1994.

[24] T.G. Cormen, C.E. Leiserson, and R.L. Rivest. *Introduciton to algorithms*, 2nd ed. MIT Press, Combridge, MA, 2001.

[25] J. Lafferty, A. McCallum, and F. Pereira, "Conditional random fields: probabilistic models for segmenting and labeling sequence data," in *Proc. 18th Intl. Conf. Machine Learning (ICML'01)*, pp. 282–289, 2001.

[26] L. Getoor and B. Taskar. *Introduction to statistical relational learning*. MIT Press, Combridge, MA, 2006.

[27] J. Fan, Y. Gao, H. Luo, and R. Jain, "Mining multilevel image semantics via hierarchical classification," *IEEE Trans. Multimedia*, vol. 10, no. 2, pp. 167–187, Feb. 2008.

[28] C.G.M. Snoek, M. Worring, and A.W.M. Smeulders, "Early versus late fusion in semantic video analysis," in *Proc. 13th ACM Intl. Conf. Multimedia (MM'05)*, pp. 399–402, 2005.

[29] W.-H. Cheng, Y.-Y. Chuang, B.-Y. Chen, J.-L. Wu, S.-Y Fang, Y.-T. Lin, C.-C. Hsieh, C.-M. Pan, W.-T. Chu, and M.-C. Tien, "Semantic-event based analysis and segmentation of wedding ceremony videos," in *Proc. 9th ACM Intl. Workshop on Multimedia Information Retrieval (MIR'07)*, pp. 95–104, 2007.